Crime Generators, Deterrents, and Attractors in Micro-Places

James R. LeDoux
Advisor: Christopher Maxwell

Boston College Department of Economics
Undergraduate Honors Thesis
May 2017

Acknowledgments

The completion of this thesis would not have been possible without the guidance of my advisor, Professor Chris Maxwell. His active role in guiding my work and research direction both kept me on track with this project and helped to navigate empirical challenges. I am grateful for his time and expertise.

Abstract

Criminal hotspots are heuristically understood, but seldom well-defined and empirically evaluated. In this thesis, I examine the concentration of crime into microgeographic hotspots, testing both the extent to which this occurs across major cities, and the relationship between spatial features and crime. I find that roughly five percent of street segments are responsible for half of crime across major cities, with this concentration level being robust to changes in total crime rate and economic conditions over time. I also find a significant relationship between the presence of spatial features such as nearby schools, bus stops, bars, and graffiti with the crime level in microgeographic units. Through a routine activity and crime pattern theoretic interpretation, such spatial models of crime can help to identify features and facilities that attract, inspire, or deter crime. These findings have policy-relevant implications for both urban planning and police strategy, offering intuition as to where crime can be expected to concentrate, and how changes to local environments impact public safety.

Table of Contents

Acknowledgments 1

Abstract 2

1 Introduction 4

2 Crime at Place 5

3 Rational Choice Theory 9

4 Routine Activity Theory 12

5 Crime Pattern Theory 14

6 Graphs and Grids 15

7 Putting Crime in its Place 18

8 VIII. Data and Methods 20

8.1 Crime Data 20

8.2 Units of Analysis 23

8.3 Measure of Distance 24

8.4 Facilities 26

8.5 Spatial Features 27

8.6 Socioeconomic Features 28

8.7 Dependent Variables 29

9 Summary Statistics 30

10 Concentration and Stability of Crime in Micro Places 33

11 Hotspot Movement Over Time 41

12 Explanatory Power of Facilities and Spatial Features on Crime 45

12.1 Features Used 45

12.2 Class of Model 49

12.3 Model Specification 50

12.3.1 Ordinary Least Squares 50

12.3.2 Beta Regression 50

12.3.3 Logit 51

12.3.4 Standardized Logit 53

12.4 Results 54

13 Discussion 64

14 References 68

15 Data Sources 71

Crime Generators, Deterrents, and Attractors in Micro-Places

1 Introduction

Economics is an inherently interdisciplinary field of study, often borrowing from the fields of psychology, sociology and government, among others. In this thesis I examine yet another blurring of disciplinary lines, at the intersection of economics and criminology.

The primary focus of this thesis is what David Weisburd calls the law of concentration of crime at place (Weisburd 2015). Weisburd observes that 50 percent of the crime in a sample of five major cities occurs on only five percent of the cities' street segments. Observing a high degree of consistency in this relationship across cities and time, he deems it a law that such a relationship occurs across all large cities.

Environmental criminological research is crucial for understanding human behavior, designing safer cities, and shaping public policy. Such research is seeing direct applications in government and police strategy today, with the cities of Los Angeles and Atlanta employing predictive policing strategies taken directly from academia (PredPol 2015), and the White House establishing the Police Data Initiative in order to inspire further advances in the field (Smith and Austin 2015). As the supply of open government data improves, the volume and significance of research in microgeographic and environmental criminology will only continue to grow.

Weisburd's 2015 paper *The Law of Crime Concentration and the Criminology of Place* serves as a natural starting point for understanding the importance of microgeographic criminology, beginning with a meta-analysis of where this subfield fits in the broader research ecosystem, and transitioning into an exposition on the law of concentration of crime at place. After a brief foray into rational expectations, the theory through which economic models made their debut in criminology, I will introduce two important theoretical frameworks through which crime concentration can be analyzed: the routine activities and crime pattern frameworks. With the theoretical

foundations in place, I will then discuss empirical methods and outline the approach I am taking in this thesis.

The goals for exploring Weisburd's law of crime concentration are twofold. First, I aim to test the existence of this phenomenon in a larger sample of cities. Following this, I will explore potential causal factors of crime concentration by modeling the relationships between the features of a street segment's local environment and its observed level of criminal activity.

2 Crime at Place

Criminological research features a wide range of units of analysis. The dominant unit of analysis, accounting for nearly two-thirds of all publications in *Criminology*, is the individual person, drawing upon sociological and psychological analyses of criminal decision making (Weisburd 2015). The other third of criminology research includes analyses of situations (15%), macro places such as cities and states (11.1%), and meso-places such as census blocks and neighborhoods (8.3%). The two lowest-featured units of analysis are micro-places, such as street segments and addresses (4.3%), and institutions (3.1%) (Weisburd, 2015). This thesis will focus on the micro-place.

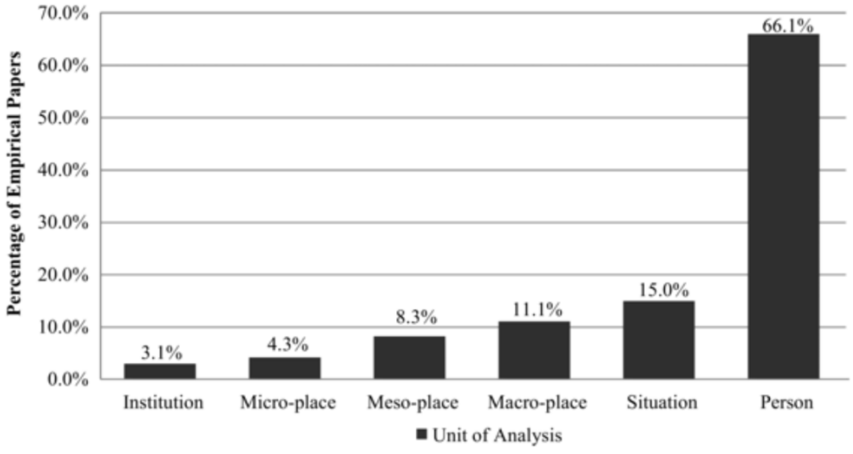


Figure 1: Composition of *Criminology* papers by unit of analysis, 1990 – 2014 (Weisburd 2015)

Analyses of micro-places have been few and far between historically, but comprise a quickly-growing portion of publications due to the recent influx of address-level data on 911 calls, incident reports, and various spatial features of cities.

Of this small but important subfield, perhaps the most important observation to date is that crime concentrates in relatively small geographic spaces. This has been known and discussed in academic research since at least as early as Guerry (1883) and Quetelet (1842), but has only recently become widely testable in well-defined units of analysis. Criminologist C. Ray Jeffery's work was among the first to empirically show the ways that crime clusters into hotspots in the early 1970s (Jeffery 1971), and later research went on to show that specific sub-categories of crime tend to have their own unique hotspot patterns (Sherman et al. 1989). As a specific example of this, Braga et al. (2010a) find that less than three percent of Boston's street segments accounted for over half of the city's instances of gun violence from 1980 to 2008, but also find that these were not necessarily the same street segments that accounted for a majority of its robbery incidents during this same period (2010b). Weisburd (2015) tests such theories of criminal hotspots at the street segment level across 8 different cities. Seeing stable and consistent ratios of the percentages of cities' street segments needed to explain fixed percentages of their total crime count, he proposes a general theory of crime concentration, the law of concentration of crime at place. In Weisburd's own words, the statement of the law is that "for a defined measure of crime at a specific microgeographic unit, the concentration of crime will fall within a narrow bandwidth of percentages for a defined cumulative proportion of crime." Specifically, Weisburd (2015) focuses on examining the percentages of street segments required to explain 50 and 25 percent of a city's crime. These are the street segments that comprise a city's principal crime hotspots.

Weisburd's sample includes five large and three small cities, with data coming from police incident reports over time periods

ranging between one and twenty years. The cities differed greatly in demographics, crime rate, population size, poverty rate, and total number of street segments. Despite this, all eight cities showed small and stable values for the percentages of street segments required to explain 25 and 50 percent of total crime, suggesting that such a law exists and that its relationship to the other factors commonly believed to affect crime rate is relatively inelastic.

The coupling of crime and place is not only observable and consistent across cities, but is also stable over time. The percentages of street segments explaining 25 and 50 percent of crime in major cities varied by little more than one percent over the time periods studied (Figure 2). These ratios remained stable despite volatile overall crime rates, which are represented by the dashed lines in Figure 2 below.

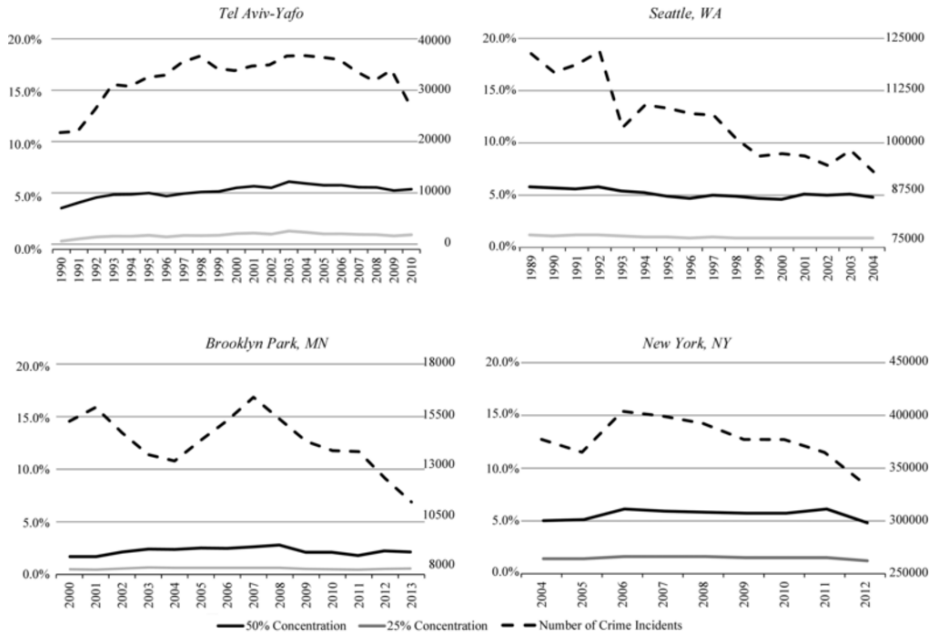


Figure 2: Crime Count, 50%, and 25% Concentration Levels in Major Cities Over Time (Weisburd 2015)

The Weisburd (2015) results also show, however, that such a law may apply to differing extents in small and large cities. The crime concentration results in Weisburd's analysis were similar across large cities, with the percentage of street segments required to explain 50

percent of crime ranging between 4.2 and 6 percent, and the percentage of segments required to explain 25 percent of crime ranging between 0.8 and 1.6 percent. The results were similarly consistent among smaller cities, with 50 percent of crime being explained by between 2.1 and 3.5 percent of street segments, and 25 percent of crime being explained by between 0.4 and 0.7 percent of street segments. There appears to be, however, a disconnect between the large and small cities in this sample, with small-city crime being more concentrated than large-city crime. This difference may suggest that this law does not hold uniformly across cities, but it is difficult to say so definitively with a small sample of only eight locations. Despite this potential sensitivity to city size, Weisburd (2015) explains that each city still shows a tight coupling of crime and place, and thus confirms his law of microgeographic crime concentration.

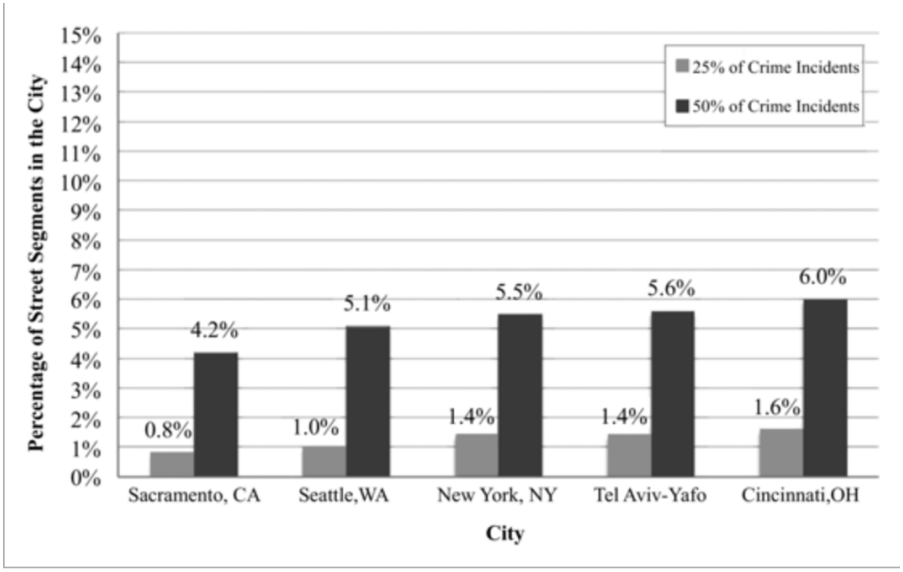


Figure 3: Crime Concentration in Large Cities (Weisburd 2015)

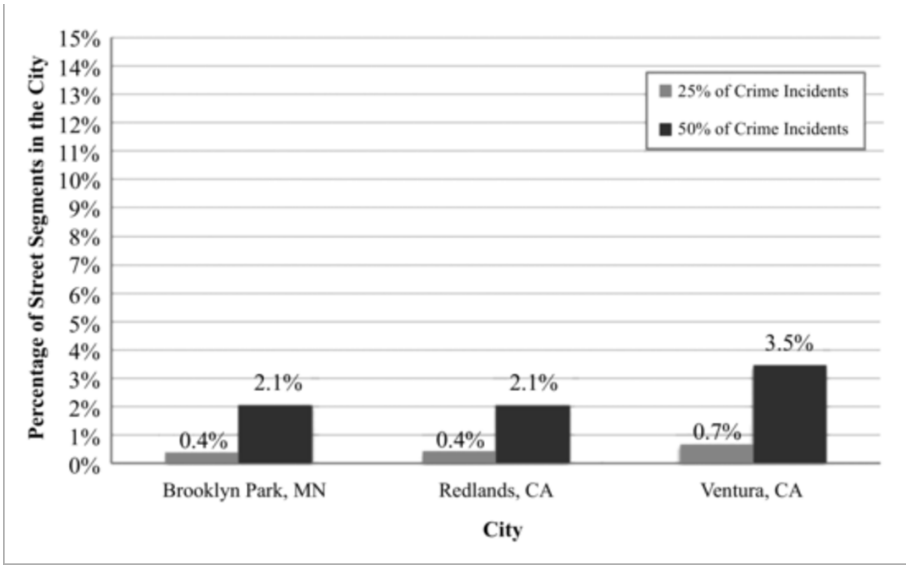


Figure 4: Crime Concentration in Small Cities (Weisburd 2015)

While there remain several unanswered and untested questions that follow from Weisburd's discussion of the law of concentration of crime at place, his work shows convincingly that crime does, in fact, aggregate into micro-places, and further that the street segment is an important unit of analysis for understanding crime patterns.

3 Rational Choice Theory

My research will focus on both an exploration of Weisburd's law of concentration of crime at place and an attempt to explain what exactly causes a street segment to become a criminal hotspot. In order to do this, it is first necessary to discuss the underlying body of theory regarding criminal decision-making, the coupling of crime and place, and the primary frameworks used to explain the relationships between geography, human psychology, and crime. The logical starting point for this is rational choice theory, developed by Nobel Prize winner Gary Becker in the late 1960s.

No legislation assumes obedience to the law; rather, it expects the opposite, focusing on what happens when it is broken. Becker's great insight was considering this legal penalty to be a cost, discussing a criminal's cost-benefit analysis, the amount of crime a government is

willing to tolerate, and how policy can be used as a tool to achieve an optimal level of law enforcement and public safety. Rational choice theory holds that

> "the optimal amount of enforcement is shown to depend on, among other things, the cost of catching and convicting offenders, the nature of punishments—for example, whether they are fines or prison terms—and the responses of offenders to changes in enforcement." (Becker 1968)

ECONOMIC COSTS OF CRIMES

Type	Costs (Millions of Dollars)
Crimes against persons	815
Crimes against property	3,932
Illegal goods and services	8,075
Some other crimes	2,036
Total	14,858
Public expenditures on police, prosecution, and courts	3,178
Corrections	1,034
Some private costs of combating crime	1,910
Overall total	20,980

SOURCE.—President's Commission (1967*d*, p. 44).

Figure 5: Costs of Crime, President's Commission 1967 (Becker 1968)

To emphasize this point on the costs of crime, Becker presents data on the reported costs incurred in the enforcement of the law from various categories, seen in Figure 5. In order to combat crime in an optimal manner, the models behind rational choice incorporate the behavioral relations underlying the costs described in Figure 5, formalizing the relationships between

1: the quantity and cost of crime,

2: the quantity of crime and severity of punishments,

3: the quantity of crime and expenditures on police and the court systems,

4: the number of convictions and the cost of imprisonments, and

5: the number of offenses and private expenditure on personal protection.

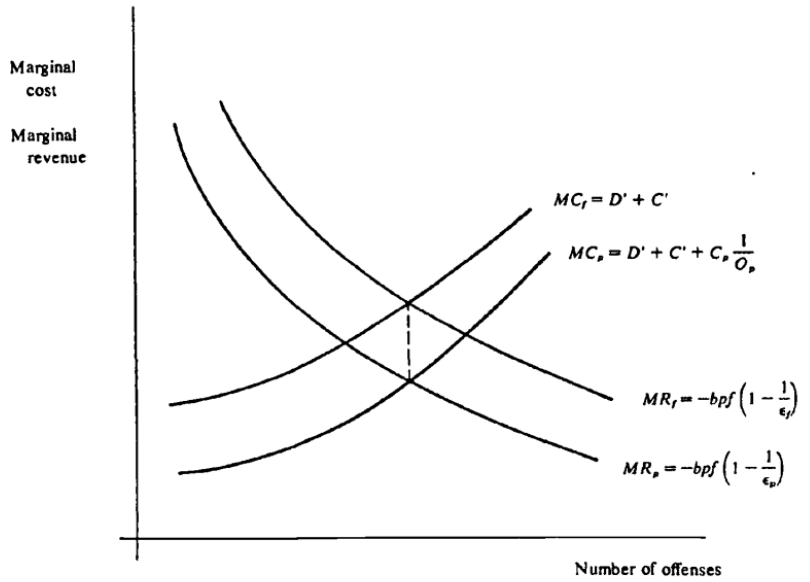


Figure 6: The equilibrium level of crime in a society (Becker 1968)

Becker's models formalize the ways that the state's actions against crime are taken into account in its decision-making through marginal cost and revenue (Figure 6). The marginal cost is dictated by features including the damage caused to society (D) and the costs of inputs such as police and judges (C), and marginal revenue is a function of the negative social loss attributed to crime (bpf). At equilibrium, the cost of an additional crime prevented will be less than the societal gain provided by a marginal increase in public safety.

The policy implications of rational choice theory come from the fact that marginal cost and revenue are not fixed. The revenue side is difficult for a government to impact, being determined by social perceptions that are slow to change. The cost side, however, comes with clear policy instruments, including the probability of conviction upon arrest and the severity of punishments. Changes to these inputs via policy and police strategy can shift the marginal cost curve upward, and thereby decrease the optimal amount of crime in a society.

4 Routine Activity Theory

While rational choice theory has little to do with the coupling of crime and place directly, its importance in the development of quantitative criminology cannot be understated. Rational choice was an instrumental step in bringing about Cohen and Felson's later work on routine activity theory, as well as an influx of economic and statistical modeling into the field of criminology.

Routine activity theory is an environmental, place-based explanation of crime, where the behavioral patterns and intersections of people in time and space influence when and where crimes occur. The theory, in short, claims that a principal driver of criminal activity is the intersection of willing offenders and suitable targets, paired with the absence of capable guardians against crime, in people's movements throughout their everyday lives (Cohen and Felson 1979).

This theory arose from the question of how urban violent crime rates could have increased from 1960 to 1975, while the factors typically attributed to the rise of violent crime had in fact decreased significantly; unemployment was down, minority education rates were up, and the income gap between races had narrowed, all while median income had risen. Despite these improving conditions, the US violent crime rate had more than doubled. Routine activity theory offers a logical framework to explain how this counterintuitive result may have occurred: a sweeping change in routine activities may have created a drastic increase in criminal opportunity.

The routine activities framework understands crime as the intersection of three key factors:

1. the presence of motivated offenders,
2. the availability of suitable targets, and
3. the absence of guardians against an offense.

If any of these factors are absent, an offense cannot occur. Further, even if the total quantities of motivated offenders, suitable targets, and guardians against crime in a macro-place remain static, changes in the routine activity patterns of any of these parties could significantly alter both the locations and quantity of crime by changing the frequency and common locations of the convergence of the three necessary factors.

Table 6. Regression Equations for First Differences in Five Index Crime Rates and Sensitivity Analyses, the United States, 1947–1974

FIRST DIFFERENCE FORM	(1) Nonnegligent Homicide	(2) Forcible Rape	(3) Aggravated Assault	(4) Robbery	(5) Burglary
Constant	−2.3632	−4.8591	−32.0507	−43.8838	−221.2303
t ratio	.3502	5.3679	7.6567	3.4497	3.7229
Proportion 15–24 (t)					
Standardized	.1667	.1425	.4941	.2320	.1952
Unstandardized	3.2190	6.4685	132.1072	116.7742	486.0806
t ratio	1.0695	.7505	3.3147	.9642	.8591
Household Activity Ratio (t)					
Standardized	.7162	.6713	.4377	.4242	.5106
Unstandardized	4.0676	8.9743	34.4658	62.8834	374.4746
t ratio	4.5959	3.5356	2.9364	1.7629	2.2474
Multiple R^2 Adjusted	.6791	.5850	.7442	.3335	.4058
Degrees of Freedom	23	25	25	25	25
Durbin-Watson Value	2.5455	2.3388	2.3446	1.4548	1.7641
1% test	Accept	Accept	Accept	Accept	Accept
5% test	Uncertain	Accept	Accept	Uncertain	Accept
AUTOREGRESSIVE FORM					
Multiple R^2 Adjusted	.9823	.9888	.9961	.9768	.9859
Durbin's h	−1.3751	−.7487	.9709	1.5490	1.1445
−1% test	Accept	Accept	Accept	Accept	Accept
−5% test	Accept	Accept	Accept	Accept	Accept
Griliches Criterion	Accept	Accept	Accept	Accept	Accept
Cochrane-Orcutt Correction, Effect upon Household Activity	Minimal	Minimal	Minimal	Minimal	Minimal
Unemployment Rate as Control, Effect Upon Household Activity	Minimal	Minimal	Minimal	Minimal	Minimal

Figure 7: Impact of Household Activity Ratio on Crime (Cohen and Felson 1979)

This theory is supported by the observation that variations in crime across times of day, days of the week, seasons of the year, and locations within a city all tend to correspond with the tempos of the related non-criminal activities of those times and places. One example of this is how gang-related violent crime tends to correspond in time and place with community leisure patterns, where and when there are large, unpoliced neighborhood parties (Cohen and Felson 1979). A more subtle example, however, is that of the relationship between daytime robberies and the household activity ratio. The household activity ratio is equal to the sum of the number of working married females and non-married adults divided by the total number of US households. This serves as a rough measurement for the portion of houses that are at risk of robbery and general property crime during the day, as the routine activities of these "active" houses' owners typically place them at work or elsewhere during the day. If routine activity theory holds true, then this ratio should have a strong positive correlation with daytime robbery and vandalism rates, as well as on most other crimes, since a high household activity ratio means a higher chance of criminals' and victims' paths intersecting in public during their routine activities. Controlling for the age distribution of the population and its unemployment rate, this is exactly what Cohen and Felson (1979) find, with the household activity ratio being highly

significant for each crime category tested (Figure 7). This ratio shows economic as well as statistical significance, with the magnitude of its effect being larger than that of the population's age distribution for every category of violent crime except for assault (Figure 7).

With compelling results, Cohen and Felson show that routine activity theory can further the tradition established by Becker, modeling criminal decision-making in a spatiotemporal framework that accounts for the placement and movement of individuals throughout a region. Beginning with a simple understanding that instances of crime require an overlap of offenders, victims, and an absence of capable guardians, the routine activities approach both takes us a step closer to the law of concentration of crime at place and offers a useful theoretical framework through which we can analyze the relationships between crime, place, and time.

5 Crime Pattern Theory

Patricia and Paul Brantingham (1993) introduce further theoretical structure to spatiotemporal crime analysis. Brantingham and Brantingham are key figures in the development of crime pattern theory, which, similar to routine activity theory, states that crime is significantly shaped by the intersection of people's routine activities, which themselves are shaped by the physical environments in which these activities take place.

The useful nuance of crime pattern theory is that it defines the types of problem spaces observed as a result of routine activities and rational choice. The first category of place is the *crime generator*. A crime generator is a location that takes people with no criminal intention and converts them into intending criminals. The second type of place is a *crime attractor*, which is a location that draws in individuals specifically intending to commit a crime. Borrowing from the routine activities framework, these types of spaces see high crime rates due to the routine presence of particularly easy targets and a low police and security presence. The third type of location is a *fear generator.* This is a space that leads individuals to believe that they are in danger of being victimized, but in reality there is little data to support the claim that the area is high in crime. Last, there are *crime*

neutral spaces, which see little-to-no criminal activity (Brantingham and Brantingham 1993).

An example of a crime generator could be a bar or pub, where the presence of alcohol makes people more likely to commit crimes, and the presence of drunk bystanders with cash on hand makes for easy targets. A crime attractor could be a shopping mall, where an intending thief knows he can steal something, or a baseball stadium, where distracted crowds make for easy pickpocketing. A fear generator could be any graffiti-covered alley that in reality poses no threat to a passersby, and a crime neutral space could be just about any area that is low in crime.

Crime pattern theory serves as a useful companion to routine activity theory, observing the same general effect of crime occurring at the intersection of routine activity patterns, and adding a structured framework through which we can analyze the physical places in which offenses occur. The pairing of these two theories provides the necessary theoretical toolkit to explain the occurrences of crime in space and time; the crime pattern theorist examines the relationship of place and environment with offense patterns, while the routine activity theorist studies the impact of the people who were critically present or absent (Eck and Weisburd 1995).

6 Graphs and Grids

With the theoretical background in place, empirical questions remain regarding the spatiotemporal models of crime required for understanding a city's criminal hotspots. Bowers et al. (2005) explore the fundamental question of how exactly crime should be modeled, lending credence to Weisburd's decision to model crime at the street-segment level by showing that crime forecasting models are significantly more effective when using street segments as their unit of analysis, as opposed to the popular grid-based alternative.

Criminal forecasting models have traditionally employed a planar approach, overlaying n-by-n blocks on a map and modeling hotspots based on criminal occurrences in these geographic squares. While this has shown some success, an alternative approach, using a network of street segments as its unit of analysis, proves to be a

superior model (Figure 8). The superiority of the street segment approach holds for all levels of coverage, where coverage is defined as the proportion of the total grid area in the case of the grid model, and total network length in the street segment case. This essentially serves as a measure for how geographically precise the model needs to be in its predictions to be considered successful, where a higher coverage value has a looser definition of a prediction being "close enough" to an actual crime's location in order to be considered an accurate prediction (Rosser et al. 2016).

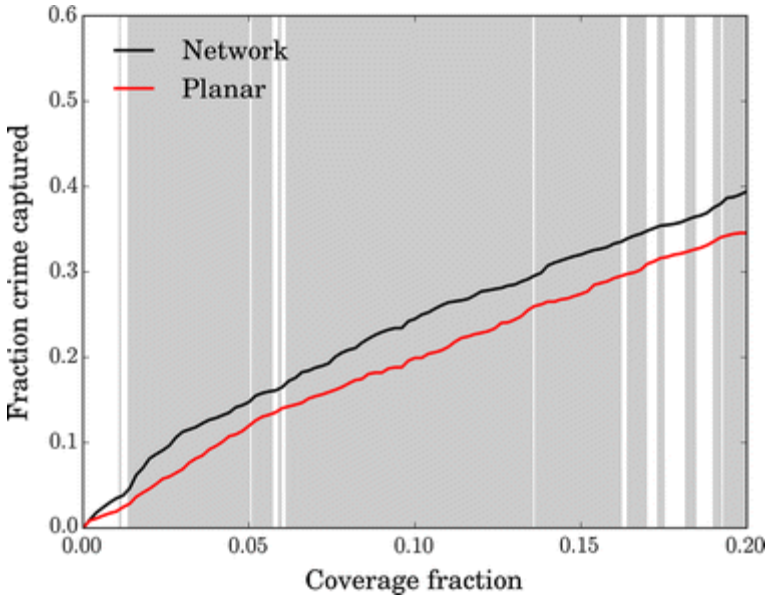


Figure 8: Accuracy of Grid ("Planar") vs. Street-Segment ("Network") Approaches to Prediction (Rosser et al. 2016)

The intuitive reason behind the comparative success of the street segment-based approach is that a network of street segments better reflects the geographic reality of the space it is modeling. While the grid-based approach to modeling crime concentration treats all n-by-n blocks as having equal chances of facilitating the intersection of motivated offenders, suitable targets, and the absence of capable guardians, the street segment approach only considers actual streets, leading to models with more uniform units of analysis in terms of public usability than those employing arbitrary grids. The relative superiority

of the street segment model over the grid model can qualitatively be seen in Figure 9, where hotspots are more precisely defined as street segments than as sections of a grid (Rosser et al. 2016).

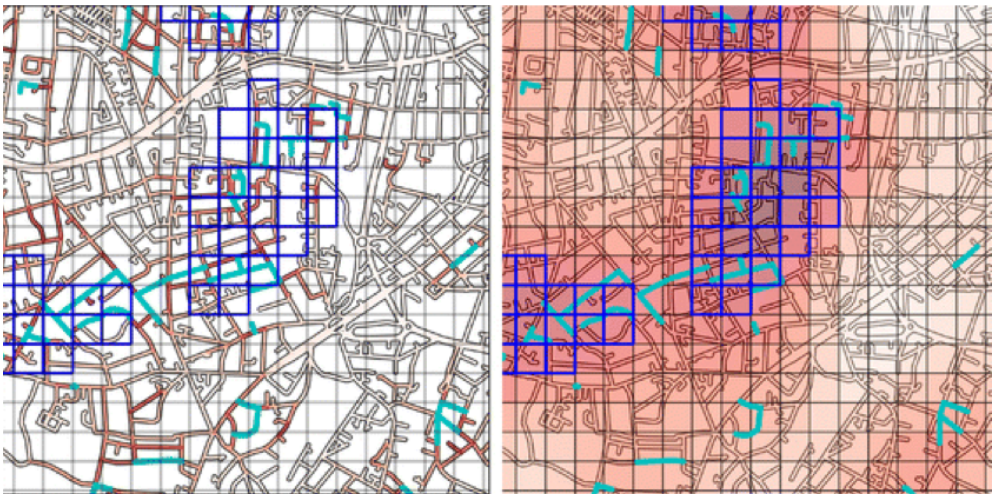


Figure 9: London's Criminal Hot Spots via Street Segment vs. Grid-Based Model (Rosser et al. 2016)

In addition to clustering at place, crime also shows a tendency to repeat in predictable intervals across time. If a house becomes a victim of burglary, for example, it is at an elevated risk for a repeat incident to occur during a short time interval after (Bowers and Johnson 2005). The increased likelihood of repeat offenses suggests an event dependency, where the conditional probability of an additional criminal incident occurring within a fixed time interval consistently increases with each additional crime that occurs (Sherman et al. 1989) (Figure 10).

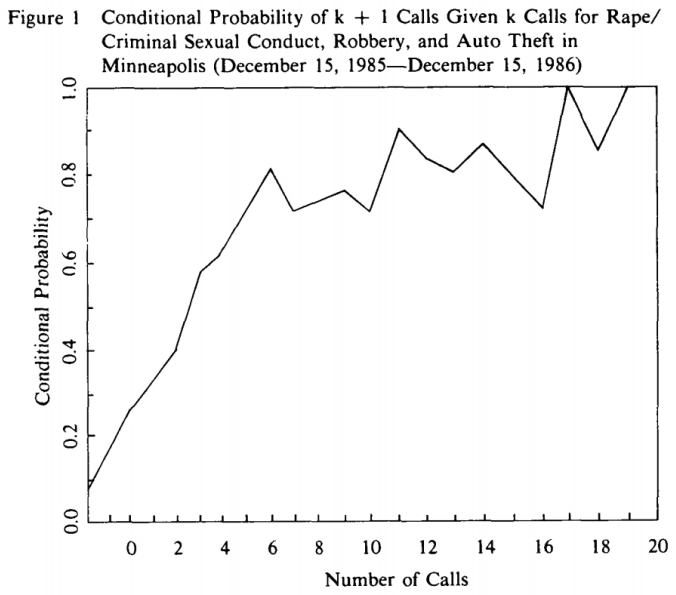


Figure 10: Conditional Probability of an Additional Incident Given Past Incident Count (Sherman et al. 1989)

This event dependency lends further support to the claim that crime clusters in the micro-place, suggesting that crime begets more crime at the street segment level within the same year, and also that recent-historical hotspots are consistently strong predictors of future crime (Sherman et al. 1989). This tendency to cluster in time as well as place is so strong, in fact, that the predictive models implemented in by the Los Angeles and Atlanta police departments in 2014 were Epidemic Type Aftershock Sequence models, borrowing from seismological models of the ways that aftershocks follow an earthquake in place and time (Mohler 2014).

7 Putting Crime in its Place

With significant backing to the theory that crime clusters in place and time, it is important that we continue to push this recently-developing body of theory, testing its generalization to unseen data points and asking the causal question of what exactly causes a criminal hotspot to appear or disappear. This, in short, is the subject of my thesis.

The econometric analysis of crime in micro-places that follows is first made possible by the recent influx of open government data. Since President Obama took office in 2009, the the US government has undertaken various initiatives for releasing federal data for research purposes, including the creation of data.gov, the Open Government Initiative, and the Police Data Initiative. Many US cities and states followed suit, leading to what has become hundreds of gigabytes of freely available police incident report data from most major US cities, including the time, location, and category of each incident that has occurred, along with other pieces of city-specific information.

With the aforementioned theoretical frameworks and empirical results as a starting point, the foundation is in place for understanding how crime is generated in and as a function of place. I will begin by replicating Weisburd's analysis from *The Law of Crime Concentration and the Criminology of Place* (2015), testing his theory of crime concentration in new locales. This first stage of analysis will focus on large cities, namely Seattle, Chicago, Los Angeles, Portland (OR), San Francisco, Philadelphia, Dallas, Washington DC, and Cincinnati. Two of these cities, Cincinnati and Seattle, overlap with Weisburd's initial sample, and the rest are yet to have had their crime concentration levels analyzed at the street segment level. I will dissect these concentration levels both comparatively against one another and over time in order to understand the strength of the law of concentration of crime at place and its stability over time. This section of analysis will attempt to answer the questions of how closely crime couples with place, and how long hotspots stay hot.

Next, I will analyze the locations in which crime concentrates through the routine activities and crime pattern frameworks, attempting to answer the causal question of why crime concentrates where it does. Focusing on the city of Chicago, the second section of this study uses geospatial data on demographics, socioeconomic status, street type, and the locations of various types of facilities such as bars, restaurants, subway stations, and retirement homes in order to better understand the ways in which the local environments of micro-places interact with their levels of crime risk. Employing both logistic and ordinary least squares regression models, the significance, direction,

and relative magnitude of these features' coefficients can be used to improve public safety.

8 VIII. Data and Methods

8.1 Crime Data

This study makes use of publicly available incident report data from the open data portals of the cities of Seattle, Chicago, Los Angeles, Portland (OR), San Francisco, Philadelphia, Dallas, Washington DC, and Cincinnati. These cities were chosen for a variety of reasons:

First, they represent a diverse cross-section of America's major cities; they vary significantly in population, racial composition, crime rate, poverty level, and street layout. With one of this paper's goals being to study the law of concentration of crime at place across a larger sample of cities than has been offered to date, a diverse sample of cities will provide the basis for a stronger claim in support of this theory, should the results be positive.

Second, the cities in this sample all provide crime data at the street segment level. While many cities outside this sample offer incident report data, these are the ones that lend themselves most readily to analysis at the street segment level. New York and Boston, for example, only offer crime locations in the form of latitude and longitude coordinates, which are computationally challenging to reverse-geocode into addresses and then street segments. In this sense, the sample I choose is also one of convenience.

ID	CaseNumber	Date	Block	IUCR	PrimaryType	Description	LocationDescription	Arrest	Domestic	Beat
2153478	HH395183	05/25/2002 08:40:00 AM	034XX W 55TH ST	0460	BATTERY	SIMPLE	RESIDENCE	false	true	822
2153479	HH401581	05/27/2002 08:00:00 AM	048XX S CICERO AVE	0610	BURGLARY	FORCIBLE ENTRY	OTHER	false	false	814

District	Ward	CommunityArea	FBICode	XCoordinate	YCoordinate	Year	UpdatedOn	Latitude	Longitude	Location
8	14	63	08B	1154500	1867935	2002	04/15/2016 08:55:02 AM	41.79343	-87.7090	(41.793431003, -87.708999558)
8	23	56	05	1145172	1872210	2002	04/15/2016 08:55:02 AM	41.80534	-87.7431	(41.805343087, -87.743097373)

Figure 11: Snapshot of Individual Crime Data (City of Chicago, 2016)

Each data set contained the block-level address of each crime (i.e. 21XX BLOCK COMMONWEALTH AVE), a description of the crime that occurred, and a latitude-longitude coordinate pair for the

incident's location. They also tended to include locale-specific encodings such as police district (Figure 11). The main area where the different cities' data sets differed was the categories into which they grouped crimes. The city of Houston's data, for example, contains only low-level crime categories such as "ASSAULT (AGG) -AGAINST SECURITY OFF (AGG), ASSAULT (AGG) -DEADLY WEAPON, ASSAULT (AGG) -DISCH FIREARM OCC BLDG/HOUSE/VEH (AGG)", whereas the city of San Francisco's reports contained only high-level categories, grouping all assaults into a single "ASSAULT" category (Figure 12). In order to compare crime concentration levels for a specific subset of crimes across all cities, it was necessary to re-encode some cities' crime categories so that all cities' data were comparable.

	Chicago	Seattle	Los Angeles	Portland	San Francisco	Philadelphia	Dallas	DC	Cincinnati
Assault Encoding(s)	ASSAULT	ASSAULT	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER, ASSAULT WITH DEADLY WEAPON, ASSAULT WITH DEADLY WEAPON, BATTERY - SIMPLE ASSAULT, CHILD ABUSE (PHYSICAL) - AGGRAVATED ASSAULT, CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT, SPOUSAL(COHAB) ABUSE - SIMPLE ASSAULT, SPOUSAL (COHAB) ABUSE - AGGRAVATED ASSAULT, OTHER ASSAULT	Aggravated Assault, ASSAULT W/DANGEROUS WEAPON, Assault, Simple	ASSAULT	Aggravated Assault Firearm, Aggravated Assault No Firearm, Other Assaults	ASSAULT, AGG ASSAULT - NFV	ASSAULT W/DANGEROUS WEAPON-GUN, ASSAULT W/DANGEROUS WEAPON-KNIFE, ASSAULT W/DANGEROUS WEAPON-OTHER	Assault -(Aggravated Assault), ASSAULT -(Aggravated Assault), Assault -(Simple Assault), ASSAULT -(Simple Assault), Domestic Violence -(Aggravated Assault), DOMESTIC VIOLENCE -(Aggravated Assault), Domestic Violence -(Simple Assault), DOMESTIC VIOLENCE -(Simple Assault), ENDANGERING CHILDREN -(Simple Assault), Felonious Assault -(Aggravated Assault), FELONIOUS ASSAULT -(Aggravated Assault), Felonious Assault -(Simple Assault)

Figure 12: Assault Encodings by City

The observations in these datasets represent police incident reports. These reports represent events that are more severe than a 911 call, and but are often less severe than an arrest. Any time an officer arrives on a scene and finds the event sufficiently important to document, the report is digitized and then released by the city as open data. For this reason, some, but not all incidents in this data represent arrests.

For cross-city analysis, two sets of crimes are selected for investigation. First, testing Weisburd's law of concentration of crime at place, it is necessary to analyze only the crimes whose categories

were included in the original paper defining this law. This means including burglary, property destruction, assault, homicide, robbery, graffiti, abandoned vehicles, drugs, prostitution, drunk driving, and hit and run incidents (Weisburd 2015). Second, the violent crimes are examined in isolation. These include assault, battery, robbery, sexual assault, homicide, and domestic violence. Due to the differing encoding systems across cities, sub-setting the data to include only these crimes required some manual searching in order to find each city's sometimes-several names for each category. For the most part, all cities' data contains the desired encodings. The one noteworthy exception is that of abandoned vehicles, which Weisburd includes in his analysis but were not available for the majority of cities in this sample. For the regressions run in the second part of this study, only the violent crimes are selected for analysis, as these are the incidents that pose the greatest threat to society and are of the highest interest to law enforcement.

The data is originally provided at the individual crime level. An observation, for example, could represent a homicide. This observation would tell which street segment the homicide occurred on, the date of the incident, its latitude and longitude, and the time of police response. To analyze the amount of crime happening at specific street segments, I then take the count of crimes happening on each street segment for each year and collapse the data so that each observation represents all crime on a a street segment, rather than an isolated incident. This operation is performed to both the general crime and the violent crime-only subsets of the original data.

A second type of data I use is each city's street centerline file. A centerline file is a shapefile that includes the polylines representative of a street network. These include latitude-longitude coordinates, street segment IDs, and metadata such as street type and street segment length. These files were converted to GeoJSON format using the free program QGIS, and then converted to a usable tabular format in the statistical programming language R. The centerline files, like the rest of this data, were provided by the cities themselves via their open data portals.

In the end, I focus on the city of Chicago for exploring potential causal relationships between facilities, socioeconomic factors, and crime. I choose Chicago for a variety of reasons. First, Chicago is among America's best open data cities, meaning that its open data portal contains a large supply of machine-readable data sets that can be brought into this analysis. Second, Chicago is of a particular degree of interest in studies of crime due to its frequent presence in the news as a city that is high in gun violence and other violent crime. Third, the Chicago Police Department's incident report data is of a higher quality than most of the other cities studied, containing minimal incomplete observations in the features being studied and having clean, interpretable encodings for its crime categories.

8.2 Units of Analysis

The unit of analysis in this study is the street segment. A street segment is defined as both sides of a street between two intersections. All but one of the cities in this sample had average street segment lengths between 354.3 and 465.7 feet, with Portland being the only outlier at 151.6 feet. The number of street segments varies by city as well, with the smallest having 13,978 street segments and the largest having 87,042.

Street segments were chosen as the preferred unit of analysis for a variety of reasons. First and foremost, the street segment holds an important place in social organization, being physically bounded from other segments and home to a common pattern of routine activities. This spatial unit serves as a psychological behavior setting in that it carries with it associated role obligations such as neighborliness, and norms which govern acceptable conduct (Taylor 1997). For this reason, paired with its small size, the street segment tends to be homogeneous in the routine activities it plays host to. There is very little overlap, for example, between the streets playing host to the activities of people's household, commercial, and night lives; the same can not be said for the larger units of analysis typically used in criminology. For this reason, understanding crime at the street segment level holds a degree of social significance that cannot be achieved with a broader, less socially cohesive unit such as zip code or police district.

Second, the street segment is the smallest geographic unit at which we can accurately measure crime. The smaller the unit of analysis is, the less we need to worry about latent features explaining the levels of crime concentration that we observe. An attempt to be more micro than street segments, on the other hand, using addresses or coordinates, would suffer from widespread inaccuracies in police reporting. Taking latitude-longitude encoded data that has been generalized to the street segment level is as micro-scale as an analysis can currently be while claiming accuracy in its underlying unit of analysis. Communities are heterogeneous entities that are challenging to define socially and geographically, so the street segment's size and social homogeneity makes it particularly well suited for crime analysis.

Last, Rosser et al. (2016) argue that larger, area-based units of analysis such as census geographies, neighborhoods, and zip codes fail to accurately represent the amount of human-usable space they contain. Where human-usable space on any two street segments can be compared against one another in feet or meters, the same can not be said of two zip codes, as one may have significantly more populated space than the other. Being micro and street-based rather than area-based in the unit of analysis is a crucial step to avoiding the problems of spatial heterogeneity encountered throughout the environmental criminology literature.

8.3 Measure of Distance

Part of this analysis depends on variables measuring the quantity of specific types of facilities within set distances of individual street segments; the number of bars, for example, within 200 feet of the 1900 block of Beacon Street. These features are created using a co-centric circles approach. For a given street segment, I first define a central point. Next, I count the number of occurrences of a facility type within the first distance threshold. Last, I take the number of facilities lying between the first and second distance threshold. The end result is two non-overlapping measures of facility counts which can be used to estimate the causal impact of proximity to certain types of facilities on crime at the street segment level. To give these variables economic significance, the distance measures chosen are equal to 200 and 600 feet, or 0.5 and 1.5 times the average street segment length in Chicago,

the city that the causal analysis will focus on (Figure 13). These distance thresholds will capture roughly the number of facilities within one and two blocks of the street segment of interest.

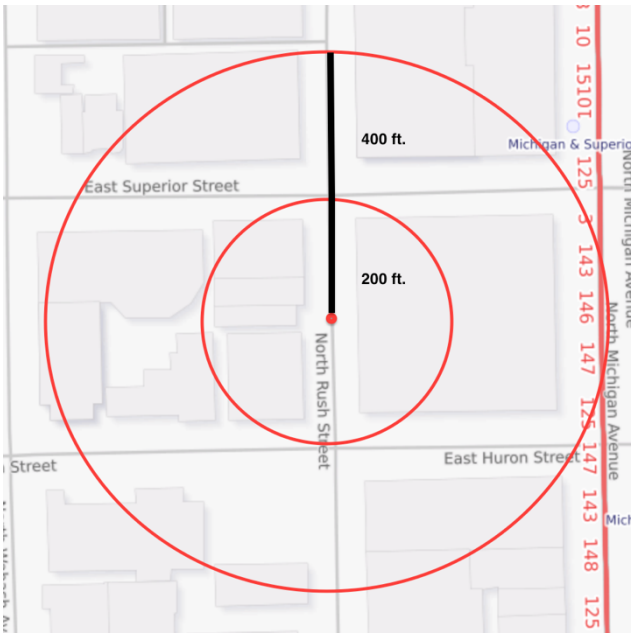


Figure 13: Distance Thresholds for Spatial Feature Generation

Distances are calculated using the haversine formula for great-circle distance. Using the diameter of the Earth and two pairs of latitude-longitude coordinates, this formula computes the distance between two locations while accounting for the curvature of the Earth. Using the R package *geosphere* (Hijmans 2016), I compute a distance matrix whose rows represent street segment centroids and columns represent facilities. Each entry in this matrix, then, is the haversine distance, in meters, between street segment i and facility j. The entries are then converted into feet, since this is the unit that the street segments are measured in, and the desired features are generated by counting the numbers of facilities that lie either between 0 and 200 or 200 and 600 feet of each street segment centroid. Haversine distance is the standard method for measuring direct distance between latitude-longitude coordinate pairs, and is analogous to the Euclidean distance that is used for points on flat coordinate planes.

It is worth noting that this is an imperfect measure. While direct distance, either Euclidean or haversine, has historically been the

default method in similar studies, the ideal method for counting the numbers of facilities within a set distance of a street segment would be to do so using street network distance, which more accurately captures the distance a person would travel between points on a street grid. For reasons of computational cost and software limitations, however, I decided to use direct distance via the haversine formula. For a thorough discussion of the comparative merits of direct and street-network distance, see Levine (2013).

One final note on the retrieval of distance-based features is that the computational and memory costs of building distance matrices on large data sets are quite high. In the context of counting the numbers of bus stops within 200 feet of each of Chicago's street segments, this means that one must compute a matrix with 52,000 rows and 11,000 columns, where each of the 572,000,000 entries is the distance between street segment i and bus stop j. This is seldom feasible and always slow on a personal computer, so it is recommended that these matrices be broken into smaller subsets of street segments and computed on a server. The size of these matrices also lends support to the use of a simple distance measurement such as haversine, since any algorithm with higher time complexity might make these features prohibitively slow to compute.

8.4 Facilities

This thesis in part measures the relationship between crime and the local environments in which it occurs. The first class of variable this focuses on is facilities. A facility, through a routine activities interpretation, is any establishment with a physical building that serves as the destination of some form of activity, commercial or otherwise. Data on facilities comes from the City of Chicago's publicly available datasets on business licenses, public parks, senior centers, grocery stores, drug treatment centers, and public schools. These facilities are converted into features by taking the counts of each facility type within a set threshold of the center point of each street segment using the already-discussed method of computing matrices of haversine distances between each street segment and each facility. The features are defined as follows:

Schools: K-12 public schools (n=672)
Drug Centers: licensed substance abuse treatment centers (n=204)
Grocery Stores: wholesale and retail grocery stores from the Chicago business license dataset (n=507)
Senior Centers: live-in retirement homes (n=21)
Restaurants: restaurants, delis, and cafes in the Chicago business license data set (n=14,473)
Bars: bars, taverns, and restaurants that become 21+ late at night (n=931)
Liquor Stores: liquor stores and retail locations authorized to sell unopened liquor (n=1,195)
Daycare Centers: licensed daycares and children's activities facilities (n=867)
Animal Care Centers: licensed animal care facilities, including veterinary clinics, grooming centers, guard dog services, and the Humane Society (n=358)
Gas Stations: self explanatory (n=434)
Pawn Shops: licensed pawnbrokers (n=50)
Arts Venues: performing arts venues, including concert halls and live theaters (n=148)
Businesses: businesses with either limited or regulated business licenses (n=28,627)

One thing to note is that there is slight spatial overlap among some of these features. Certain restaurants, for example, become bars after a certain hour and are legally licensed under both categories. Similarly, certain grocery stores are also licensed liquor retailers. This introduces a slight concern that the variables may be highly correlated, leading to a multicollinearity problem. The vast majority of facilities, however, represent independent single-facility locations, and correlations between coefficients are discussed in detail in Section 12.1.

8.5 Spatial Features

A second class of feature created from this data is non-facility spatial features. This class of feature includes any feature of a local

environment that does not fit the definition of a facility. These are features of street segments that characterize how the space is used.

Subway Stations: stations from the CTA's L system (all lines, n=110)
Bus Stops: public bus stops provided by the CTA (n=11,593)
Distance to City Center: log-haversine distance to Willis Tower
Parks: public parks managed by the Chicago Parks District (n=577)
Length: street segment length in feet, taken from street centerline file
Graffiti: closed 311 service requests for graffiti removal within one year of the hotspot-year being tested (n=758,612)

Some of these features are noticeably large in number. Chicago has a particularly expansive public transportation system, and has also recorded over 700,000 graffiti removal requests since it began keeping track. This class of feature is also slightly different than the facility-based features in that it includes two features, distance to city center and street segment length, which are continuous rather than count variables.

8.6 Socioeconomic Features

The last class of feature used in the models of crime is socioeconomic. The socioeconomic data used in this thesis are provided by the City of Chicago at the community area level, and also by the US Census Bureau at the census tract level. The city has 77 community areas, serving as the primary geographic unit for urban planning, as well as 866 census tracts. The socioeconomic variables considered are:

Percent of Housing Crowded: percent of housing units with more than one occupant per room, provided at the community area level
Per Capita Income: calculated as the sum of tract-level aggregate incomes divided by total population, provided at the community area level

Percent of Households Below Poverty: calculated using the federal poverty level, provided at the community area level

Age Quantiles: age composition of the population at the census tract level, provided in four-year quantiles (e.g. percent aged 15-19, 20-24, and so on)

Percent Aged 25+ without High School Diploma: provided at the community area level

Hardship Index: an index of socioeconomic hardship, calculated by standardizing the above-mentioned community area-level variables and taking their average (Nathan and Adams 1989)

The community area-level features were calculated for the time period of 2008 – 2012, which is the most recent period for which the city has provided this data. The age quantiles come from the most recent census (U.S. Census Bureau 2010).

8.7 Dependent Variables

In order to frame the problem in two different ways and better understand the underlying causes of criminal hotspots, both binary and discrete dependent variables will be used for modeling. The binary dependent variable, to be used in logistic regression models, will come from Weisburd's definition of a hotspot. The variable will equal one if the street segment is among those accounting for 25 percent of the city's total violent crime, and will equal zero otherwise. The discrete variable will simply be the number of crimes that have occurred on a street segment in the year being tested.

segment_id	community_area	Block	Latitude	Longitude	subway_stations	bus_stops	parks	graffiti	Length
2463	25	002XX S LOTUS AVE	41.87721	-87.76172	0	6	0	10	826.9784
29030	65	037XX W 62 ST	41.78070	-87.71581	0	4	0	217	669.2291
9614	73	011XX W 100TH ST	41.71230	-87.65057	0	4	0	2	329.9007

log_dist_city_center	bars	schools	grocery_stores	senior_centers	businesses	parking_garages	liquor_stores	childcare
9.594584	0	1	0	0	10	0	1	0
9.167888	1	0	0	0	5	0	1	0
7.797353	0	0	0	0	0	0	0	1

animal_care	gas_stations	drug_treatment_centers	pawn	arts_venues	restaurants	pct_housing_crowded	pct_houses_poverty
0	0	0	0	0	5	6.3	28.6
0	0	0	0	0	3	5.8	14.9
0	0	0	0	0	1	1.1	16.9

pct_16plus_unemployed	pct_25plus_no_diploma	pct_under16_over64	pc_income	hardship_index	crime_count	is_hotspot
22.6	24.4	37.9	15957	73	14	1
9.6	33.6	39.6	16907	56	0	0
20.8	13.7	42.6	19713	48	0	0

Figure 14: Snapshot of Spatial, Crime, Facility, and Socioeconomic Data at Street Segment Level

9 Summary Statistics

As is mentioned in an earlier section, this sample of cities was chosen due to data availability and compatibility with this study's unit of analysis. It would be advantageous, however, if this sample also happened to represent a diverse subset of US cities across the major factors attributed to macro-scale crime rates. Table 1 demonstrates that this is the case.

	Chicago	Seattle	Los Angeles	Portland	San Francisco	Philadelphia	Dallas	DC	Cincinnati
Population (in thousands)	2,179	652	3,884	609	837	1,553	1,258	658	297
Crime per 1k residents	39.22	63.76	30.52	43.26	70.24	41.80	42.16	59.45	65.41
Violent crime per 1k residents	9.08	6.02	6.45	5.26	7.85	10.30	6.98	12.69	9.26
Percent black	31.9	7.3	9.2	6.1	5.7	43.0	24.6	49.6	43.5
Percent below poverty line	22.7	14.0	22.4	18.3	13.3	26.7	24.1	18.2	30.9
Percent age 18 - 24	10.9	11.3	11.3	9.1	8.5	12.4	10.4	13.1	14.3

Table 1: Summary Statistics on Sample Cities (data: Neighborhood Scout)

A few things stick out upon viewing the summary statistics for this sample. First, while these would all be classified as large cities, there is significant variation among them in population. The population of Los Angeles is roughly an order of magnitude larger than that of Cincinnati, with the other cities' populations being distributed between those two. The sample is well stratified in terms of crime rate

as well, with the highest-crime cities seeing twice as much crime as the safest. Racial composition varies between these cities as well, with African American populations being as low as 5.7 percent of the population in San Francisco, and as high as 49.6 percent in Washington, D.C.. The cities' young adult populations, measured by the percent of the population aged 18 to 24, varies from being 8.5 percent in San Francisco to 14.3 percent in Cincinnati.

While there are other dimensions along which these cities could be compared, those shown in Table 1 represent an important subset in the criminology literature. Population is among the most important ways a city is characterized, serving as a proxy for city size and activity level. Per-resident crime levels matter for obvious reasons, as one might hypothesize that a higher crime rate would affect concentration levels. It is also important to have an understanding of socioeconomic and demographic indicators in the sample, since the relationships between poverty, race, and crime are widely debated in sociology and economics (Buonanno 2006). A well-stratified sample across these factors carries the benefit of a strengthened claim of crime concentration if the concentration levels are consistent across cities, and will present potential directions for further research if the results are negative or inconclusive.

	Chicago	Seattle	Los Angeles	Portland	San Francisco	Philadelphia	Dallas	DC	Cincinnati
Date Range	2001 - 2016	2008 - 2016	2012 - 2015	2004 - 2014	2003 - 2016	2006 - 2016	2015 - 2016	2011 - 2016	2012 - 2015
Number of Street Segments	52887	23895	84018	37355	13978	41020	97932	13462	10061
Average Segment Length (ft.)	415.98	429.42	465.74	151.66	NA	354.37	NA	NA	445*
Avg. 50% Concentration Level (%)	6.55	5.28	5.48	5.11	4.05	8.09	4.76	5.75	5.55
Avg. 25% Concentration Level (%)	1.99	1.25	1.65	1.26	0.79	2.71	0.99	1.7	1.69
Avg. Violent 50% Concentration Level (%)	5.65	0.95	3.21	1.24	1.75	6.27	0.76	4.44	3.3
Avg. Violent 25% Concentration Level (%)	1.78	0.24	1.02	0.32	0.44	1.98	0.25	1.35	1.02

*Taken from Weisburd (2015)

Table 2: Statistics on Crime Concentration and Street Segments

Table 2 presents each city's crime concentration level, along with summary statistics on the data used to generate these values. The data sets vary significantly in the duration of years for which they provide crime data. Dallas provides only two years of data, while Chicago tops the list by providing 16. The duration of the period studied is not a major factor in gathering concentration levels, but longer time periods will be useful in studying the longevity of hotspots in a later section of this analysis. Information on the street segments themselves is provided as well, because this represents the nature of the underlying unit of analysis. The cities vary significantly in number of street segments, but are more or less similar in average street segment length, with the only outlier being Portland with a well-below-average mean street segment length of 151.6 feet. I was not able to obtain reliable street segment length numbers for the cities of Dallas and Washington, D.C. Most importantly, the statistics for crime concentration are presented at the bottom of Table 2. These numbers include concentration levels for general crime, in accordance with Weisburd's crime categories discussed in Section 8.1, and violent crime, discussed in the same section. These statistics represent the percentage of each city's street segments required in order to explain a set proportion of its total crime, reported at the 50 and 25 percent total crime concentration levels. These concentration levels, and the method used to obtain them, are discussed in detail in Section 10.

It is a helpful visual aid to see how each city's crime is distributed at the street segment level (Figure 15). It is clear from this figure that crime follows a negative exponential distribution, with the majority of crime taking place in a small number of street segments. In measuring the percentage of street segments required in order to explain a set percentage of a city's crime, this is the effect that is implicitly being measured. The distributions are standardized to a 0-1 scale on each axis for comparability. Here a one on the y-axis represents the city's highest-crime street segment, and the x-axis simply represents the ranking of segments by crime level, where the segments near x=0 are the highest in crime and those near x=1 are the lowest.

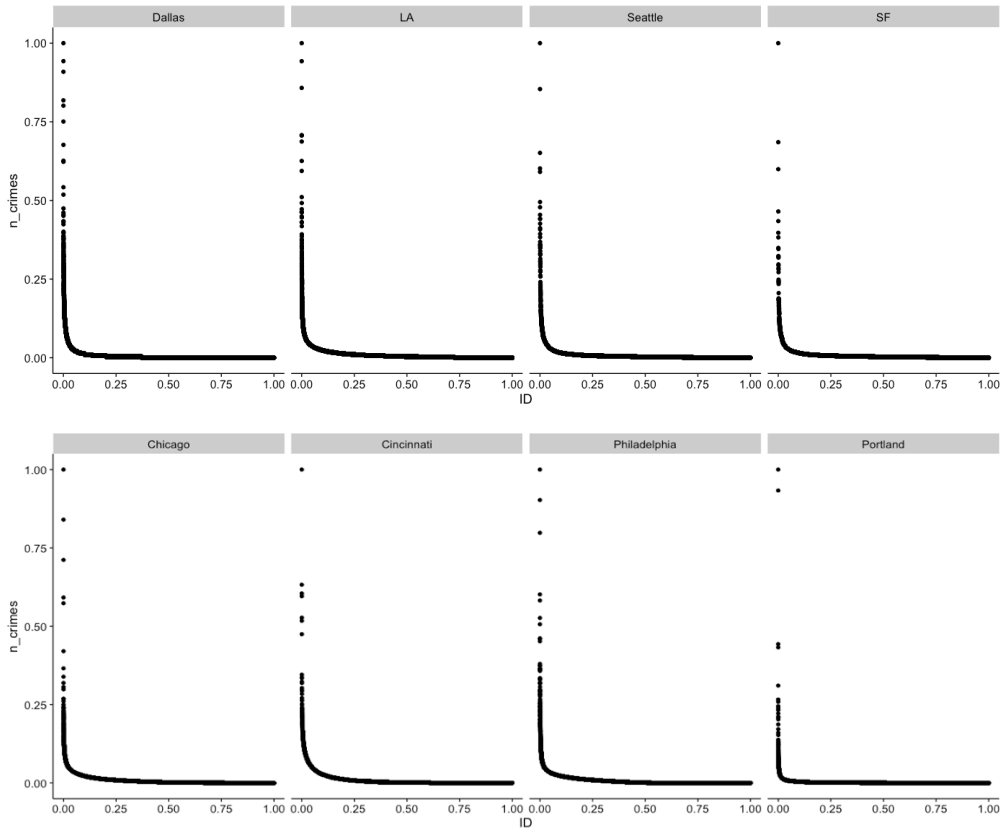


Figure 15: Distribution of Crime Count at the Street Segment Level Across Cities

The cities in this study follow near-identical distributions of crime across street segments, with slight differences visible in the near-vertical left tail of the plots. The following section will assign a number to this similarity in discussing the concentration levels at set cumulative proportions of crime.

10 Concentration and Stability of Crime in Micro-Places

The law of concentration of crime at place, stating a city's concentration of crime will fall within a narrow bandwidth of percentages for a defined cumulative proportion of crime, is originally tested by Weisburd (2015) in a sample of five large and three small cities. Focusing only on large cities due to data availability and sample sizes, I test this relationship in an expanded sample of nine cities. This expanded sample contains two cities from Weisburd's 2015 analysis –

Seattle and Cincinnati – in order to verify that my method is able to achieve a sound replication.

The method for arriving at these numbers is simple. First, the data is filtered by crime type so that only the categories we are interested in remain. Second, I create a table of the city's street segments, sorted by the number of crimes occurring on each segment. Last, I find the number of crimes that represents the cumulative portion of the city's crime we are looking to explain by multiplying the total crime count by that percentage, and then find the number of street segments needed in order to explain this percentage of the city's crime by taking a cumulative sum over the sorted table. This number is divided by the total number of street segments in the city so that it represents the percentage of street segments required in order to explain the set proportion of crime, rather than the raw number of segments. I repeat this process for each year in the data, and the mean of the concentration levels is the value which is reported in Table 2 and Figure 16.

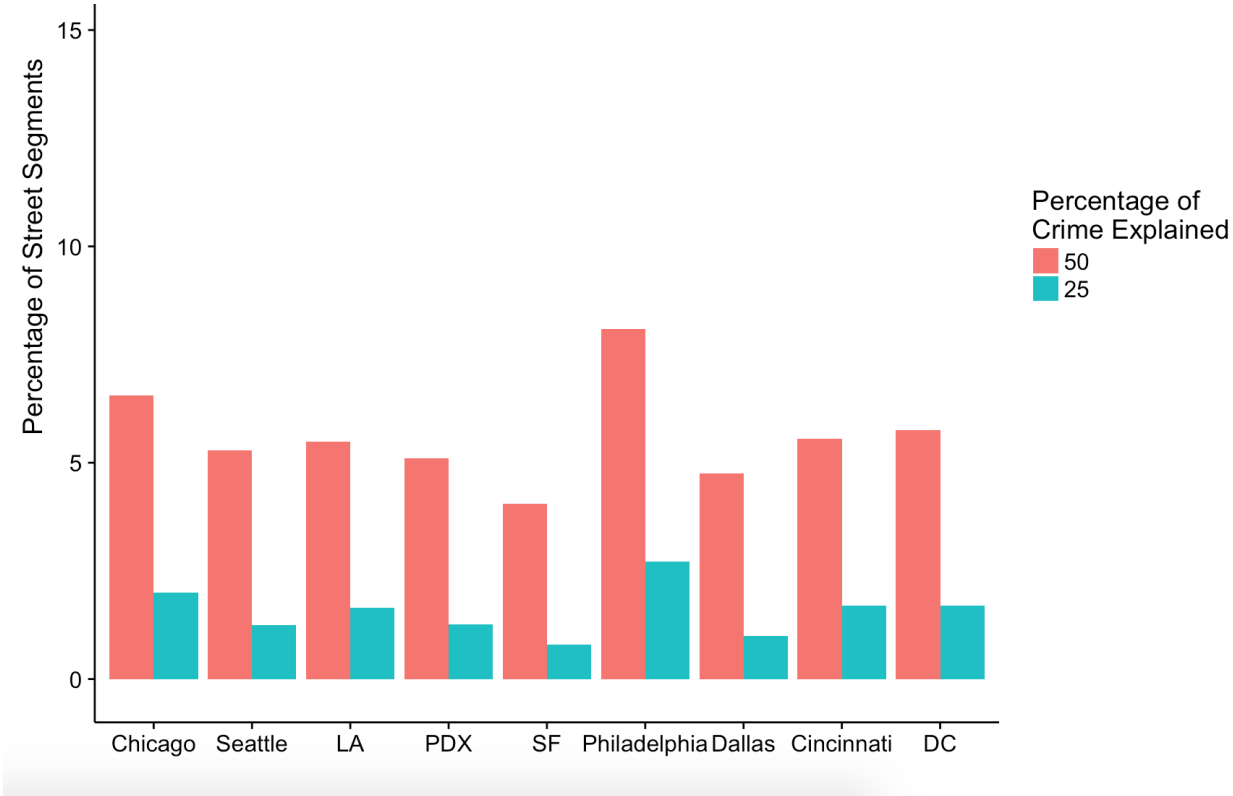


Figure 16: Crime Concentration Levels across Cities

Looking at the concentration level results, it is clear that crime concentrates within a small number of street segments in major US cities. Despite this being a well-stratified sample, these results show that 50 percent of crime concentrates at between 4.05 (San Francisco) and 8.09 (Philadelphia) percent of a city's street segments, with the mean 50 percent concentration level across cities being 5.62 percent. As is expected, crime is more than twice as concentrated at the 25 percent level, with this percentage of cities' crime coming from between 0.79 (San Francisco) and 2.71 (Philadelphia) percent. This sample shows that, on average, 25 percent of a major city's crime comes from only 1.56 percent of its street segments, and 50 percent of its crime can be explained by just 5.62 percent.

These findings are generally in line with those of Weisburd (2015), with the mean 50 and 25 percent concentration levels across cities each being slightly higher that those from the original study. My sample has mean 25 and 50 percent concentrations level of 1.56 and 5.62 percent, where Weisburd shows an average of 1.24 and 5.28 percent concentration at these same levels across similar cities.

These findings, qualitatively less stable than those of the original study, suggest that the narrow bandwidth of percentages that concentration levels fall within may be slightly larger than was initially hypothesized. Crime is certainly highly concentrated in this extended sample, but with a higher variance in concentration levels than the original sample of five cities suggests. It is also possible, however, that Philadelphia, with its lower levels of crime concentration, is an exception to a broader rule. If we are to ignore this observation, this nine-city sample looks remarkably similar to the original five-city sample. This, however, would only be speculation, and is an indication that a still-larger sample of cities may be necessary in order to understand the distribution of concentration levels across large cities. Variance aside, the means of the two samples are quite similar, with this extended sample further confirming the high concentration level of crime in major cities.

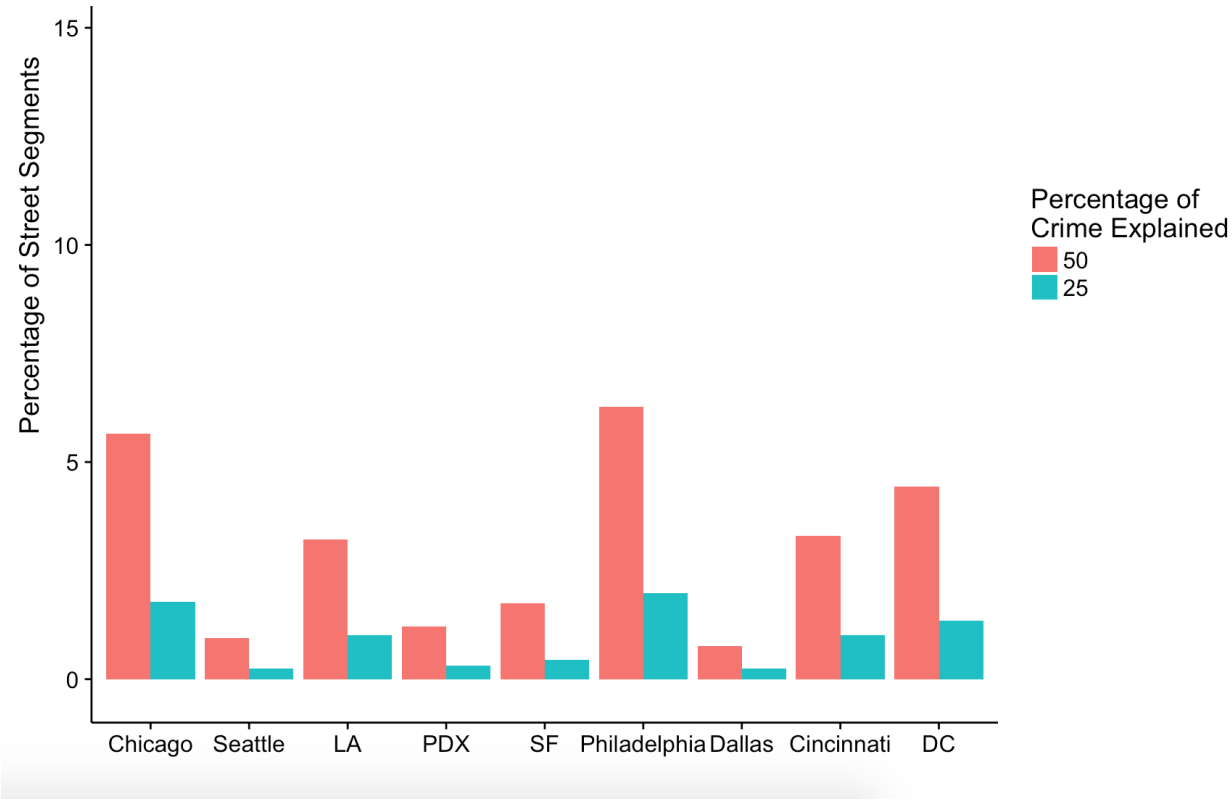
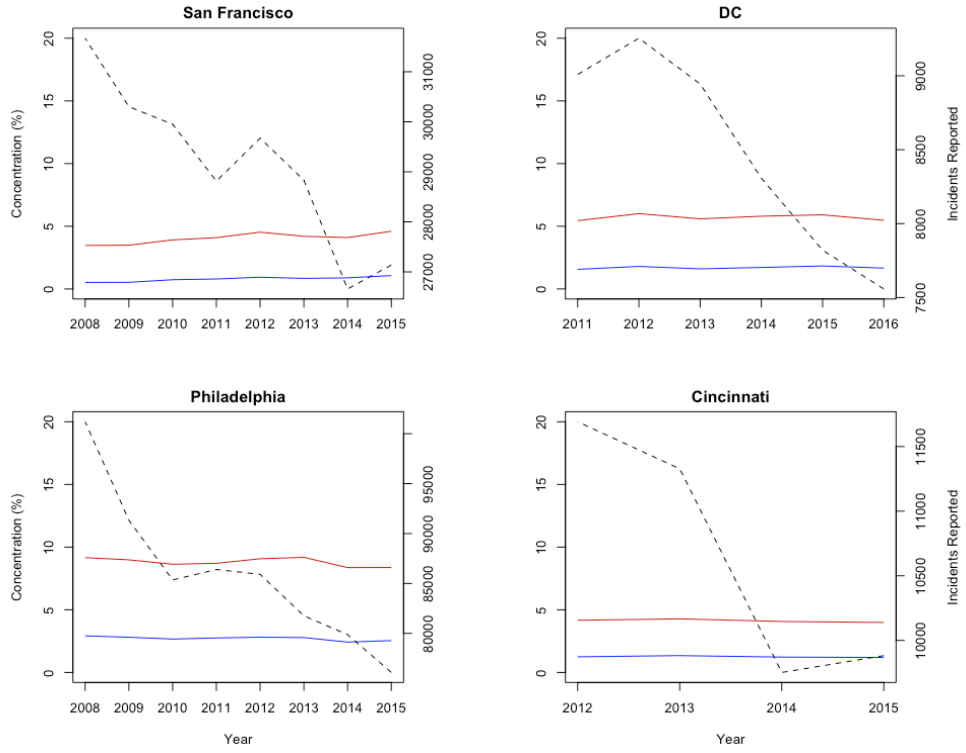


Figure 17: Violent Crime Concentration Levels across Cities

An additional point of intrigue is whether crime concentration differs by violent and nonviolent classification. The violent concentration numbers in Table 2, presented in Figure 17, show that this is indeed the case. There are two notable takeaways from viewing violent in comparison with overall crime concentration. First, it is clear that violent crime sees a higher degree of concentration than crime in general. The mean 50 percent concentration level in the violent-only sample is 3.06 percent, and is 5.62 percent for all categories combined, showing that violent crime is significantly more concentrated than crime in general. Second, the concentration level of violent crime at hotspots is far less consistent across cities than is the case with crime categories in aggregate. This means that, while violent crime may represent a greater opportunity for understanding and policing hotspots in that it is more concentrated, it does not conform nicely to the law of concentration of crime at place.

A natural next step to measuring concentration levels is to examine their stability over time. The data show that concentration

levels are surprisingly consistent over time in each city in the sample (Figure 18). The city with the highest variance in concentration level is Portland, but even this example varies by only 1.33 percentage points, with a maximum of 5.85 and a minimum of 4.52 percent of its street segments being needed to explain half the city's crime over the nine years tested. A better representation of the overall sample is Seattle, whose 50 percent concentration level stayed between 4.73 and 5.63 percent over its eight years tested (Figure 18). The data for Dallas is omitted from the following figures because the two year period that it spans is not large enough to observe a meaningful trend.



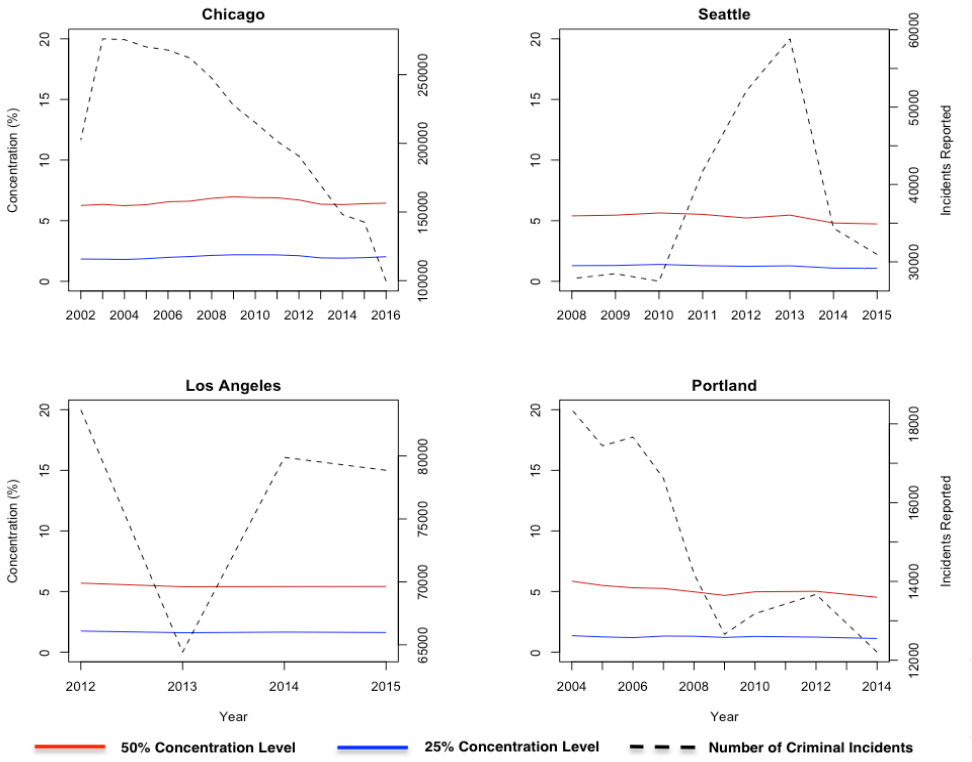


Figure 18: Stability of Crime Concentration Levels Over Time

Examining these trends for violent crime alone, a similar pattern emerges (Figure 19). While the concentration levels are less consistent across cities for violent crime, they are almost perfectly stable over time. Washington, D.C. has the most volatile violent crime concentration level, with its 50 percent concentration level varying between 4.14 and 4.80 percent of street segments, but even this would have placed it among the most stable examples from the previous sample which includes all categories of crime. This finding is at least somewhat surprising due to the violent crime concentration levels being inconsistent across cities. Because violent crime does not fit a ratio of street segments needed to explain a fixed portion of crime that is consistent across cities, there is little reason to expect the concentration levels of this class of crime to be just as stable is the case with all categories of crime combined.

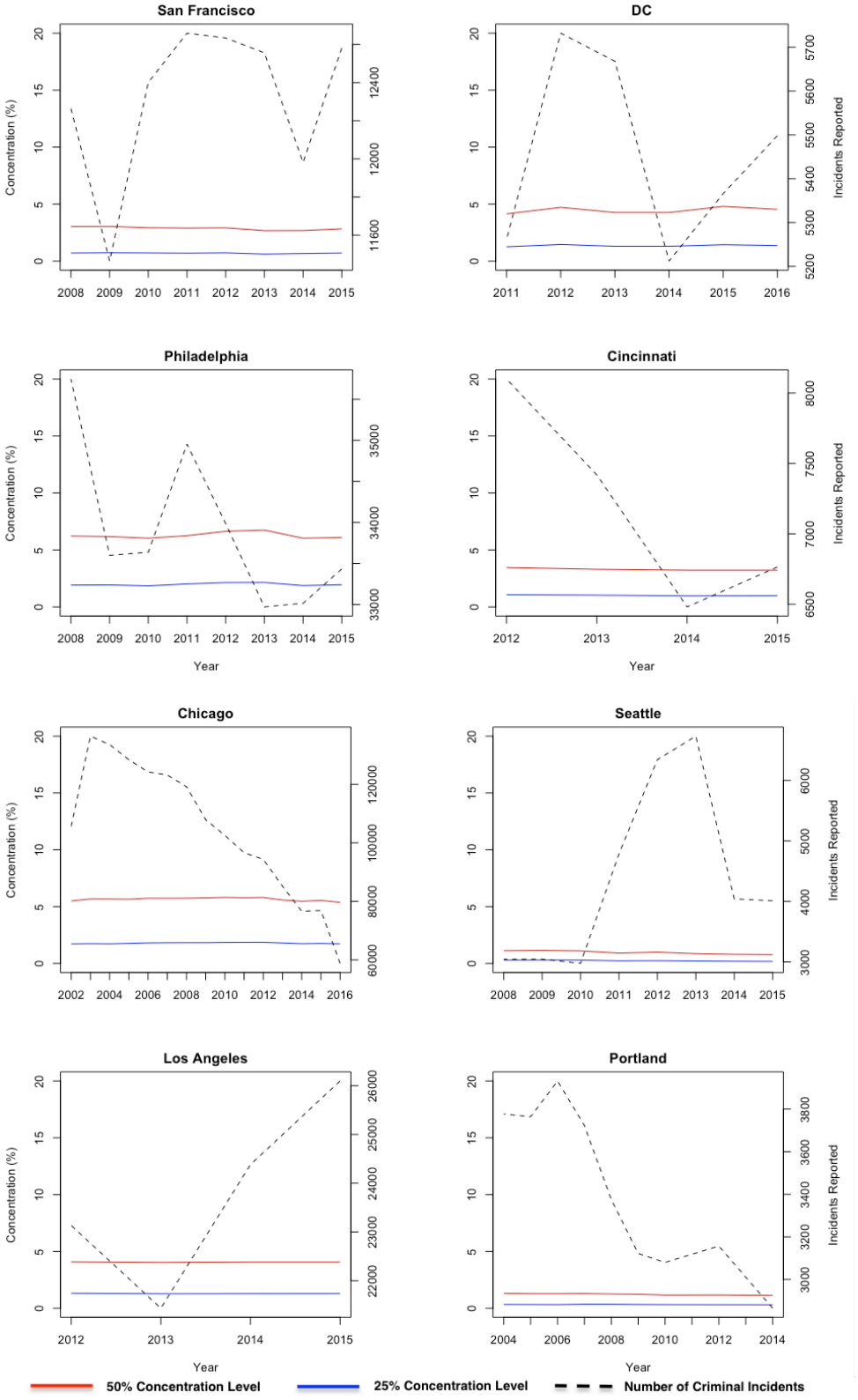


Figure 19: Violent Crime Concentration Levels Over Time

It is worth noting that a small gross change in crime concentration level can still be large in percentage terms due to the scale of these numbers. An increase from 1 to 1.5 percent of street segments being needed in order to explain 25 percent of a city's crime, for example, is both a large relative change, representing a 50 percent increase, and a small gross change, at only half of a percentage point. While the relative changes in crime concentration are larger than the gross changes, gross change in concentration level is a better reflection of the changes being observed in the city with respect to the concentration of crime in micro-scale hotspots.

It is especially interesting that the concentration levels shown in Figures 18 and 19 are resilient not only to time, but to changes in overall crime rate as well. Represented by the dashed line in each plot, the crime rates are shown to move significantly during the time periods examined. In the presence of a serious crime wave or crime drop, such as that observed in Seattle from 2013 to 2015, one would expect to observe a change in the city's concentration level.

In the longer time series in this sample we observe that these ratios are also robust to volatile macroeconomic conditions. San Francisco, Philadelphia, and Portland all show stable concentration levels throughout the Great Recession and subsequent recovery. Chicago shows stability through both the Great Recession and dotcom bust of 2001 and 2002. This shows that crime concentration levels are resilient to changes in unemployment and market performance.

This stability in concentration levels suggests that both crime waves and significant decreases in crime might affect the various sections of a city equally. This is contrary to what one would expect, where it is typically assumed that a movement in the overall crime rate is driven by either social disorganization or improved policing in high crime areas, either of which one would expect to affect the level of crime concentration. The fact that crime concentration levels are resilient to major changes in the crime rate, however, suggests that the distribution of crime across a city's street segments sees minimal change during a crime wave or drop. While this is certainly not proof

of such a phenomenon, this finding does motivate such a question for further research.

Regardless of the distribution of a crime wave's impact across a city, it seems to be the case that each city has a natural level at which crime concentrates in micro-places. Seeing that concentration levels are unaffected by recessions, recoveries, crime waves and declines, it is fair to assume that no reasonably common phenomenon would produce a noticeable change in crime concentration. The reason for this stability, however, is not clear. One potential explanation is that crime concentration is closely related to infrastructural features of cities which see little change over time. A city's layout and road quality, for example, could impact its policeability due to the ways these factors impact intra-city mobility. What I find more likely, however, is that crime is driven by routine activity patterns, which have seen little change in the past few decades for which this data exists. Short of a new shock to daily transportation, working and leisure habits comparable to when the automobile went mainstream in the early 20th century, I would expect crime concentration levels to see little change going forward.

11 Hotspot Movement Over Time

Thus far I have shown a close coupling of crime and place across several cities, confirming the relationship that approximately five percent of a city's street segments explain 50 percent of its crime, and that between one and two percent of street segments explain 25 percent of crime. Further, the concentration level in each city has been shown to be resilient to macroeconomic conditions, time, and changes in the overall crime level. It is of both theoretical and applicable interest, however, whether the hotspots composing the 25 and 50 percent concentration levels are the same segments each year, or whether criminals randomize their behavior in effective ways so that hotspots can not be easily targeted by police. In this section I measure annual hotspot change across cities, and then visualize patterns of hotspot movements in Chicago.

For a measure of hotspot change, I first calculate the hotspots at the 25 percent concentration level for each city. Give these baseline hotspots, I then re-calculate the hotspots at the same concentration level for each subsequent year. Using this procedure, I am able to calculate the percentage of the original year's hotspots that remain hotspots at each point in time. If criminals did not change their behavior whatsoever, close to 100 percent of the original year's hotspots would remain high in crime each year. If criminal behavior was perfectly adjusted to avoid hotspot detection, close to zero percent of the original year's hotspots would be detectable the next year. What we observe in practice is something in-between these two extremes.

In any given year, between 40 and 60 percent of a city's hotspots from the previous year are still classified as such. While the one-year dropoff rate for hotspots is steep, the remaining hotspots tend to stabilize thereafter, with between 30 and 40 percent of the original year's hotspots remaining hot throughout the remaining years for which there is public data (Figure 20). The street segments that remain high in crime after a year passes are the chronically problematic street segments that law enforcement is most concerned with.

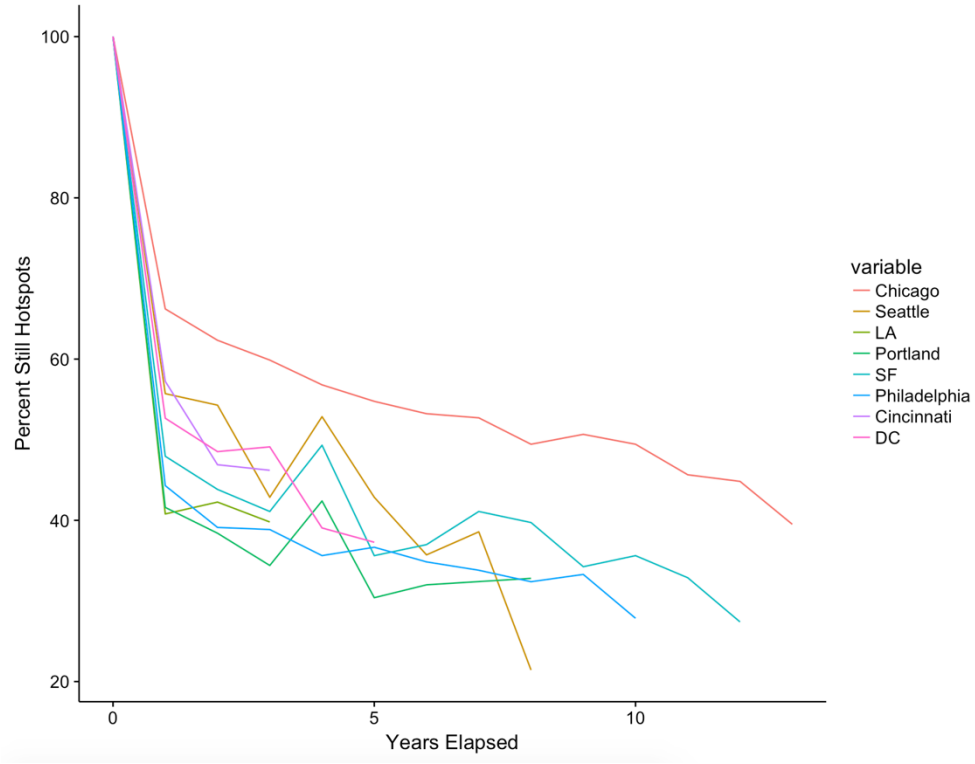


Figure 20: Hotspot Dropoff Rate for Each City

The one exception to the general trend of hotspot change is Chicago. While every other city sees 40 to 60 percent of its hotspots change after one year, Chicago exhibits a noticeably higher number of repeat hotspots. This remains the case as time moves forward, with Chicago seeing an overall lower rate of change among its highest-crime street segments (Figure 20). This suggests that Chicago's criminal hotspots are more persistent than those of other major cities, which is yet another reason to focus further analysis on this city.

Seeing that over half of a city's hotspots change each year, it is important to know whether the relocated hotspots are appearing near the original ones. If hotspots travel long distances when they shift, their usefulness to police will be minimal, as this would show that criminals effectively randomize their behavior in at least half of their common activity spaces. If the hotspots that change each year stay within the same blocks and neighborhoods, however, then this would not be meaningful movement, as a police officer positioned in one of

the original hotspots would still be able to act upon a crime occurring within the same block or neighborhood.

Viewing Chicago's hotspots at the 25 percent concentration level side by side, it becomes clear that although half of the city's hotspots change locations each year, they are staying in the same general areas of the city. Particular areas of the west side, south side, and the coast along Lake Michigan, for example, are consistently filled with high-crime street segments. While the specific street segments identified as hotspots within these areas change each year, they are typically replaced by new high-crime streets within a small number of blocks (Figure 21).

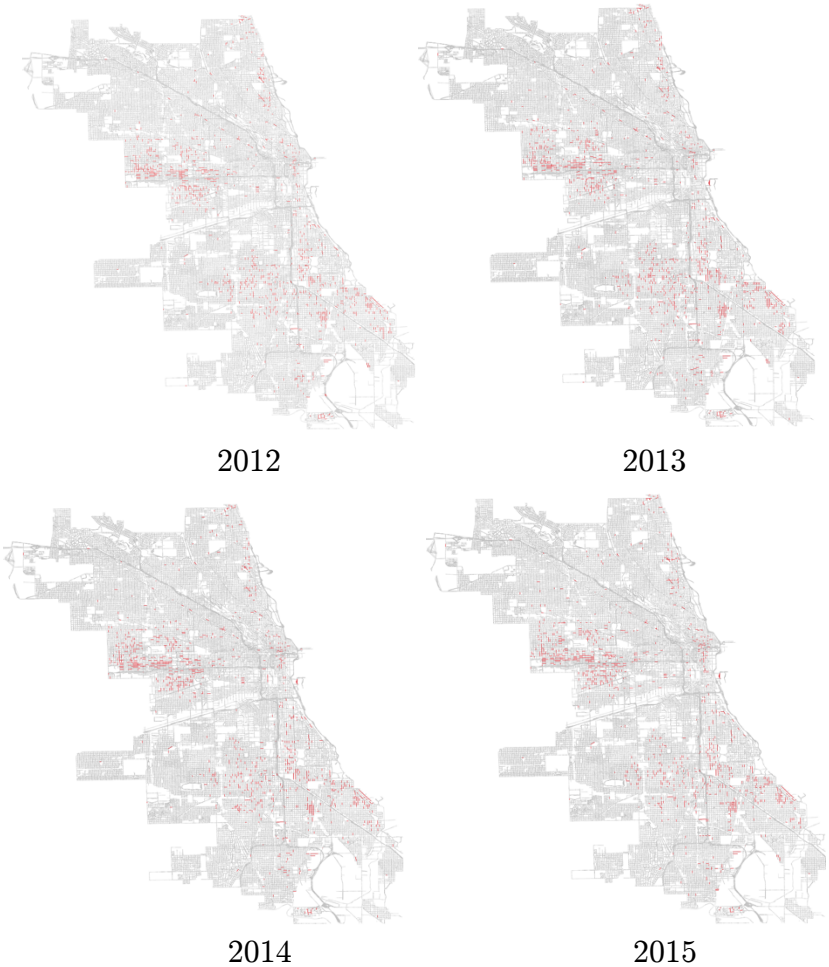


Figure 21: Chicago's Hotspots Over Time

12 Explanatory Power of Facilities and Spatial Features on Crime

Crime at street segments lends itself uniquely well to causal inference due to the reduced noise of this unit's purpose in its broader community. While a higher level unit of analysis such as a police district misses the underlying subtlety of its sub-regions, a street segment is self-contained and socioeconomically homogeneous.

In evaluating models of crime concentration in micro-places, there are two primary model features that should be considered. First, this topic lends itself to more than one potential class of model. Of those available, I consider ordinary least squares, beta regression, and a logit model. Second, it is important to choose a set of variables to use in the models in order to maximize explanatory power and interpretability.

12.1 Features Used

Given the several available variables in this data, it is not necessarily best to use them all in a model. First, I address the multicollinearity problem. Multicollinearity is defined as the existence of a high degree of correlation between independent variables which makes it impossible to determine an independent variable's impact on the dependent variable. When a model's variables display a high degree of multicollinearity, their features lose explanatory power. To think of this intuitively, an OLS coefficient represents the expected change in the dependent variable when increasing the variable of interest by one and holding everything else constant. When a particular variable is highly correlated with other variables in the model, it becomes irrational to consider such a situation where the variable of interest moves and those correlated do not. The result of this phenomenon is that the coefficient can take on unexpected values, such as having the wrong sign or displaying a nonsensical magnitude.

The two primary methods for solving problems of multicollinearity are to find more data and to remove variables that are highly correlated with the others in the model. Because I am

already using all of Chicago's available crime data, I am left with the latter of these solutions.

My method for identifying both the extent to which my models show multicollinearity and the variables responsible is to use the variance inflation factor, or VIF. VIFs measure the severity of multicollinearity in an OLS model. A model's VIF scores are calculated by first running an OLS model for each variable j where j is the dependent variable, with all of the model's other covariates as independent variables. Once this R_j^2 is calculated, the VIF for each variable is equal to $\frac{1}{1-R_j^2}$. A variable that is highly correlated with the rest of a model's covariates will have a high VIF, and a variable that is perfectly independent from the model's other features will have the minimum VIF score of one. As a rule of thumb, a score above 10 is said to be high, below five is safe, and and close to one is ideal. A score between five and 10 is somewhat of a grey area.

It became immediately apparent that using spatial features from varying tiers of distance would result in high multicollinearity, as one would expect. For this reason, the models I run use either the 200 or 600 foot distance measurements in exclusivity, rather than layering them.

Examining VIF values for using 200 and 600 foot (also referred to as one block and two block) distance measures, it is clear that variables counting facilities within two blocks of street segments display a higher degree of multicollinearity (Figure 22). This means that the one-block features will have better explanatory power than the two-block features, because it better satisfies the OLS assumption that a model's regressors are linearly independent from one another.

Name	VIF One Block	VIF Two Block
Percent Aged 25+ Without HS Diploma	9.063304	9.570238
Restaurants	7.025168	24.110412
Percent Housing Units Crowded	6.700657	7.025293
Per Capita Income	6.229225	7.538911
Percent Aged 16+ Unemployed	4.835377	5.184857
Percent Households Below Poverty Line	4.361988	4.659402
Businesses	3.866417	9.806027
Parking Garages	3.052160	9.581123
Liquor Stores	2.690516	6.614500
Graffiti	2.039027	3.424629
Log Distance to City Center	2.011920	2.363623
Bus Stops	1.967973	2.960881
Bars	1.966300	4.422101
Gas Stations	1.562057	2.056679
Arts Venues	1.487420	2.855785
Grocery Stores	1.389757	2.004844
Daycare Centers	1.234876	1.963038
Animal Care Centers	1.231189	2.540679
Subway Stations	1.178680	2.003891
Pawn Shops	1.153886	1.445108
Drug Rehab Centers	1.071132	1.318023
Schools	1.055363	1.434519
Parks	1.053111	1.400569
Length	1.027444	1.024231
Senior Centers	1.020964	1.074632

Figure 22: Variance Inflation Factors with All Variables in Model

The first thing that stands out in Figure 22 is that the socioeconomic features are highly correlated with the rest of the model. This is presumably because indicators such as income, unemployment, housing crowdedness, and education level are all highly related with one another at the community level. For this reason, I drop all individual socioeconomic features and replace them with a single representative feature called the *Intercity Hardship Index*. This statistic, defined by Nathan and Adams (1989), represents the average of the standardized ratios of crowded housing, houses below the poverty line, unemployment, residents without high school diplomas, percent aged either under 18 or over 64, and negative per capita

income. The result is a metric bounded between 0 and 100, capturing six of the major hardship indicators without the concern of including collinear features in the model (Nathan and Adams 1989).

Additional to this, I remove the feature for restaurant count from the model due to its high VIF score. Calculating new VIF scores after these changes yields a model with far less correlation among its regressors (Figure 23).

Variable	VIF
Parking Garages	2.693567
Businesses	2.592867
Liquor Stores	2.342813
Hardship Index	1.937027
Bus Stops	1.930709
Bars	1.785382
Graffiti	1.747554
Gas Stations	1.559414
Arts Venues	1.463577
Log Dist. to City Center	1.400719
Grocery Stores	1.376167
Animal Care Centers	1.226437
Daycare Centers	1.224845
Subway Stations	1.164204
Pawn Shops	1.151082
Drug Rehab Centers	1.061326
Schools	1.046777
Parks	1.039937
Length	1.027085
Senior Centers	1.018200

Figure 23: Variance Inflation Factors of Final Feature Set

While the ideal situation would be for the model to have VIF scores of close to one across the board, this is not often achievable with real-world data. All things considered, I am surprised by the lack of multicollinearity between spatial features, and argue that the observed VIF scores allow us to accept the coefficients of a model using this data as being reliable.

12.2 Class of Model

The second open question is which class of model is most appropriate for understanding crime concentration. Here I consider two classes of model: ordinary least squares and logit. I then extend these models with two classes of coefficient: standardized and non-standardized.

The question of model type comes down to the questions of the extent to which interpretability matters and whether to formulate crime concentration as a problem of crime count or the existence of a high-crime low-crime dichotomy. On the first question, interpretability is of high interest, as reliable and understandable coefficients will contribute to an understanding of the relationship between facilities, spatial features, and crime. These relationships have interesting implications for crime pattern and routine activity theory, which lends support to the use of an ordinary least squares model and its easy-to-understand mapping between coefficients and the dependent variable.

Considering dependent variables, however, it seems most appropriate to consider crime concentration as a binary dependent variable problem. The accuracy of a model of crime in micro places will inevitably be low, and therefore it may not be appropriate to have the illusion of accuracy given by the continuous output of OLS. In the case of crime concentration, a useful binary dependent variable could be set to one when a street is a criminal hotspot at the 25 percent concentration level, and set to zero otherwise. This way, the model would be predicting whether a street is high in crime, rather than attempting to predict exactly how many crimes would happen at a particular street in a given year. Modeling binary hotspots is far more realistic than predicting discrete crime counts, and for this reason I prefer the logit model to OLS despite the challenges of interpretability caused by its nonlinearity.

The second model-related consideration was whether to use standardized coefficients. The benefit of using standardized coefficients is that they allow a direct comparison between variables with different units and scales. Rather than measuring the expected impact of a one-unit change in a regressor on the dependent variable, a standardized coefficient instead measures the expected impact of a one-standard-deviation change. This allows the magnitudes of coefficients of different

units and scales to be measured against one another, which is of particular use due to the differing scales and units of features such as street segment length, the hardship index, and the counts of retirement homes and storefronts within one block.

While the standardized coefficients are clearly useful, the non-standardized coefficients still have their place due to the economic significance of a coefficient representing a unit change in a variable's original unit. Similarly, while the logit model is preferred as a more appropriate formulation of the crime problem, the OLS model remains useful due to its superior interpretability. For these reasons, I employ all of the above models: an OLS regression and a logit model, each with both standardized and non-standardized coefficients.

12.3 Model Specification

12.3.1 Ordinary Least Squares

The equation fit for ordinary least squares is:

$$y_i = \hat{\beta}_0 + \sum_{j=1} \hat{\beta}_j x_{ij} + \widehat{\mu_i},$$

Where y_i is the value of the dependent variable at observation i, $\widehat{\beta_0}$ is the estimated intercept term, $\widehat{\beta_j}$ is the estimated coefficient for variable j, x_{ij} is the value of variable j at observation i, and $\widehat{\mu_i}$ represents the model's error at observation i. In fitting the coefficients that minimize the model's sum of squared residuals, it yields an unbiased estimator where a one-unit increase in variable j at observation i represents a $\hat{\beta}_j$ increase in the estimated output $\hat{y}_i$.

12.3.2 Beta Regression

A beta regression fits the same equation as OLS, with the slight modification that the dependent variable y and set of independent variables $(x_1, \ldots, x_j)$ are all standardized with mean zero and standard deviation one. Formally, beginning with the OLS estimator, we first subtract the means of each term so that

$$(y_i - \bar{y}) = \sum_1 \hat{\beta}_j (x_{ij} - \bar{x_j}) \quad + \hat{\mu}_i.$$

Note that the intercept term $\widehat{\beta_0}$ disappears in this model, since all variables are standardized to mean zero and the intercept always runs through the point $(\bar{X}, \bar{Y})$. Also keep in mind that $\widehat{\mu_i}$ has sample mean zero, allowing the term to stay in the model as-is.

Next, we divide both sides by σ_y, the sample standard deviation of y, and then both multiply and divide each right hand side coefficient by the sample standard deviation of x_j, denoted σ_j.This yields:

$$\frac{(y_i - \bar{y})}{\sigma_y} = \sum_{j=1} \frac{\sigma_j}{\sigma_y} \widehat{\beta_j} \frac{(x_{ij} - \bar{x_j})}{\sigma_j} + \widehat{\mu_i},$$

where $\frac{(x_{ij} - \bar{x_j})}{\sigma_j}$ is the z score of x_{ij}. Beta regression, then, can be written as:

$$z_y = \sum_{j=1} \frac{\sigma_j}{\sigma_y} \widehat{\beta_j} z_{ij} + \widehat{\mu_i},$$

and then simplified to:

$$z_y = \sum_{j=1} \widehat{b_j} z_{ij} + \widehat{\mu_i},$$

Where $\widehat{b_j}$ is the standardized beta coefficient for variable j (Wooldridge 2013). With both the right and left hand side variables converted to z-scores, the beta coefficients now represent the expected standard-deviation change in y given a one standard deviation change in x. As was mentioned earlier, the benefit of this is that variables of differing scales and units of measurement can be directly measured against one another when standardized this way, allowing for the judgment of which features have the largest impact on crime.

12.3.3 Logit

Ordinary least squares is no longer an appropriate model when the dependent variable is binary. While it is possible to run a linear probability model, regressing a set of independent variables on a binary dependent variable in an OLS model, the result of this would be a

probabilistic model whose values may either exceed one or fall below zero for much of the function's domain. The logit model is a solution to this, being linear in parameters and bounded between zero and one, yielding valid probabilistic estimates.

To get from ordinary least squares to the logit model, begin with the linear probability model:

$$\pi_i = \hat{\beta}_0 + \sum_{j=1} \hat{\beta}_j x_{ij} + \hat{\mu}_i,$$

Where π_i is the predicted probability that $y_i = 1$. Converting π_i into the odds ratio $\frac{\pi_i}{1-\pi_i}$, one can then obtain the log odds, also called the logit:

$$\eta_i = \ logit(\pi_i) = log \frac{\pi_i}{1 - \pi_i}.$$

Finally, taking the inverse of the logit gives:

$$\pi_i = \frac{e^{\eta_i}}{1 \ + \ e^{\eta_i}}.$$

If we assume the logit of the underlying probability that the dependent variable equals one is a linear function of the predictors, we obtain our logit model of:

$$\widehat{\pi_i} = \frac{e^{(\widehat{\beta}_0 + \sum_{j=1} \widehat{\beta}_j x_{ij})}}{1 + e^{(\widehat{\beta}_0 + \sum_{j=1} \widehat{\beta}_j x_{ij})}},$$

transforming the original linear probability model into a nonlinear function bounded between zero and one (Rodriguez 2007). The coefficients of this model are not nearly as interpretable as those from OLS due to its nonlinearity. A unit increase in variable j no longer means an expected $\hat{\beta}_j$ increase in y; the new value now needs to be passed through the model in order so see what impact the change will have.

One possible solution to this is to take the marginal effects of the model's variables at the mean. The logit model flattens at its tails as it nears zero and one respectively, but its effects are relatively close to linear at its features' means (Figure 24).

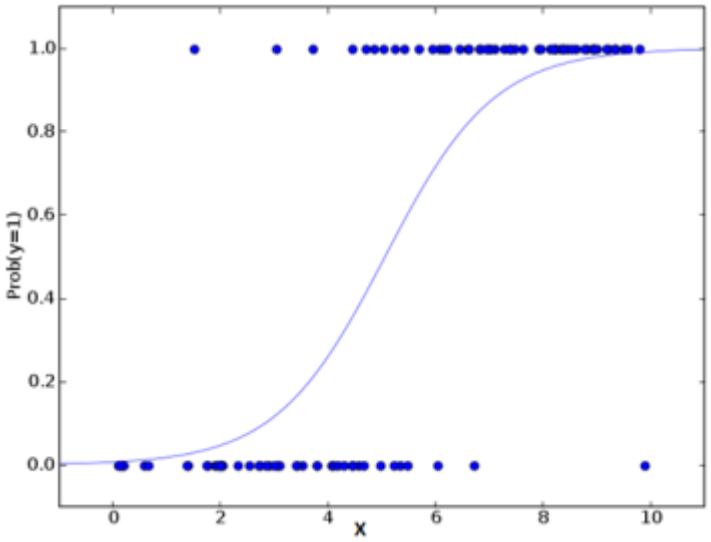


Figure 24: Predicted Logit Probabilities vs. Binary Target Values (Source: Analytics Vidhya)

Knowing this, the marginal effect of a variable at its sample average will provide something close to the expected increase in probability caused by a unit-increase in that particular variable. While this relationship will be different at all other values that the feature takes on, the marginal effect is useful in that it gives this model a degree of interpretability. To calculate these values, take the partial derivative of y with respect to the variable of interest at its sample mean.

12.3.4 Standardized Logit

The logistic regression equivalent of beta regression is slightly different because the dependent variable is binary. Where beta regression standardizes both the left and right hand sides of the OLS equation, it is complicated to fully standardize a logit model in this same way, and it does not necessarily make sense to do so (Menard, 2011). A similar effect, however, can be obtained by standardizing the model's independent variables and running the same logit model on the z scores of the original variables. While the standardized logit coefficients will be uninterpretable for the same reason as the non-standardized ones, their marginal effects will be comparable despite differing units and scales.

12.4 Results

The OLS and beta regression coefficients are reported in Table 3. For each model, an observation is a Chicago street segment in the year 2015, and the independent variable is the number of crimes happening on this segment during that year. The OLS coefficient is reported in the *OLS* column, and the standardized beta regression coefficient is reported in the *Beta* column. Note that, as expected, the intercept value is zero for the beta regression.

	Dependent variable:	
	Crime Count	
	OLS	Beta
	(1)	(2)
Intercept	4.097^{***}	0.000^{***}
	(0.615)	(0.615)
Subway Stations	0.575^{***}	0.042^{***}
	(0.059)	(0.059)
Bus Stops	0.013^{***}	0.029^{***}
	(0.003)	(0.003)
Parks	0.110^{***}	0.018^{***}
	(0.025)	(0.025)
Graffiti (Num. Reports)	-0.001^{***}	-0.070^{***}
	(0.0001)	(0.0001)
Length (ft.)	0.001^{***}	0.070^{***}
	(0.0001)	(0.0001)
Log Dist. City Center (ft.)	-0.117^{***}	-0.034^{***}
	(0.016)	(0.016)
Bars	0.109^{***}	0.038^{***}
	(0.016)	(0.016)
Schools	0.334^{***}	0.055^{***}
	(0.025)	(0.025)
Grocery Stores	0.224^{***}	0.037^{***}
	(0.028)	(0.028)
Senior Centers	0.224	0.006
	(0.142)	(0.142)
Businesses	0.001	0.005
	(0.001)	(0.001)
Parking Garages	0.071^{***}	0.028^{***}
	(0.017)	(0.017)
Liquor Stores	0.130^{***}	0.047^{***}
	(0.017)	(0.017)
Daycare Centers	0.062^{***}	0.014^{***}
	(0.019)	(0.019)
Animal Care Centers	-0.029	-0.004
	(0.030)	(0.030)
Gas Stations	0.140^{***}	0.026^{***}
	(0.027)	(0.027)
Drug Treatment Centers	0.296^{***}	0.034^{***}
	(0.036)	(0.036)
Pawn Shops	0.071	0.004
	(0.082)	(0.082)
Arts Venues	0.035	0.004
	(0.043)	(0.043)
Hardship Index	0.017^{***}	0.144^{***}
	(0.001)	(0.001)
Observations	57,426	57,426
R^2	0.062	0.062
Adjusted R^2	0.062	0.062
Residual Std. Error (df = 57388)	3.274	3.274
F Statistic (df = 37; 57388)	102.741^{***}	102.741^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3: OLS and Beta Regression Coefficients

The first thing one notices is that almost every feature tested is significant in this model. The only variables that are not significant at traditional levels are the numbers of arts venues, pawn shops, businesses, and senior centers within one block of a street segment, with senior center count being close with $p < .12$. The insignificance of pawn shops is most surprising among these, as pawn shops typically carry a reputation for selling stolen goods and being a center for interpersonal conflict.

Other coefficients confirm existing suspicions. It has long been known, for example, that bars, public parks, gas stations, and liquor stores are home to large amounts of crime. Similarly, it does not come as a surprise that socioeconomic hardship has a positive and significant association with crime count.

Still more variables one might have no prior assumption about. Bus stops, daycare centers, and businesses, for example, seldom enter the discussion on the topic of crime concentration. These models show, however, that controlling for a wide array of socioeconomic and spatial features, these variables all have positive and significant relationships with the crime level in micro-places.

The sign and significance level of the variable for graffiti presence might be the most surprising in this model. While one would expect graffiti presence to have a positive relationship with crime due to its association with gang activity, the coefficient was in fact negative and highly significant. This could be the case for a variety of reasons, but the most likely are that either crime in graffiti-covered areas is low because police presence on these streets is high, or we have an irrational fear of these areas which is simply not consistent with the level of crime that is observed in reality. Additionally, it is also possible that graffiti is reported most often in highly supervised neighborhoods, and that the 311 calls for graffiti removal are implicitly picking up the effect of neighborhood watches and citizens' concerns for their local environments.

Last, the intercept term is slightly larger than one might expect, at just over four crimes. Due to the large coefficients on the *distance to city center* and *hardship index* variables, however, this could make intuitive sense. This combination of coefficients could mean, for example, that streets close to downtown and in impoverished areas are

expected to see a high baseline level of crime, while the well-off neighborhoods outside the city center still see a low expected crime count.

Turning to the standardized coefficients, we see that the hardship index, street segment length, graffiti presence, liquor stores, and schools have the largest impacts on crime count at the street segment level. The control variable for street segment length is expected to be among the most important, because longer streets have more room for human activity, criminal and otherwise. It is also in line with expectations that the hardship index be high in magnitude, because the factors constituting this explain several factors of people in the area's living conditions and expected routine activities. The high beta coefficient on school count is at least in part due to the relatively small number of schools in the city, and the disproportionately large impact that a school has on its local environment. The presence of a school essentially guarantees a high amount of activity on a street segment, and also draws in the particularly crime-heavy younger age groups. The only surprising feature among these, again, is the amount of graffiti that has been reported on a street, and that is because the impact is negative when one would expect it to be positive.

The results of the logit and standardized logit models are shown in Table 4. As was the case with OLS and beta regression, the column *Logit* represents the original logistic regression coefficients, and the *Standardized Logit* column represents the coefficients when the model's independent variables are standardized to their z scores.

	Dependent variable:	
	Crime Count	
	Logit	Standardized Logit
	(1)	(2)
Intercept	−1.300	−4.499***
	(1.336)	(0.044)
Subway Stations	0.305***	0.076***
	(0.091)	(0.023)
Bus Stops	0.014***	0.107***
	(0.005)	(0.036)
Parks	0.190***	0.105***
	(0.052)	(0.029)
Graffiti (Num. Reports)	−0.002***	−0.399***
	(0.0003)	(0.050)
Length (ft.)	0.001***	0.179***
	(0.0001)	(0.022)
Log Dist. City Center (ft.)	−0.094**	−0.092**
	(0.037)	(0.036)
Bars	0.108***	0.127***
	(0.024)	(0.028)
Schools	0.432***	0.241***
	(0.047)	(0.026)
Grocery Stores	0.139***	0.078***
	(0.049)	(0.028)
Senior Centers	0.203	0.020
	(0.274)	(0.027)
Businesses	−0.0004	−0.014
	(0.001)	(0.033)
Parking Garages	0.112***	0.146***
	(0.031)	(0.041)
Liquor Stores	0.146***	0.177***
	(0.031)	(0.038)
Daycare Centers	0.083**	0.066**
	(0.039)	(0.031)
Animal Care Centers	−0.101	−0.051
	(0.081)	(0.041)
Gas Stations	0.160***	0.099***
	(0.047)	(0.029)
Drug Treatment Centers	0.329***	0.130***
	(0.056)	(0.022)
Pawn Shops	0.044	0.008
	(0.123)	(0.022)
Arts Venues	−0.150*	−0.058*
	(0.078)	(0.030)
Hardship Index	0.019***	0.528***
	(0.002)	(0.049)
Observations	57,426	57,426
Log Likelihood	−4,475.993	−4,475.993
Akaike Inf. Crit.	9,027.986	9,027.986
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4: Original and Standardized Logistic Regression Coefficients

Due to the challenges associated with interpreting the coefficients of a logit model, the marginal effects of both the original and standardized logit models' coefficients, reported at the means, are shown in Table 5.

Name	Estimate	Standardized
Businesses	-0.000005	-0.000155
Parking Garages	0.001216	0.001586
Liquor Stores	0.001584	0.001921
Graffiti	-0.000025	-0.004336
Log Dist. to City Center	0.001019	-0.001002
Bus Stops	0.000155	0.001167
Bars	0.001178	0.001379
Gas Stations	0.001742	0.001081
Arts Venues	-0.001637	-0.000630
Grocery Stores	0.001501	0.000851
Daycare Centers	0.000908	0.000716
Animal Care Centers	-0.001097	-0.000555
Subway Stations	0.003316	0.000825
Pawn Shops	0.000478	0.000086
Drug Rehab Centers	0.003581	0.001409
Schools	0.004705	0.002620
Parks	0.002063	0.001146
Length	0.000009	0.001950
Senior Centers	0.002436	0.002436
Hardship Index	0.000205	0.005742

Table 5: Original and Standardized Logit Marginal Effects

The largest non-standardized marginal effects are for school count, rehab facilities, and subway stations with marginal effects of .0047, .0035, and .0033 respectively. This means that the addition of a school within one block of a street segment, all else held equal, increases the expected probability of the street being a criminal hotspot by 0.47 percent. Similarly, adding an additional rehab facility adds 0.35 percent to this probability, and an additional subway station adds 0.33 percent.

The standardized marginal effects have slightly different interpretations. These effects measure the expected increase in probability resulting from a one standard deviation change in an independent variable. Similar to beta regression, this allows the impacts of independent variables of different units and scales to be measured directly against one another. The largest standardized marginal effects in absolute terms are the hardship index, graffiti count, and school presence, with marginal effects of .0057, -.0043, and .0026. These correspond to expected 0.57, -0.43, and 0.26 percent changes in the probability of a street being a hotspot at the 25 percent level resulting from one-standard-deviation changes in each of these features.

Upon examining the marginal effects, it initially appears that the impacts of the individual variables on the probability of a street being a criminal hotspot are quite small. Considering these effects in context, however, this should not be surprising. Recalling the definition of a hotspot, there are very few of these in any given city relative to its total number of street segments. In Chicago specifically, hotspots for violent crime at the 25 percent level make up only 1.78 percent of street segments. With this in mind, a change in probability on the scale of tenths of a percent could still hold economic significance, as a small change can still be meaningful in relation to the baseline probability of a street segment being a hotspot.

Despite formulating the crime problem in two different ways, one as a binary problem of modeling high vs. low crime street segments and the other as an estimator of crime count, the two classes of model widely agree on the significance and direction of effects. Senior centers, businesses, animal care facilities, and pawn shops are all insignificant at the ten percent level or higher in both models. The only variable whose significance differs between the two models by traditional standards is arts venues, which is significant in the logit model and not in the OLS model.

The two classes of model agree on the signs of coefficients as well. The only variables whose signs differ are those which are insignificant in either one model or both. The signs of the variables for businesses and arts venues both differ between model classes, for example, but the business coefficient is not statistically different from

zero in either model, and the arts venue variable is insignificant in the OLS model. Apart from these two variables, the two model classes agree on the signs of all other coefficients.

Moving beyond sign and significance, we can also see the extent to which the model classes agree on which variables have the largest impact. By ranking the absolute values of the standardized coefficients for the beta and standardized logit models, we can use a measure of rank similarity called Kendall's tau coefficient. Kendall's tau, also called Kendall correlation, measures the ordinal association between two measured quantiles. Formally, the coefficient is defined as:

$$\tau = \frac{c - d}{n(n-1)/2},$$

Where c is the number of concordant pairs, d is the number of discordant pairs, and the denominator is equal to the number of total pair combinations. A Kendall correlation of one indicates perfect agreement between the two orderings of coefficient magnitude, a correlation of negative one indicates perfect disagreement between the two orderings, and a correlation of zero indicates independence between the two sets.

The magnitudes of the standardized coefficients of the two models have a Kendall correlation of 0.59 ($p < 0.001$), indicating significant positive agreement between the two rankings. Both models rank the controls for socioeconomic status and street segment length highly, along with the coefficients for graffiti and school presence. Similarly, the models both agree that pawn shops, performing arts venues, and daycare centers have relatively little impact on crime compared to the other regressors. The most significant disagreements between the models are that the logit model places a higher relative importance on the impacts of senior centers and parking garages, while the beta regression places higher relative importance on subway stations and grocery stores (Table 6). A Kendall correlation of greater than 0.5 indicates a high degree of agreement between the two models, which can be qualitatively seen in Table 6.

Rank	Standardized Logit	Beta
1	Hardship Index	Hardship Index
2	Graffiti	Graffiti
3	Schools	Length
4	Senior Centers	Schools
5	Length	Liquor Stores
6	Liquor Stores	Subway Stations
7	Parking Garages	Bars
8	Drug Rehab Centers	Grocery Stores
9	Bars	Drug Rehab Centers
10	Bus Stops	Log Distance to City Center
11	Parks	Bus Stops
12	Gas Stations	Parking Garages
13	Log Distance to City Cent	Gas Stations
14	Grocery Stores	Parks
15	Subway Stations	Daycare Centers
16	Daycare Centers	Senior Centers
17	Arts Venues	Businesses
18	Animal Care Centers	Animal Care Centers
19	Businesses	Arts Venues
20	Pawn Shops	Pawn Shops

Table 6: Standardized OLS and Logit Coefficients Ranked from Highest to Lowest

The last important piece to mention about these models is their goodness-of-fit measures. For the OLS and beta regressions, their r-squared values are quite low, at 0.06. The logit model is not as simple to evaluate, as the pseudo r-squared metric is not as meaningful as its OLS analog. Two alternative ways to evaluate fit for this model are classification accuracy and area under the ROC curve (AUC). Classifying all observations with predicted probabilities greater than 0.3 as hotspots yields 98.2 percent in-sample accuracy, with a 7:16 true positive to false positive ratio. Classifying only observations with outputs greater than 0.5 as hotspots is not recommended with this model, because this predicts very few positive outcomes. This is a result of the earlier-discussed problem of this being a highly imbalanced data set with a low baseline probability of a street being a hotspot. Accuracy

in general is a poor metric for the evaluation of this logit model, since even a naïve classifier that always predicts zero will be highly accurate. For this reason, the AUC score is a better metric for the fit of this model.

AUC score is defined as the area under the ROC curve, where ROC stands for receiver operating characteristic. This curve plots the true positive rate, also known as the model's sensitivity, against its false positive rate, defined as one minus specificity. A naïve classifier, even in an imbalanced data set, will only gets a score of 0.5 for the area under this curve. The closer the area under the ROC curve is to one, the better the fit of the model is said to be (Fawcett 2004). This logit model has an AUC score of 0.7807 in-sample, which is considered to be relatively high (Figure 25).

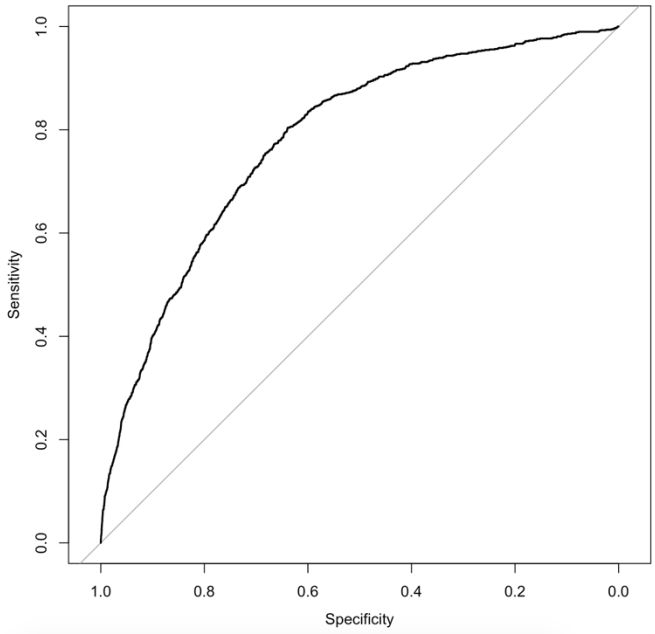


Figure 25: ROC Curve for Logit Model

The closeness of fit for the linear models is overall quite poor, but the classification accuracy of of the probabilistic models is reasonably high, as is shown by the accuracy and AUC scores. In the end, however, goodness of fit is not incredibly important in these models. The purpose of these models, rather than predicting

accurately, is to assess the impacts of various spatial and facility-based features on expected crime level. In this purpose, each model tested shows economic and statistical significance.

13 Discussion

In this thesis I have demonstrated a close coupling of crime and place and explored the relationships between spatial features and crime. Making use of government-provided open data on crime incidents, socioeconomics, city infrastructure, and commerce, I have conducted an analysis of the interaction between local environments and criminal activity that has only recently become possible.

My findings reaffirm the Weisburd (2015) claim of a law of concentration of crime at place. Testing the extent to which crime clusters into small geographic spaces within cities, I find that only 5.62 percent of street segments in major cities are responsible for half of observed crime, and that as few as 1.56 percent of street segments are responsible for 25 percent. This means that crime, rather than being problematic across entire cities, is densely concentrated into a small number of micro-places.

Further, my results show that the crime concentration levels in cities are stable over time and robust to changes in both economic conditions and the overall crime rate. Slightly more than half of hotspots change year-to-year, but they tend to stay within the same general areas of a city when they move. Further, a significant portion remain high in crime over time, showing that persistent criminal hotspots exist in cities despite their eventual detection by law enforcement.

Searching for causal factors of crime concentration, I found that the numbers of schools, subway stations, bus stops, rehab centers, grocery stores, public parks, parking garages, liquor stores, daycare centers, and gas stations all have a positive and significant relationship with the amount of crime observed on a street segment in the city of Chicago. The two classes of model tested disagreed on the significance of performing arts venues, but this feature may have had an impact on crime level as well. These results were found while controlling for the sizes of street segments, their proximity to downtown, the age

composition of the community areas they rested within, and the degree of socioeconomic hardship observed in the surrounding area.

The fact that most variables have a positive and significant relationship with crime is not surprising from a routine activity theoretic perspective. According to this theory, the chief driver of crime is the intersection of motivated offenders, easy targets, and a lack of capable guardians against crime. While other factors may play a role, the primary cause of this overlap in necessary factors is having people's routine activities overlap in space and time. For this reason, any facility that regularly attracts large numbers of people should be more likely than other locations to observe a high level of crime. While grocery stores, daycare centers, and performing arts venues may not be sites typically associated with criminal activity, the mere increase in foot and car traffic generated by their existence on a block or street segment is enough to indirectly drive crime by increasing the intersections of motivated offenders and easy targets in their local environments.

From the perspective of crime pattern theory, these locations would be considered crime generators. They are locations that draw people in for non-criminal purposes, but create criminal opportunities nonetheless that are tempting enough to sway people's intentions toward the unlawful. By drawing in significant crowds with innocent intentions, and causing an increase in crime despite this, these results show that daycare centers, grocery stores, and performing arts venues are crime generators.

The other spatial features with positive impacts on crime are not particularly surprising. Bars, gas stations, liquor stores, bus stops, subway stations, parking garages, and schools are all known for being hosts to criminal activity. Gas stations, bars, and liquor stores, for example, are common sites for robberies, where potential offenders know they can find an easy target. Similarly, schools serve as centralized locations for drug deals and interpersonal conflict among students. What these facility types have in common is that they all have properties that make them appealing sites for crime. As a result, motivated offenders seek these locations out, causing higher levels of crime in their surrounding areas. For this reason, they are considered crime attractors.

It is worth noting briefly that it is possible for a location to be both a crime generator and attractor. A school, for example, can clearly perform both roles. It generates offenders because it plays host to a captive audience five days per week, most of which has innocent intentions, while also attracting offenders who know that this audience will allow them to achieve their criminal goals. Bars are a similar case; while they certainly do attract already-motivated offenders, the presence of alcohol also serves to modify the previously-benign intentions of some toward the criminal.

The last class of location defined under crime pattern theory is the fear generator. Fear generators are society's red herrings; they are accused of causing crime due to their perceived levels of danger, but the data show that these fears are unfounded. Fear generators show themselves in models of crime as variables with either insignificant or negative coefficients, where positive and significant coefficients were expected. The key fear generators identified in this thesis are streets containing graffiti and pawn shops. The number of graffiti reports on a street segment has a negative and significant relationship with the observed number of criminal incidents, and the relationship between the number of nearby pawn shops and crime was not statistically significant. The coefficient on graffiti was expected to be positive because of its perceived relationship with gang violence and narcotics. Pawn shops, similarly, are expected to be sites of robberies and other types of conflict, due to the cash and goods they keep in inventory and their reputation for selling stolen goods. Neither of these features show a positive relationship with crime, however, which suggests that society's fears surrounding pawn shops and the presence of graffiti may be unfounded.

The findings of this thesis are of interest to those involved in both the police force and urban planning. Whether a facility is considered a crime generator or attractor, it is important to understand the expected impact on public safety when approving a building permit or business license. These results show that seemingly innocent facilities can have unforeseen impacts on crime within their local environments. Further, it is important for police agencies to understand both the principal criminal hotspots and fear generators in their cities. This

thesis shows that the vast majority of a city's crime comes from a small number of its street segments. Directing police officers away from fear generators and toward criminal hotspots could both save taxpayer money and improve public safety.

There remains room for improvement in the current state of the art for microgeographic models of crime, but this does not mean that we should not consider implementing such models today. While further research is needed with respect to differing micro-level units of analysis, classes of statistical model, and implementation strategies, recent advances in the criminology of place have shown significant and actionable results. This thesis represents yet another proof of concept for employing such models in US cities, demonstrating that econometric models of crime with microgeographic units of analysis can be used to design safer cities.

14 References

1. Analytics Vidhya. "Simple Guide to Logistic Regression in R." *Analytics Vidhya*. N.p., 08 July 2016. Web. 05 May 2017.
2. Becker, Gary S. "Crime and Punishment: an Economic Approach." *The Economic Dimensions of Crime* (1968): 13-68. Web.
3. Bowers, K. J. "Domestic Burglary Repeats and Space-Time Clusters: The Dimensions of Risk." *European Journal of Criminology* 2.1 (2005): 67-92. Web.
4. Braga, Anthony A., Andrew V. Papachristos, and David M. Hureau. "The Concentration and Stability of Gun Violence at Micro Places in Boston, 1980–2008." *Journal of Quantitative Criminology* 26.1 (2010): 33-53. Web.
5. Braga, A. A., D. M. Hureau, and A. V. Papachristos. "The Relevance of Micro Places to Citywide Robbery Trends: A Longitudinal Analysis of Robbery Incidents at Street Corners and Block Faces in Boston." *Journal of Research in Crime and Delinquency* 48.1 (2010): 7-32. Web.
6. Brantingham, P.J., and P.L. Brantingham. "Environment, Routine, and Situation: Toward a Pattern Theory of Crime." *Routine activity and rational choice*. By R. V. G. Clarke and M. Felson. New Brunswick, NJ: Transaction Publishers, 1993. Web.
7. Buonanno, Paulo. "The Socioeconomic Determinants of Crime: a Review of the Literature." *Working Paper Series n. 63, Department of Economics, University of Milan-Bicocca, Milan* (2006). Web.
8. Cohen, Lawrence E., and Marcus Felson. "Social Change and Crime Rate Trends: A Routine Activity Approach." *American Sociological Review* 44.4 (1979): 588. Web.
9. Eck, John E., and David Weisburd. "Crime Places in Crime Theory." *Crime Prevention Studies, Vol. 4* (1995). Web.

10. Fawcett, Tom. "ROC Graphs: Notes and Practical Considerations for Researchers." Tech. Rep. HPL-2003-4, *HP Labs*, 2004.

11. Guerry, A.-M, Hugh P. Whitt, and Victor W. Reinking. "A translation of Andre-Michel Guerry's Essay on the moral statistics of France (1883): a sociological report to the French Academy of Science. Lewiston, NY." Edwin Mellen Press, 2002. Web.

12. Hijmans, Robert J. "geosphere: Spherical Trigonometry. R package version 1.5-5" (2016).

13. Jeffery, C. Ray. "Crime prevention through environmental design." *Beverly Hills: Sage Publications*, 1971. Web.

14. Levine, Ned. "CrimeStat IV: A Spatial Statistics Program for the Analysis of Crime Incident Locations, Version 4.0." Tech. no. 242960. N.p.: National Criminal Justice Reference Service, n.d. Web.

15. Menard, S. "Standards for Standardized Logistic Regression Coefficients." *Social Forces* 89.4 (2011): 1409-428. Web.

16. Mohler, G.O., M.B. Short, S. Malinowski, M. Johnson, G.E. Tita, A.L. Bertozzi, and P.J. Brantingham. "Randomized controlled field trials of predictive policing." *J Am Stat Assoc* 110(512) (2015): 1399–1411. Web.

17. Nathan, Richard P., and Charles F. Adams. "Four Perspectives on Urban Hardship." *Political Science Quarterly* 104.3 (1989): 483. Web.

18. PredPol. "UCLA Study on Predictive Policing." *PredPol*. N.p., 29 Nov. 2015. Web. 05 May 2017.

19. Quetelet, Adolphe. "A treatise on man and the development of his faculties." *William and Robert Chambers*, 1842. Web.

20. Rodríguez, Germán. "Lecture Notes on Generalized Linear Models." *Generalized Linear Models*. Data.princeton.edu. 01 September 2007. Web. 05 May 2017.

21. Rosser, Gabriel, Toby Davies, Kate J. Bowers, Shane D. Johnson, and Tao Cheng. "Predictive Crime Mapping: Arbitrary Grids or Street Networks?" *Journal of Quantitative Criminology* (2016). Web.
22. Smith, Megan, and Roy Austin Jr. "Launching the Police Data Initiative." *National Archives and Records Administration.* National Archives and Records Administration, 18 May 2015. Web. 05 May 2017.
23. Sherman, Lawrence W., Patrick R. Gartin, and Michael E. Buerger. "Hot Spots Of Predatory Crime: Routine Activities And The Criminology Of Place." *Criminology* 27.1 (1989): 27-56. Web.
24. Taylor, R. B. "Social Order and Disorder of Street Blocks and Neighborhoods: Ecology, Microecology, and the Systemic Model of Social Disorganization." *Journal of Research in Crime and Delinquency* 34.1 (1997): 113-55. Web.
25. Weisburd, David. "The Law Of Crime Concentration And The Criminology Of Place." *Criminology* 53.2 (2015): 133-57. Web.
26. Wooldridge, Jeffrey M. "Introductory econometrics: a modern approach." *Cengage Learning*, 2013. Print.

15 Data Sources

1. Chicago Park District. *Parks – Locations.* Chicago Data Portal, 2016. Web. 01 December 2016.
2. Chicago Police Department. *Crimes – 2001 to Present*. Chicago Data Portal, 2016. Web. 01 December 2016
3. Chicago Public Schools. *Chicago Public Schools – School Locations SY1415*. Chicago Data Portal, 2016. Web. 01 December 2016.
4. Chicago Transit Authority. *CTA – Bus Stops - Shapefile*. Chicago Data Portal, 2016. Web. 01 December 2016.
5. Chicago Transit Authority. *CTA – Ridership – 'L' Station Entries – Daily Totals.* Chicago Data Portal, 2016. Web. 01 December 2016.
6. Cincinnati Police Department. *Crime Incidents*. City of Cincinnati, 2016. Web. 01 December 2016.
7. City of Chicago. *Business Licenses – Current Liquor and Places of Amusement Licenses*. Chicago Data Portal, 2016. Web. 01 December 2016.
8. City of Chicago. *Business Licenses*. Chicago Data Portal, 2016. Web. 01 December 2016.
9. City of Chicago. *Senior Centers*. Chicago Data Portal, 2016. Web. 01 December 2016.
10. City of Chicago. *311 Service Requests – Graffiti Removal.* Chicago Data Portal, 2016. Web. 01 December 2016.
11. City of Chicago. *Street Center Lines.* Chicago Data Portal, 2016. Web. 01 December 2016.
12. City of Dallas GIS Services. *Streets Shapefile*. Dallas City Hall. N.d. Web. 01 February 2017.
13. Dallas Police Department. *All Crime*. Dallas Open Data, 2017. Web. 01 February 2017.
14. DC GIS. *Street Centerlines*. Open Data DC, 2017. Web. 01 March 2017.
15. District of Columbia Metropolitan Police Department. *Crime Incidents in 2011 – 2016*. Open Data DC, 2017. Web. 01 March 2017.

16. Los Angeles Bureau of Engineering. *Street Centerline*. Los Angeles Open Data Portal, 2016. Web. 01 December 2016.
17. Los Angeles Police Department. *Crime Data From 2010 to Present*. Los Angeles Open Data Portal, 2016. Web. 01 December 2016.
18. Neighborhood Scout. "The leading all-in-one real estate market data platform in the USA." *NeighborhoodScout*. N.p., n.d. Web. 05 May 2017.
19. Philadelphia Police Department. *Crime Incidents*. OpenDataPhilly, 2016. Web. 01 December 2016.
20. Philadelphia Streets Department. *Street Centerlines*. OpenDataPhilly, 2017. Web. 01 February 2017.
21. Portland Police Bureau. *Crime Statistics*. Portland Police Bureau, 2016. Web. 01 December 2016.
22. Portland Maps. *Street Centerlines*. Portland Open Data, 2016. Web. 01 December 2016.
23. San Francisco Department of Technology. *San Francisco Basemap Street Centerlines*. DataDF, 2016. Web. 01 December 2016.
24. San Francisco Police Department. *Police Department Incidents*. DataSF, 2016. Web. 01 December 2016.
25. Seattle Police Department. *Seattle Police Department Police Report Incident*. City of Seattle Open Data Portal, 2016. Web. 01 December 2016.
26. Seattle Public Utilities. *Street Network Database.* City of Seattle Open Data Portal, 2016. Web. 01 December 2016.
27. United States Census Bureau. *Census Data – Selected Socioeconomic Indicators in Chicago, 2008 - 2012.* Chicago Data Portal, 2016. Web. 01 December 2016.
28. United States Census Bureau - American FactFinder. *QT-P1 – Age Groups and Sex: 2010.* 2010 Census. U.S. Census Bureau, 2010. Web. 01 March 2017.