INTO THE CITY

A Multi-Disciplinary Investigation of Urban Life

MARCO DE NADAI

Department of Information Engineering and Computer Science
Università degli studi di Trento

CYCLE XXXI
2019

This dissertation is submitted for
the degree of Doctor of Philosophy

Marco De Nadai: *Into the city,* © January 2019

SUPERVISORS:
Bruno Lepri
Nicu Sebe

EXAMINERS:
Cecilia Mascolo
Dino Pedreschi
Andrea Passerini

ABSTRACT

Cities are essential crucibles for innovation, novelty, economic prosperity and diversity. They are not a mere reflection of individual characteristics, but instead the result of a complex interaction between people and space. Yet, little is known about this self-organized and complex relationship. Traditional approaches have either used surveys to explain in detail how a few individuals experience bits of a city, or considered cities as a whole from their outputs (e.g. total crimes). This tide has however tuned in recent years: the availability of new sources of data have allowed to observe, describe, and predict human behaviour in cities at an unprecedented scale and detail.

This thesis adopts a "data mining approach" where we study urban spaces combining new sources of automatically collected data and novel statistical methods. Particularly, we focus on the relationship between the built environment, described by census information, geographical data, and images, and human behaviour proxied by extracted from mobile phone traces. The contribution of our thesis is two-fold. First, we present novel methods to describe urban vitality, by collecting and combining heterogeneous data sources. Second, we show that, by studying the built environment in conjunction with human behaviour, we can reliably estimate the effect of neighbourhood characteristics, predict housing prices and crime. Our results are relevant to researchers within a broad range of fields, from computer science to urban-planning and criminology, as well as to policymakers.

The thesis is structured in two parts. In the first part, we investigate what creates urban life. We *operationalize* the theory of Jane Jacobs, one of the most famous authors in urban planning, who suggested that the built environment and vitality are intimately connected. Using Web and Open data to describe neighbourhoods, and mobile phone records to proxy urban vitality, we show that it is possible to predict vitality from the built environment, thus confirming Jacob's theory. Also, we investigate the effect of safety perception on urban vitality by introducing a novel Deep Learning model that relies on street-view images to extract security perception. Our model describes how perception modulates the relative presence of females, elderly and younger people in urban spaces. Altogether, we demonstrate how objective and subjective characteristics describe urban life.

In the second part of this dissertation, we outline two studies that stress the importance of considering, at the same time, multiple factors to describe cities. First, we build a predictive model showing that objective and subjective neighbourhood features drive more than 20% of the housing price. Second, we describe the effect played by a neighbourhood's characteristics on the presence of crime. We present a Bayesian method to compare two alternative criminology theory, and show that the best description is achieved by considering together socio-economic characteristics, built-environment, and mobility of people. Also, we highlight the limitations of transferring what we learn from one city to another.

Our findings show that new sources of data, automatically sensed from the environment, can complement the lengthy and costly survey-based collection of urban data and reliably describe neighbourhoods at an unprecedented scale and breath. We anticipate that our results will be of interest to researchers in computer science, urban planning, sociology and more broadly, computational social science.

PUBLICATIONS

This thesis is based on the following papers:

[1] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. In *WWW '16*, 2016.

[2] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *MM '16*. ACM, 2016.

[3] Marco De Nadai and Bruno Lepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment. In *DSAA '18*, 2018.

[4] Marco De Nadai, Yanyan Xu, Emmanuel Letouzé, Marta C. González, and Bruno Lepri. Socio-economic, built environment, and mobility conditions associated with crime: A study of multiple cities. *Under submission to Nature Human Behaviour*, 2019.

This Ph.D. was instrumental to study other topics, which I chose not to include in this manuscript:

[1] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2015.

[2] Simone Centellegher, Marco De Nadai, Michele Caraviello, Chiara Leonardi, Michele Vescovi, Yusi Ramadian, Nuria Oliver, Fabio Pianesi, Alex Pentland, Fabrizio Antonelli, and Bruno Lepri. The mobile territorial lab: A multilayered and dynamic view on parents' daily lives. *EPJ Data Science*, 5(3), 2016.

[3] Marco Mamei, Francesca Pancotto, Marco De Nadai, Bruno Lepri, Michele Vescovi, Franco Zambonelli, and Alex Pentland. Is social capital associated with synchronization in human communication? an analysis of italian call records and measures of civic engagement. *EPJ Data Science*, 7(1):25, Jul 2018.

[4] Marco De Nadai, Angelo Cardoso, Antonio Lima, Bruno Lepri, and Nuria Oliver. Apps, places and people: strategies, limitations and trade-offs in the physical and digital worlds. *Under review in Scientific Reports*, 2019.

[5] Emanuele Strano, Filippo Simini, Marco De Nadai, Thomas Esch, and Mattia Marconcini. Precise mapping, density and spatial structure of all human settlements on earth. *Under submission to Nature Communications*, 2018.

[6] Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, and Bruno Lepri. Gesture-to-gesture translation in the wild via category-independent conditional maps. *Under review in ACM MM '19*, 2019.

CONTENTS

Part I

INTRODUCTION

We begin our narrative by explaining the importance of cities, and their effect. We then briefly outline the history of urban studies, to then move to more recent quantitative research based on new sources of data. Finally, we succinctly present the research questions and contributions that will become clearer in the next parts.

1

STUDYING CITIES, STUDYING PEOPLE

Cities have the capability of providing something for everybody; only because, and only when, they are created by everybody.

— Jane Jacobs [10]

For most of human history, there were no cities. At first, our ancestors lived either as nomads or in small villages. Just five thousand years ago, once more complex societies emerged, cities have become the primary attraction of human life for both economic and political reasons [11]. Now, cities are perhaps the ultimate way for humans to colonize Earth [12].

People aggregation is the cause and effect of cities. When individuals live together in a settlement, their social interactions grow exponentially, resulting in a systematic acceleration of urban life [13]. As cities get larger, they get wealthier, more innovative, tolerant, but also unequal, chaotic and affected by crime [14–16]. In search for prosperity, novelty and jobs, people have always migrated to urban areas, generating an increasing complexity in managing, organizing and understanding cities [17]. For this reason, throughout history, cities have been studied, planned and imagined to find the ingredients for the perfect city.

The first discussions on the ideal city can be traced back to the Renaissance, when an orderly society based on a perfectly planned and geometric balanced city was imagined [18]. At first, cities were described as *systems*, defined and organized entities composed of elements and interactions in a state of equilibrium. Planning and management processes were the tools believed to control the rigid hierarchical structure of urban areas [19]. However, cities are not in equilibrium, as they are forever changing. Nor they are rigid on their hierarchy; they are emergent phenomena composed of millions of individuals constantly deciding, moving, evolving. Thus, in the late 20th century scholars started thinking of cities as a living *organism* rather than static *machine*. This analogy was indeed the appropriate stratagem to formalize cities. In biology the enormous diversity of animal kingdom can be characterized by simple mathematical equations that tie together the metabolic rate to the mass of each animal in a predictable way. Surprising as it may seem, cities are not different and the pace of urban life is predicted to increase with size. As a city's population doubles the Gross domestic product (GDP) grows by almost 127% while requiring only 85% of new infrastructure [14]. Similarly, as a city grows, the innovation almost triplicates. By adopting this new perspective, researchers could for the first time describe emergent phenomena such as income, streets, flu outbreaks, and homicides using very simple equations [14, 20, 21]. The recipe for the perfect city seemed at hand.

The existing framework, however, considered cities as a whole, neglecting the fact that extremely diverse areas often coexist within the same city. Cities have always been highly spatial and ethnic segregated [22, 23], economically unequal [24–26], and unevenly plagued by crime [27]. These differences, at the level of the neighbourhood, mediate social life. People are indeed not just subjected to these local patterns, they react to them, and these reactions shape the perception, social interactions and behaviour beyond the boundary of neighbourhoods [28]. Thus, researchers from a variety of disciplines have studied the diversity present within cities, suggesting a range of theories and approaches. An abundant literature has long studied the sociological *neighbourhood effect* [29] on crime, showing through the *Social disorganization theory* that the lack of communication and cooperation between neighbours hinders the realization of common values and the maintenance of the social controls to promote public order [30–35]. Similarly, urban planners and activists, such as Jane Jacobs, have stressed the importance of the physical environment (e.g., sidewalks, parks, *walkability*) and its direct effect on social life [10, 36, 37]. Others, focused on the effects of disadvantaged districts on individual's outcomes such as drug use, low birthweight, and cognitive ability [38, 39]. Until recently, however, theoretical advancements were limited by the high cost and sampling bias of census and surveys data.

This is why the advent of new data sources and methods has acted as a real game-changer in urban studies. The *digitalization* of society has allowed to observe, understand and even predict many aspects of human mobility and social behaviour [40, 41]. Scholars have shown that individual mobility is recurrent [42, 43], heterogeneous [44, 45], and characterized by a set of few favourite places [46], which are tightly coupled to individuals' social relations [47, 48]. Collective mobility traces well describe the rhythm of a city [49–51], the different neighbourhoods [52], the temporal dynamics [49] and the social networks of places [53, 54], but also predict socio-economic indicators such as economic development [55, 56], deprivation [57, 58], and crime [59, 60]. Concurrently, the availability of online geographical data and crowd-sourced experiments has allowed to describe areas and test urban theories at an unprecedented scale. Cities and neighbourhoods' physical environment are now extensively mapped through OpenStreetMap [61]. Street-level imagery has been linked to human perception [62, 63], urban change [64] and crime [65], while social media data has been used to describe the Point Of Interests (POIs) needs [66], socio-economic indexes [67], tourists attractiveness [68], happiness [69] and even urban smell [70]. Altogether, scholars have shown that mobile communication records and geographical data provide a reliable proxy for human behaviour and urban spaces. These concepts, however are not independent. Nor it is clear what creates and influence urban life at the level of the neighbourhood.

This thesis investigates urban life, providing extensive results that describe how the neighbourhood characteristics influence human behaviour, specifically perception, crime, and the presence of people. Our work is best placed in an emerging interdisciplinary field called "urban computing" [71], which combines computer science approaches with more traditional fields like urban planning, urban economy, and urban sociology. First, we describe urban areas, particularly neighbourhoods, their built environment, and socio-economic conditions, relying on public and commercial data. Then, we measure human behaviour, using mobile phone records. Our contribution is two-fold. First, we present novel methods to collect and combine heterogeneous sources of data and describe urban vitality. Secondly, we show that by studying the built environment in conjunction with other socio-economic aspects of human behavior, we can reliably estimate the role played by differences at the level of the neighbourhood and predict housing prices and crime. This dissertation is based on statistics and data mining, as tools to find and study emergent properties of cities. We develop new models that aims at extracting predictive information from geographical data, high-level human perception from images, and tools to analyse cities at scale. We anticipate that our methodology and results promise to have a great multi-disciplinary impact especially in urban studies, where the neighbourhood effect has been long studied using traditional data. To help readers following the line of research, we describe below the four research questions that underlie this thesis.

1.1 RESEARCH ISSUES AND CONTRIBUTIONS

What are the objective characteristics that thrives urban life?

Understanding the conditions that promote urban vitality is fundamental in order to design better cities. In her influential book "The Death and Life of Great American Cities" [10], written in 1961, Jane Jacobs proposed four conditions that promote life in a city. Supposedly, neighbourhoods characterized by a mix of land uses, small blocks, high density, and diversity of buildings create virtuous processes that increase the diversity of people and, thus, urban life. However, these conditions were not empirically tested until recently – mainly due to the lack of suitable data about "city life". Recently, the city of Seoul collected pedestrian activity through surveys at an unprecedented scale, with an effort spanning more than a decade, allowing researchers to conduct the first study testing Jacobs's conditions [72].

In Chapter 2, we test at scale Jacobs' theory. We contribute to the field by identifying a valuable alternative to the lengthy and costly collection of activity survey data: mobile phone data. Also we collect land use and socio-demographic information from the Italian Census and Open Street Map, and test the four conditions in six Italian cities. Although these cities are very different from the places for which Jacobs's conditions were spelled out (i.e., great American cities) and from the places in which they were recently tested (i.e., the Asian city of Seoul), we find those conditions to be indeed associated with urban life in Italy as well.

The chapter is based on research published in De Nadai *et al.* [1].

How does human perception influence vitality?

Policy makers, urban planners, architects, sociologists, and economists are interested in creating urban areas that are both lively and safe [10, 32, 34, 73–77]. But are safety and liveliness of neighbourhoods two independent characteristics? Or are they just two sides of the same coin?

In Chapter 3, we contribute to the field by connecting the levels of human activity with the neighbourhood's safety perception through automatically collected data. We combine mobile phone data (as a proxy for activity or liveliness) with scores of perceived safety estimated using a Convolutional Neural Network trained on a dataset of Google Street View images, scored using a crowdsourced visual perception survey [65]. We find that: (i) safer looking neighbourhoods are more active than what is expected from their population density, employee density, and distance to the city centre; and (ii) that the correlation between appearance of safety and activity is positive, strong, and significant, for females and people over 50, but negative for people under 30, suggesting that the behavioral impact of perception depends on the demographics of the population. Finally, we use occlusion techniques to identify the urban features that contribute to the appearance of safety, finding that greenery and street-facing windows contribute to a positive appearance of safety (in agreement with Oscar Newman's *defensible space* theory [74]).

The chapter is based on research published in De Nadai *et al.* [2].

Is the housing value influenced by the neighbourhood?

Housing costs have a significant impact on individuals, families, businesses, and governments. Estimating housing value is a challenging multi-dimensional problem that involves the appraisal of many facets of a property, its neighbourhood, and its city. In recent years, new sources of data and new methods have allowed to estimate property values from millions of similar recently sold properties [78–80]. However, the role of neighbourhood characteristics on housing value remains an open question.

Chapter 4 builds upon previously documented contributions of some neighbourhood's factors to housing price [81–85], testing them almost one at a time. Our contribution is three-fold. First, we study these factors together shading light on the role played by the neighbourhood's characteristics such as *walkability* and security perception. Secondly, we relax the linearity assumptions of previous studies. Finally, we formalize this problem in a data mining setting and now-cast housing prices from Open data, without the need for historical transactions. Experiments involving 70,000 houses in 8 Italian cities highlight that the neighbourhood's *vitality* and *walkability* seem to drive more than 20% of the housing value. Moreover, the use of this information improves the now-cast by 60%.

The chapter is based on research published in De Nadai *et al.* [8].

How can crime be explained by the neighbourhood? Is this knowledge transferable?

According to the *Social disorganization theory*, the inability of a community to realize the common values of its residents creates the ecological conditions favorable to crime. In criminology, studies have documented that weakened friendships, poverty, heterogeneity, and rapid population turnover are often linked to a weakened community's values and, thus, higher crimes in a neighbourhood [31–35]. In urban planning, the concept of local organization, care, and common values are believed to be a virtuous consequence of the neighbourhood's built environment [10, 36, 37].. However, neighbourhoods are not to be considered islands unto themselves, as they are embedded in a city-wide system of social interactions. On a daily basis, the people's routine exposes residents to different conditions, possibilities [86] and neighbourhoods. The existing literature, however, has focused on a single city at a time, considering only a limited number of factors (such as only socio-economical characteristics). Hence, our understanding of the factors influencing crime across cultures and cities is very limited.

In Chapter 5, we contribute to the field by exploring how crime is related not only to socio-economic factors but also to the built environmental (e.g. land use) and mobility characteristics of neighbourhoods. To that end, we integrate multiple open data sources with mobile phone traces in diverse cities. We find that the mobility information and physical characteristics of the neighbourhood effectively explain the emergence of crime, and improve the performance of the traditional approaches. Moreover, we show that the ecological factors of neighbourhoods relate to crime very differently from one city to another.

The chapter is based on research published in De Nadai *et al.* [4].

1.2 OUTLINE

This thesis is organized in two parts, one devoted to vitality and one to the outcome of neighbourhoods. In Part ii, we first (Chapter 2) present the objective and physical urban characteristics of neighbourhoods that influence urban vitality. Here, we operationalize the Jane Jacobs' theory through commercial, open and mobile phone data. We end by analyzing how perception influence vitality, through the eye of the safety perception extracted from street-level images (Chapter 3). Part iii concerns the research questions that are directly linked to the outcome of the city. We depart from Chapter 4) analyzing tens of thousands houses and their neighbourhoods to find the *neighbourhood effect* on housing price. Then, Chapter 5 shows how socio-economic characteristics, the built environment and the people's mobility can, together, explain crime. We also show the intrinsic limits in transferring the knowledge from one city to the other.

Finally, Part iv ties everything together by highlighting the learned lesson and foreseeing potential new directions.

1.3 MISCELLANEOUS NOTES

1.3.1 *Reproducibility*

Valuable research produces models that can be applied to new data and reproduce results. With this in mind, we released the code, intermediate data and instructions for all the articles presented in this manuscript. Data and code are freely available online in permanent repositories.

To emphasize the development of new methods and analysis, we also released a multi-source dataset of mobile phone data, social network conversations and weather information. The data is not extensively presented here, but it can be found in Barlacchi *et al.* [5].

For convenience, we here list the online repositories of the publications:

- *The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective*[1];
- *Are Safer Looking neighbourhoods More Lively? A Multimodal Investigation into Urban Life*[2];
- *The economic value of neighbourhoods: Predicting real estate prices from the urban environment*[3];
- *Socio-economic, built environment, and mobility conditions associated with crime: A study of multiple cities*[4].

1.3.2 *Tools*

The research presented in this manuscript was supported by free software such as Python 3.5 [5], PostgreSQL 9.4 [6], PostGIS 2.4 [7], QGIS[8] and OSRM[9]. However, many Python packages were used to analyze the data. Notably, Pandas, Matplotlib, Jupyter, and PyMC3 [87], but also Apache Spark[10].

This document was typeset using TeXstudio and LaTeX. The template used is the typographical look-and-feel `classicthesis` developed by André Miede[11], and modified by Rémi Louf[12].

1 Available at https://github.com/denadai2/jacobs_urban_planning
2 Available at https://github.com/denadai2/google_street_view_deep_neural
3 Available at https://github.com/denadai2/real-estate-neighbourhood-prediction
4 Available at https://github.com/denadai2/crime-multi-city
5 Available at http://www.python.org
6 Available at https://www.postgresql.org
7 Available at https://postgis.net
8 Available at https://www.qgis.org
9 Available at http://project-osrm.org
10 Available at https://spark.apache.org
11 Available at http://code.google.com/p/classicthesis/.
12 Available at https://github.com/rlouf/phd-thesis.

Part II

VITALITY

Diverse cities enable convenient social interactions, face-to-face encounters, and a spontaneous sense of community that spurs economy, security and urban life. This idea does not come from empirical evidences or long-term studies on communities. Instead, it is drawn from the common sense and observation of Jane Jacobs, who wrote the book "The Death and Life of Great American Cities" in 1961. Many scholars have theoretically challenged this book. However, the lack of data have hindered empirical validations of both the original and alternative theories.

In this part we first analyse how Jane Jacobs' theory can be empirically verified through geographical and passively collected data. Thus, we analyse how objective characteristics of neighbourhoods are correlated with presence of people. Finally, we present a model that analyse how subjective features, automatically extracted from street-level imagery, might influence the presence of people.

2

THE CONDITIONS FOR A VITAL PLACE

To seek "causes" of poverty in this way
is to enter an intellectual dead end
because poverty has no causes.
Only prosperity has causes.

— Jane Jacobs [10]

A fundamental research question that urban planners, sociologists and economists are investigating is what creates urban life. The urban sociologist Jane Jacobs in *The Death and Life of Great American Cities* [10], one of the most influential books in city planning, introduced the urban physical environment (the *urban fabric*) as an essential factor for urban vitality [10]. As the book title says, Jacobs dealt with the death and life of American cities. In its most basic form, she argued that death was caused by the elimination of pedestrian activity (e.g., by highway construction, large-scale development projects), and that life was created by the presence of pedestrians at all times of the day.

She argued that, to promote urban life in large cities, the physical environment should be characterized by *diversity* at both the district and street level. Diversity, in turn, requires four essential conditions: (i) *mixed land uses*, that is, districts should serve more than two primary functions, and that would attract people who have different purposes; (ii) *small blocks*, which promote contact opportunities among people; (iii) *buildings diverse in terms of age and form*, which make it possible to mix high-rent and low-rent tenants; and (iv) sufficient *dense concentration* of people and buildings.

Despite their importance, those conditions have not been empirically tested all together until recently, mainly because it is hard to collect data about which neighbourhoods have full urban life and which have little of it. After an effort lasting more than a decade, the city of Seoul recently managed to collect pedestrian activity through surveys, and local researchers performed the first study testing Jacobs's conditions in the city [72]. The researchers found her claims to hold in Seoul too: mixed-use, old buildings, and density all contributed to urban life.

In this chapter, we put forward the use of mobile phone data to complement the lengthy and costly survey-based collection of activity data. We extract human activity measurements from mobile phone records in six Italian cities - Bologna, Florence, Milan, Palermo, Rome, and Turin. These cities are very different from the places in which Jacobs's conditions were spelt out (i.e., great American cities) and from the places in which they were recently tested (i.e., the Asian city of Seoul). Despite that, our results show that the four conditions for promoting urban life do hold for Italian cities as well.

The chapter is based upon De Nadai *et al.* [1] and it is organized as follows. Section 2.1 reviews the relevant literature; Section 2.2 presents the Jane Jacobs' theory; Section 2.3 describes the datasets and the methods; and in Section 2.4 we provide a comprehensive assessment of the predictive power of Jacobs's diversity conditions in the Italian context. In Section 2.5 we identify our work's theoretical and practical implications, and its limitations.

2.1 STATE OF THE ART

The idea of testing urban theories using novel sources of data (e.g., social media, online images and videos, mobile phone data) has received increasing attention [62, 88, 89]. The urban sociologist Kevin Lynch showed that people living in an urban environment create their own personal "mental map" of the city based on features such as the areas they visit and the routes they use [90]. Hence, he hypothesized that, the more recognizable a city, the more navigable the city. To test Lynch's theory, Quercia *et al.* built a web game that crowd-sourced Londoners' mental images of the city [88]. They showed that areas suffering from social problems such as poor living conditions and crime are rarely present in residents' mental images.

Researchers also investigated which urban elements people use to visually judge a street to be safe, wealthy, and attractive using web crowd-sourcing games [62, 89, 91], and they also studied how to identify walkable streets using the social media data of Flickr and Foursquare (e.g., unsafe streets tended to be photographed during the day, while walkable streets were tagged with walkability-related keywords [92]).

The recent availability of large-scale data sets, such as those automatically collected by mobile phone networks, opens new possibilities for studying city dynamics at finer and unprecedented granularity [40]. Mobile phone data represents a highly valuable proxy for human mobility patterns [41, 45, 93]. Such data was recently used to map functional uses [94, 95], to identify places that play a major role in the life of citizens [96], to compare cities based on their spatial similarities and differences [49], and to predict socio-economic indicators [55, 57], including crime [59, 60, 97].

Researchers recently tested Jane Jacobs's observations about crime [59, 60, 97]. They verified her *natural surveillance* theory. She posited that people diversity and presence of visitors in a neighborhood contribute to natural surveillance and, as such, discourage crime: as people are moving around an area, they effectively become "eyes on the street" able to observe what is going on around them.

To go beyond crime, we set out to conduct a comprehensive validation of Jacobs's four conditions for urban life. To do that, we gather several sets of data.

2.2 DEATH AND LIFE OF CITIES

In Jane Jacobs' words, "a well-used city street is apt to be a safe street and a deserted city street is apt to be unsafe" [10]. This means that streets must be used relatively continuously during the day, both by residents and by strangers. These "eyes on streets" generate a virtuous loop which, in turn, increases public safety, creates diverse face-to-face interactions, and contributes to the local economy.

Jacobs spelt out four diversity requirements essential for the generation and maintenance of urban life. First, a district should serve more than one primary function, preferably more than two. This attracts people with diverse purposes who end up sharing common facilities. Moreover, when primary uses are combined in a way that attracts people at different times of the day, then that positively impacts the local economy. For instance, a neighbourhood where people are present only during office hours is likely to provide only a few, if any, leisure facilities: it would be neither efficient nor attractive for residents. Conversely, should a secondary function flourish by, e.g., the presence of theatres, museums or night-life places, residents would likely benefit.

Second, Jacobs argued that street blocks must be short. She observed that large rectangular blocks are apt to thwart the effective mixture of use and people, resulting in paths that meet each other too infrequently. By contrast, smaller blocks provide more intersections and thus slow down automobiles, which, she argued, are inclined to destroy urban vitality by discouraging the presence of pedestrian activity.

Third, in Jacobs's view, the buildings in a district should be diverse in terms of age and form. This ensures the diversity of economic activity. Diverse buildings make it possible to have a mixture of jobs (i.e., high-, medium- and moderate-income jobs) and a mixture of tenants (e.g., high-rent and low-rent tenants) [10, 98–100]. Her view came from observing large-scale projects in the New York of her time: she saw that large-scale buildings did not change over time and hardly adapted to the environment that surrounded them; by contrast, old buildings helped to cultivate new primary uses in the neighbourhood (e.g., new companies were likely to start and initially grow in old and low-rent buildings).

The fourth and final condition is about a dense concentration of people and buildings. The idea is that density fosters "lively" district that can attract people for different purposes.

Jacobs emphasized that *all* four factors are necessary: density alone cannot create urban diversity, and mixed-land use would not flourish in areas with big blocks and of low density.

Further, in addition to those four conditions, Jacobs talked about 'vacuums', which are patches of land dedicated to a single use (e.g., transport facilities, parks). Those elements might be good or bad. Small railway stations, bus stops and small parks may act as hubs for the pedestrian activity; but, at the same time, if not well managed, a park could be used only at a particular time of the day, exposing it to deprivation and criminality.

2.3 OUR APPROACH

Our approach leverages mobile Internet activity from call data records to extract proxies for *urban vitality*, and web data from public entities (e.g., national census) and commercial ones (e.g., Foursquare) to extract *urban diversity* as per Jacobs's four conditions. Our goal is to study the relationship between urban vitality and diversity. Next, we spell out the metrics and the regression models we used to meet that goal.

In the 2011 Italian census, the smallest administrative unit of cities is represented by the census sections (*sezioni di censimento*), also called *blocks*. These sections are delimited by street segments and are grouped with other nearby sections to form a census area (*area di censimento*) called *district*. The boundaries of a district are drawn on the basis of socio-economic conditions in a way that districts can be compared with each other in terms of, for example, population: districts have between 13,000 and 18,000 inhabitants[1]. This is a loose constraint, however: city administrations may ignore it and define their own districts. Empirically, we verify that, in our data, a district covers an average area of 2.47 km^2 with an average population density of 10,000 people per km^2 (Table 1). Jacobs did not define any strict criterion concerning district size: she simply proposed that the edge of an administrative district should not exceed

1 We exclude districts that are not densely populated (e.g. the national park in Rome, the volcano in Naples) These are marked by the Italian census with an identifier `ace` equal to zero.

City	#Districts	Size (avg)	Population (avg)
Bologna	23	3.34	15,918
Florence	21	2.89	16,633
Milan	85	1.72	14,551
Palermo	43	2.01	15,075
Rome	146	3.24	17,312
Turin	56	2.00	15,543

Table 1: Number, average size (in km^2), and population of the districts in our six cities.

2.4 km, and that each district should have a minimum population of 50,000 people. Compared to that definition, our districts are similar in terms of area and smaller in terms of population.

2.3.1 *Sets of data*

Mobile Phone Activity. Every time a mobile phone communicates, a radio base station delivers the communication through the network, and a new call data record is created. This record reports the time of the communication, and the radio base station that handled it. Different types of call records exist, recording various types of communication (e.g., incoming calls, Internet, outgoing SMS). We decided to focus on Internet activity, since it allows for the passive reconstruction of people's mobility: even in the absence of direct user activity (e.g., making/receiving a call, receiving/sending an SMS), mobile phones can be tracked since they are likely to be connected to the Internet for background traffic and push notifications.

Call records have been provided by the Semantics and Knowledge Innovation Lab of Telecom Italia Mobile, which is the largest mobile operator in Italy with 34% of the entire market share[2]. Our data is aggregated every 60 minutes, comes from both TIM customers and roaming customers in the six cities, and covers the time ranging from February to October 2014.

OpenStreetMap. OpenStreetMap[3] (OSM) is a global project that aims to collaboratively create a detailed map of the world. With more than 2.1 million contributors, it has become the most valuable and openly accessible source of

2 http://bit.ly/1LtNrFY
3 http://www.openstreetmap.com

ISTAT code	Description
P1	number of residents
E2	number of used buildings
E3	number of residential buildings
E8-E16	number of buildings in 9 age bands in the range [1919, 2011]
E17-E20	number of buildings of 4 types based on their number of floors
E21-E26	number of buildings of 6 types based on their internal apartments
ADDETTI	number of employees

Table 2: ISTAT open data variables under study. The last variable comes from the census of industry and services (*censimento dell'industria e dei servizi*), while the rest come from the census of population and housing (*censimento della popolazione e delle abitazioni*).

information about the physical world. OSM data is composed by three data primitives: *nodes*, *ways* and *relations*. *Nodes* define points in space like Point Of Interests (which we call "places"), *ways* define roads and other linear features, and *relations* reflect how those elements are related to each other. Since we focused on the geographic unit of a *block* (i.e., the smallest area that is surrounded by street segments), we filtered out suburban elements: *ways* with tag `highways` using Overpass API[4], `footpaths`, `primary` roads, and `proposed` roads.

Census Data. The Italian National Institute for Statistics (ISTAT) provides, under open data licences, information gathered via the 2011 Italian Census[5]. We used the census' variables related to people and buildings (Table 2), which were defined at various geographic levels.

Land Use. Urban land-cover and land-use for Large Urban Zones (LUZs) is mapped by the Urban ATLAS European project[6]. This project exploits satellite images to categorize the city in 20 classes (e.g., agricultural areas, continuous urban fabric) with a precision between 0.25 and 1 hectares, and accuracy above 80%. The dataset is built and validated for 2006, while its version updated to 2012 will be released at the end of 2015. However, for Turin and Rome, the new data is already available, and we used it.

4 http://overpass-turbo.eu/
5 http://www.istat.it/it/archivio/104317
6 http://www.eea.europa.eu/data-and-maps/data/urban-atlas

Infrastructures. In addition to the Urban ATLAS dataset, we used the ISTAT statistical atlas of infrastructures (*Atlante statistico delle infrastrutture*)[7], which provides details on the logistic facilities (e.g., presence of railways) in Italy.

Foursquare Data. Created in 2008, Foursquare is the world leading location-based social network. It has attracted more than 30 million users[8]. Through a gamification system, users are encouraged to *check-in* their geographical position into the social network and sharing the places they visit (*venues*). Venues are places whose geographical information is enriched with semantic labels specifying details about them. Three hierarchical levels for place categories exist in Foursquare: an abstract level grouping *venues* in macro categories (e.g., “food"); a more specific one (e.g., “restaurant"); and a third highly detailed level (e.g., “Thai restaurant")[9]. We extracted Foursquare *venues* from the public API[10] for the categories in Table 3.

Group	Foursquare category
NightLife	Brewery, Champagne Bar, Cocktail Bar, Dive Bar, Gay Bar, Hookah Bar, Lounge, Night Market, Night club, Other Nightlife, Pub, Strip Club, Whisky Bar, Wine bar, Nightlife Spot
Art-night	Laser Tag, Movie Theater, Music Venue, Performing Arts
Services	Medical Center, Library, Government Building, Military Base, Post Office
Eating-drinking	Food
Org. activity	Comedy Club, Country Dance Club, Salsa Club, Club, Community Center, Cultural Center, Library, Social Club, Spiritual
Outside	Park, Plaza, Pedestrian Plaza
Commercial	Shop & Service (excluding ATM, Construction, EV Charging, Gas Station / Garage, Newsstand)

Table 3: The Foursquare categories of the venues under study.

7 http://www.istat.it/it/archivio/41899
8 http://bit.ly/1VNf4jQ
9 https://developer.foursquare.com/categorytree
10 https://api.foursquare.com

2.3.2 *Jacobs's metrics*

We extracted several variables to quantify the four conditions (Table 4), and we will detail them next.

2.3.2.1 *Land use*

The first of Jacobs's four conditions is to have mixed primary uses in a district [10]. Primary categories include residential buildings, offices, industrial facilities, entertainment places, education facilities, recreation facilities, museums, libraries, and galleries.

We computed Land Use Mix (LUM) [101] in district i as:

$$\text{LUM}_i = -\sum_{j=1}^{n} \frac{P_{i,j}\log(P_{i,j})}{\log(n)} \tag{1}$$

where $P_{i,j}$ is the percentage of square footage with land use j in district i, and n is the number of possible land uses (in our case, $n = 3$). If district i's land is dedicated to one use only, then LUM_i is zero; instead, if the land is used equally in all n ways, then LUM_i is one. The higher LUM_i, the more mixed i's land use. We defined the three land uses following the recommendations in [102]: the first land use is "residential"; the second includes the categories "commercial", "institutional and governmental", and "resource and industrial"; and the third includes "park and recreational" and "water".

Since well-managed parks might well function as hubs for pedestrian activity, we defined the average distance from the nearest small park (area smaller than 1 km^2) for each district i:

$$\text{Closeness to SM}_i = \Big(\frac{1}{|B_i|}\sum_{j\in B_i} \text{dist}(j, \text{closest}(j, SM))\Big)^{-1} \tag{2}$$

where B_i is the set of blocks in district i, closest(j, Y) is a function that finds the geographically closest element in set Y from block j's centroid, SM is the set of small parks, and dist(a, b) is the geographic distance between two elements' centroids a and b.

We also computed the Residential/Non-Residential (RNR) balance in district i as:

$$\text{RNR}_i = 1 - \Big|\frac{Res_i - NonRes_i}{Res_i + NonRes_i}\Big| \tag{3}$$

where Res_i is the area occupied by residential buildings in district i, and $NonRes_i$ is the area occupied by non-residential ones. The higher RNR_i, the more balanced the district in terms of residential *vs.* non-residential uses.

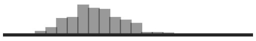
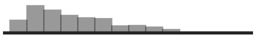
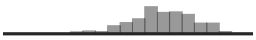
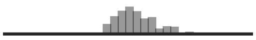
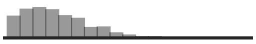
Land use	**Distribution**	μ
(1) Land use mix		0.73
(2) Closeness to small parks (SPs)		1×10^{-3}
(3) Residential *vs.* Non-Res.		0.67
(4) Housing types		4.98
(5) Commercial		0.30
(6) Nightlife		0.10
(7) Nightlife density		1×10^{-5}
(8) Daily		0.02
(9) 3^{rd} Places		1×10^{-4}
Small blocks		
(10) Block area		9.61
(11) Intersections density		1×10^{-4}
(12) Anisotropicity		0.38
Aged buildings		
(14) $\overline{\text{building age}}$		50.34
(16) $\sigma_{\text{building age}}$		12.96
(17) $\overline{\text{Employees per company}}$		6.98
Concentration		
(18) Population density		0.01
(19) Employment density		5×10^{-3}
(20) $\frac{\text{population density}}{\text{employee density}}$		3.72
(21) $\frac{\lvert\text{internal apartments}\rvert}{\lvert\text{buildings}\rvert}$		15.31
(22) Density of daily places		5×10^{-3}
(23) Density non-daily places		3×10^{-3}
Vacuums		
(24) Closeness to large parks		3×10^{-3}
(25) Closeness to railways		1×10^{-3}
(26) Closeness to highways		6×10^{-4}
(27) Closeness to water		1×10^{-3}
Mobile phone activity		
(28) Activity density		3.84

Table 4: Urban diversity metrics plus mobile Internet activity. Most variables are not normally distributed. Density measures are computed with surface area in m^2, and closeness measures are the inverse of the distance (1/m).

To go beyond horizontal land use and look at vertical development, we computed the average number of floors per building in district i and called it 'housing types' (as [72] did):

$$\text{Housing types}_i = \frac{\sum_c h_{c,i} \cdot z_c}{\sum_c h_{c,i}} \tag{4}$$

where $h_{c,i}$ is the number of buildings that are in height category c in district i, and z_c is the number of floors corresponding to height category c. The sums were repeated over all height categories (i.e., over the categories `E17-E20` in Table 2).

The previous definitions have characterized spatial use in terms of land use and building use. However, activities are important too. Jacobs argued for mixing primary uses so that people are on the street at different times of the day. To characterize spatial use in terms of activities, upon Foursquare data, we determined whether each place is used daily (e.g., convenience stores, restaurants, sport facilities) or not, and whether it is used at night or daylight [72]. Based on that, we defined:

$$\text{Commercial}_i = \frac{|\text{non daily-use places}_i|}{|\text{places}_i|} \tag{5}$$

$$\text{Nightlife}_i = \frac{|\text{nightlife places}_i|}{|\text{places}_i|} \tag{6}$$

$$\text{Nightlife density}_i = \frac{|\text{nightlife places}_i|}{\text{area}(N_i)} \tag{7}$$

$$\text{Daily}_i = \left(\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, D))\right)^{-1} \tag{8}$$

where $\text{area}(N_i)$ is the surface area of neighbourhood N_i, D is the set of places that are used daily, and B_i is, again, the set of street blocks in district i.

Also, not all activities are equal. Some activities are more 'social' than others. To capture this, we resorted to the concept of *third places*. These are defined by Oldeburg [103] as the "great, good places" that foster community and communication among people outside home (the first place) and work (the second place): "they are places where people gather primarily to enjoy each others' company". Third Places function as unique public spaces for social interaction, providing a context for sociability, spontaneity, community building and emotional expressiveness [104]. Therefore, we computed:

$$3^{rd} places_i = \frac{|3^{rd}\text{places}_i|}{|\text{places}_i|} \tag{9}$$

We determined whether a place is a third place or not by following the 4-category classification proposed by Jeffres *et al.* [105]. Third places fall into these four categories: *eating, drinking and talking* (e.g., coffee shops, bars, pubs, restaurants, and cafes); *organized activities* contributing to social capital [106] (e.g., places of worship, clubs, organizations, community, and senior centres); *outdoor* (e.g., plazas and parks); and *commercial venues* (e.g., stores, malls, shopping centres, markets, beauty salons, and barber shops).

2.3.2.2 *Small blocks*

Jacobs listed the presence of small blocks as the second necessary condition for diversity. Small blocks are believed to support stationary activities and provide opportunities for short-term and low-intensity contacts, easing interactions with other people in a relaxed and relatively undemanding way. Specifically, she stated that "lowly, unpurposeful and random as they may appear, sidewalk contacts are the small change from which a city's wealth of public life may grow". She criticized super-blocks and rectangular blocks, which constrain urban mobility with high travel distances and limited opportunities of cross-use.

The easiest way to identify small blocks is to compute the average block area among the set B_i of blocks in district i:

$$\overline{\text{Block area}}_i = \frac{1}{|B_i|} \sum_{j \in B_i} \text{netarea}(N_j) \tag{10}$$

where $\text{netarea}(N_j)$ is the $\text{area}(N_j)$ excluding unpopulated patches such as rivers and natural parks.

Since a district with high intersections density is likely to contribute to random contacts, we also computed:

$$\text{Intersection density}_i = \frac{|\text{intersections}_i|}{\text{netarea}(N_i)} \tag{11}$$

Finally, since block size is distributed as a power law $P(\text{area}) \sim \frac{1}{\text{area}^r}$ with $r \sim 2$, we characterized a district i by its average shape anisotropicity [107] of the blocks B_i within it:

$$\text{District anisotropicity}_i = \frac{1}{|B_i|} \sum_{j \in B_i} \Phi_j \tag{12}$$

where Φ_j is the ratio between the area of the block j and the area of its circumscribed circle $\mathcal{C}_j$:

$$\Phi_j = \frac{\text{area}(B_j)}{\text{area}(\mathcal{C}_j)} \tag{13}$$

The quantity Φ_j is always smaller than one, and the larger its value, the less anisotropic block j, the more opportunities for contacts the block creates.

2.3.2.3 *Aged buildings*

Jacobs stressed the importance of having old buildings in a district. If a district has only new buildings, then it would have only enterprises that can support the high costs of new constructions. If it has old buildings too, instead, it would be able to incubate new small enterprises that cannot afford high rents, and that will benefit the local economy in the long run: "If the incubation is successful enough, the yield of the building can, and often does, rise" [10].

As a first measure, we computed the average building age in district i:

$$\overline{\text{Building age}_i} = \frac{\sum_b |\text{building}_{b,i}| \cdot \text{age}(b)}{\sum_b |\text{building}_{b,i}|} \tag{14}$$

This measures how old, on average, a building in district i is. Since age is expressed in age bands in the Italian census (the first band is `E8` in Table 2, and the last is `E16`), we needed to compute the temporal length of each band. More specifically, for each band b in the range $[\text{start}_b, \text{end}_b]$, we computed:

$$\text{age}(b) = \frac{(\text{last} - \text{start}_b) + (\text{last} - \text{end}_b)}{2} \tag{15}$$

where last is the year of the latest built building in the whole census data. Then, in equation (14), we weighted age band b's temporal length (old_b) by the number of buildings in that band ($|\text{building}_{b,i}|$), and that gave us the average age of the district's buildings.

We also computed the corresponding standard deviation as:

$$\sigma_{\text{building age}_i} = \frac{\sum_b |\text{building}_{b,i}|^2}{\left(\sum_b |\text{building}_{b,i}|\right)^2} \cdot \sigma^2 \tag{16}$$

where σ^2 is the standard deviation of the number of buildings in each band. The higher the standard deviation, the higher the district's heterogeneity in terms of building age.

Finally, we computed the average number of employees per company in district i:

$$\overline{\text{Employees per company}_i} = \frac{1}{|F_i|} \sum_{j \in F_i} \text{employees}(j) \tag{17}$$

where F_i is the set of companies in district i.

2.3.2.4 *Concentration*

Jacobs's fourth and final condition is about having concentration of (both residential and non-residential) buildings and of people. Similarly to Sung *et al.* [72], we computed two sets of concentration measures: one for people, and the other for buildings.

First, population and employment density measures were calculated dividing the number of people (employees) by the district's net area.

$$\overline{\text{Population density}_i} = \frac{|\text{Population}_i|}{\text{area}(N_i)} \tag{18}$$

$$\overline{\text{Employment density}_i} = \frac{|\text{Employed people}_i|}{\text{area}(N_i)} \tag{19}$$

We then computed the interaction term between those two measures as the ratio between population density and employment density:

$$\frac{\text{Population density}_i}{\text{Employment density}_i} \tag{20}$$

The higher it is, the more residents (as opposed to employees) in the district. Different values correspond to districts with different social textures.

To add to those people-based concentration measures one building-based measure, we computed:

$$\text{internal}_i = \frac{|\text{internal apartments}_i|}{|\text{buildings}_i|} \tag{21}$$

which is the average number of apartments per building.

Finally, as we have done previously, to go beyond people and buildings and look at activities, we computed:

$$\text{Density of daily places}_i = \frac{|\text{daily-use places}_i|}{\text{area}(N_i)} \tag{22}$$

$$\text{Density of non-daily places}_i = \frac{|\text{non-daily-use places}_i|}{\text{area}(N_i)} \tag{23}$$

These two quantities are not totally anti-correlated to each other since not all places can be classified as being fully daily *vs.* non-daily.

2.3.2.5 *Vacuums*

Border vacuums are places that act as physical obstacles to pedestrian activity. For instance, parks can be a hub of pedestrian activity, if efficiently managed [10], but they could also be deplorable places in which criminality flourishes (especially at night). In a similar way, the proximity to expressways may discourage pedestrian activity or may effectively connect different parts of the city. This is what Jacobs called "the curse of border vacuums".

Therefore, we needed to identify large areas with single use. From OSM, we took parks[11], and, from the Urban ATLAS, we took railway areas, fast transit zones, rivers, lakes, and highways. That extraction allowed us to build the sets of large parks LP, railways R, highways H and water areas W.

To verify the impact of a type of vacuum area, say, that of large parks on a district, we calculated the average distance between a district's block (i.e., smallest area surrounded by street segments) and its closest large park:

$$\text{Closeness to LP}_i = \left(\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, LP))\right)^{-1} \tag{24}$$

where $\text{dist}(j, \text{closest}(j, LP)$ is the distance between block j and its closest large park. The sum of distances is done over all blocks in district i (B_i is indeed the set of district i's blocks).

The next type of vacuum area is that of railways. Our dataset did not differentiate between rails and stations. However, in terms of pedestrian activity, the two differ: often, rails act as obstacles of activity, while stations act as hubs. To incorporate that distinction, we excluded the stations (i.e., the 600-meter buffer areas around them) from the set R of railways, and computed:

$$\text{Closeness to R}_i = \left(\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, R))\right)^{-1} \tag{25}$$

In a similar way, we computed the average distance between a district's block and its closest highway, and its closest water area:

$$\text{Closeness to highways}_i = \left(\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, H))\right)^{-1} \tag{26}$$

$$\text{Closeness to water}_i = \left(\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, A))\right)^{-1} \tag{27}$$

where H and A are the sets of highways and of water areas.

2.3.3 *Activity density*

As it has been extensively done in previous work, we used mobile Internet activity as a proxy for *urban vitality*.

To roughly estimate the number of Internet connections that fell into each district, we represent the urban space as a set of 2-dimensional, non-overlapping,

---

11 We took areas identified by `leisure=park`, filtering out the parks with `barrier`.

and non-convex polygons. These polygons come from a Voronoi tessellation based on the radio stations' positions.

Then, to estimate the number of Internet call records $S_i(t)$ in each district i at time t we defined:

$$S_i(t) = \sum_v R_v(t) \frac{\text{area}(v \cap i)}{\text{area}(v) - \text{area}(v \cap W')} \quad (28)$$

where v is the Voronoi polygon of the antenna's coverage, and $R_v(t)$ is the number of Internet connections in v at time t. The count of Internet connections is weighted by $\frac{\text{area}(v \cap i)}{\text{area}(v) - \text{area}(v \cap W')}$, which is the proportion of the antenna's area that falls into district i, divided by the total area of the antenna's coverage, where we removed sea areas denoted by W' as we consider negligible all the Call Details Record produced in the sea.

Then, having $S_i(t)$, we computed a district's *activity density* as the average number of Internet connections throughout a typical business day, divided by the district's area. The results for Milan and Rome are shown in Figure 1. The normalization of surface area makes it possible to compare the activities of districts of different sizes.

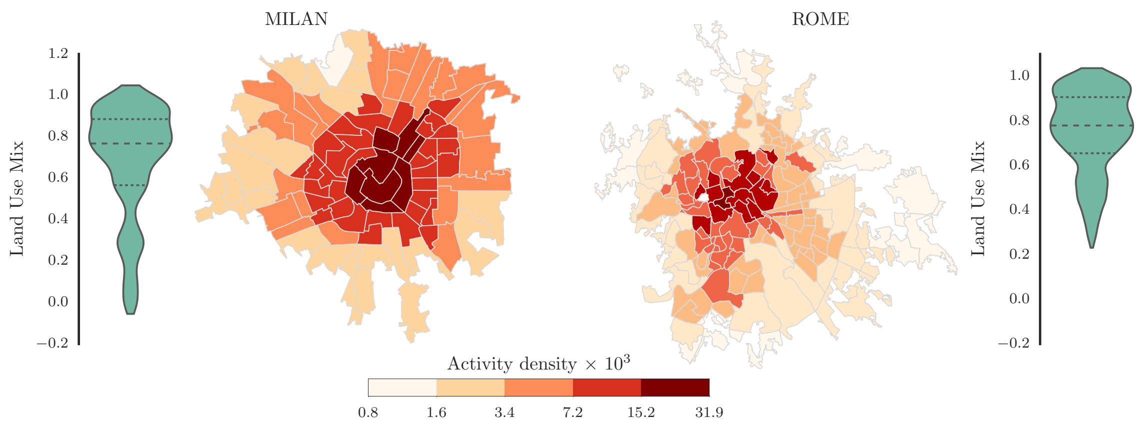


Figure 1: District activity density in Milan (left) and Rome (right), and their corresponding values of mixed land use (LUM). The activity density maps show, perhaps as expected, that vitality is higher in city centre than in the peripheral parts of the city. The violin plots show that the neighbourhoods in Rome exhibit, on average, high land-use mix, while Milan has a wider distribution showing even neighbourhood with only one land use.

2.3.4 *The regression model*

We used an Ordinary Least Squares (OLS) model to evaluate the relationship between each district's structural diversity (Section 2.3.2) and the district's activity density (Section 2.3.3).

Since most of the regression variables were skewed, we transformed them. More specifically, we log-transformed activity density using the natural logarithm, and transformed the structural diversity metrics using the Box-Cox method [108]. To avoid over-fitting, we did split the data into training set (75%) and test set (25%), and repeated our measurements 1000 times using a shuffle split cross-validation.

We created six linear models. First, we created five models, each of which had *one* of the five sets of Jacobs's metrics as independent variables (i.e., we separately analyzed land use, small blocks, aged buildings, concentration, and vacuums). Those models always had activity density as dependent variable. The resulting coefficients are shown in the first five columns of Table 5. We then had a combined model by selecting the independent variables through both recursive feature elimination and stability selection [109]. The resulting selected variables and corresponding coefficients are shown in the last column of Table 5.

2.4 RESULTS

2.4.1 *Land use*

Mixed land use matters only in cities in which functional uses were historically separated like in the case of Milan. By looking at the first two predictors of land use in Table 5, one learns that mixed land use does not contribute neither positively nor negatively. That is because not all cities are equal: land use in Rome is mixed (Figure 1), while Milan is separated in functional areas. Consequently, in Milan, vitality is experienced only in the mixed districts. However, the highest beta coefficient concerning land use is found to be related to the presence of third places (β = 0.3972). Daily errands are important, but third places are more so (with a Spearman's correlation of 0.8337 with activity density). These places are public places in which people can hang out for good company and lively conversations, putting aside the concerns of home and work (their first and second places) [103]. Examples of such places include pubs, coffee shops, and taverns.

	Land use	Small blocks	Aged buildings	Concentration	Vacuums	Combined
$AdjR^2$	0.70	0.63	0.40	0.73	0.19	0.77
Land use mix (1)	0.023					
Closeness small parks (SPs) (2)	0.040					
Residential *vs.* Non-Res. (RNR) (3)	-0.018					
Housing types (4)	0.195***					0.185***
Commercial (5)	0.013					
Night-life (6)	0.011					
Night-life density (7)	-0.110*					
Daily (8)	0.044					
3^{rd} Places (9)	0.397***					
Block area (10)		0.061				
Intersections density (11)		0.738***				0.191***
Anisotropicity (12)		0.046				
$\overline{\text{building age}}$ (14)			0.368***			
$\sigma_{\text{building age}}$ (16)			0.226***			
$\overline{\text{Employees per company}}$ (17)			-0.077			
Population density (18)				-0.148		
Employment density (19)				0.740***		0.434***
$\frac{\text{population density}}{\text{employee density}}$ (20)				0.302*		
$\frac{\lvert\text{internal apartments}\rvert}{\lvert\text{buildings}\rvert}$ (21)				0.022		
Density of daily places (22)				0.204**		
Density non-daily places (23)				0.106		
Closeness large parks (24)					0.138**	
Closeness railways (25)					0.174***	
Closeness highways (26)					-0.345***	-0.101***
Closeness water (27)					0.099	
3^{rd} Places × Closeness to highways						0.068**
Closeness to SPs × Closeness to highways						-0.078**

Table 5: Linear regression models that predict district activity density. Each column is a different model. For each model's column, the table reports the model's predictive power (Adjusted $\mathbf{R^2}$ in the first row) and the β coefficients. Blank cells correspond to variables that are absent from the model. As for statistical significance, we use the following notation: $^*\mathbf{p < 0.05}$, $^{**}\mathbf{p < 0.01}$, $^{***}\mathbf{p < 0.001}$.

.

2.4.2 *Small blocks*

Continental European cities do not have the super-blocks typical of American cities. This is especially true in Italy, so it comes at no surprise that block size does not greatly matter (β = 0.0615 in model "small blocks" of Table 5). By contrast, the density of intersections does matter: vibrant urban areas are those with dense streets, which, in fact, slow down cars and make it easier for pedestrian to cross [110], creating what Jane Jacobs called the "sidewalk ballet". This 'ballet' goes beyond simple pedestrian activity. It is about informal contacts and public trust (e.g., children playing on the street, adults other than their parents paying attention to them, leaving home keys with shopkeepers). To increase the potential for such types of interactions, it is essential to be able to look into each others' eyes, and only small streets could foster that.

2.4.3 *Aged buildings*

In large American cities, Jacobs observed that neighbourhoods with aged buildings tended to have a diversified local economy. By mingling building of varying age and conditions (i.e., mixing expensive building with "a good lot of plain, ordinary, low-value old buildings, rundown old buildings"), a neighbourhood attracts not only standardized and high-earning enterprises but also ordinary and innovative ones. Aged buildings were a rarity in American cities. Back in her days, Jacobs noted: "in Miami Beach, where novelty is the sovereign remedy, hotels ten years old are considered aged and are passed up because others are newer" [10]. Decisions about which buildings get to stay and which get to go depend on culture: they are influenced by the relationship that a culture has with time and place. In a neighbourhood, the patina of time may be *retained*, *imitated*, or *removed*, as Kevin Lynch puts it [111]. In Western Europe, the idea of preservation appeared about 1500 and, since then, retention is by far preferred over removal. As a result, Italian districts are defined by age. Central areas are the ancient ones and their surrounding areas followed over time - each district ended up having its own era. Consequently, in the Italian context, mixing buildings of different eras is not as important as (or, rather, as possible as) it is in the American context.

2.4.4 *Concentration*

The last of Jacobs's conditions is about dense concentration of people and buildings. In our cities, the most informative purpose is that of office work (highest beta coefficient for the column named "Concentration" of Table 5). This is reasonable as the previous three conditions (embedded in the regression) speak to the vitality contributed by *residents*, and this fourth condition adds to it the vitality contributed by *dwellers* who likely work in the district.

2.4.5 *Border vacuums*

Finally, looking at the "vacuums" model in Table 5, we observe little effect of border vacuums. Surprisingly, railways or rivers - which might hamper pedestrian activity at times - seem, instead, to be effectively integrated in the social fabric of active districts. However, as one expects, highways are not: in general, being close to them ends up being detrimental ($\beta = -0.3457$ in the "vacuums" column and $\beta = -0.1018$ in the "combined" column).

Figure 2 shows to which extent (in terms of β coefficients) each of Jacobs's four conditions explains district activity.

2.5 DISCUSSION AND IMPLICATIONS

Taken together, our results suggest that Jacobs's four conditions for maintaining a vital urban life hold for Italian cities as well (see Figure 2). From Figure 2 and the last column of Table 5, we see that as much as 77% of the variability of district activity is explained by simple structural and static features. As Figure 3 details, even individual features are strongly associated with activity.

Also, the extent to which the different features matter does not dramatically change across cities. To partly verify that, we took the largest and smallest cities: Rome and Florence. As Figure 4 suggests, all the main features mattered to a very similar extent.

Based on our findings, to paraphrase the four conditions in the Italian context, we might say that:

> *Active Italian districts have dense concentrations of office workers, third places at walking distance, small streets, and historical buildings.*

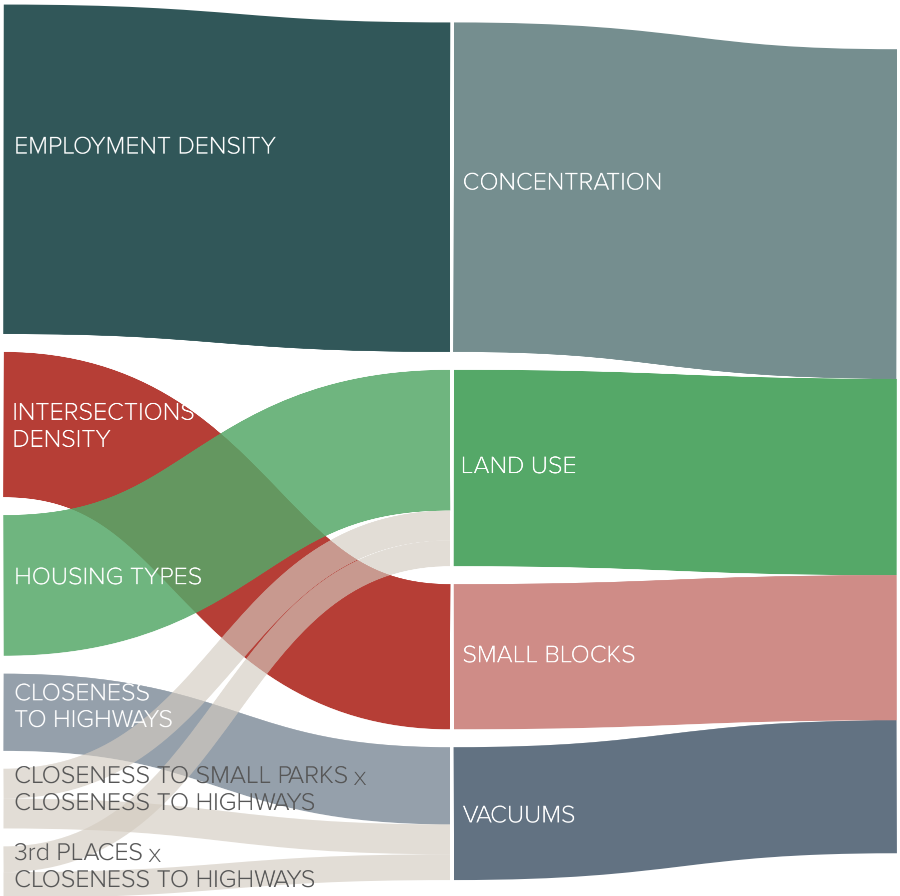


Figure 2: The predictive power of Jacobs's features. The bar size is proportional to the absolute β value in the linear regression model. The first column shows individual features, and the second shows grouped features. The three groups - concentration, land use, and small blocks - plus vacuum areas explain 77% of the activity density variability.

After having operationalized Jane Jacobs's four conditions, we now discuss our work's practical and theoretical implications, and the questions it left open.

Practical Implications. We foresee several practical implications of our work:

City dashboards. Part of our work could inform the selection of which features should go into a city dashboard. We have shown that a variety of structural features of the built environment are closely linked to district activity. Given the importance of those features and the ease with which they could be computed upon web and official data, city dashboards could show and track them and, in so doing, could support well-informed decisions by policy makers, urbanists, and architects.

Quantifying regulatory interventions. In his book 'A City is Not a Tree' [112], Christopher Alexander argued that the variety and diversity essential for

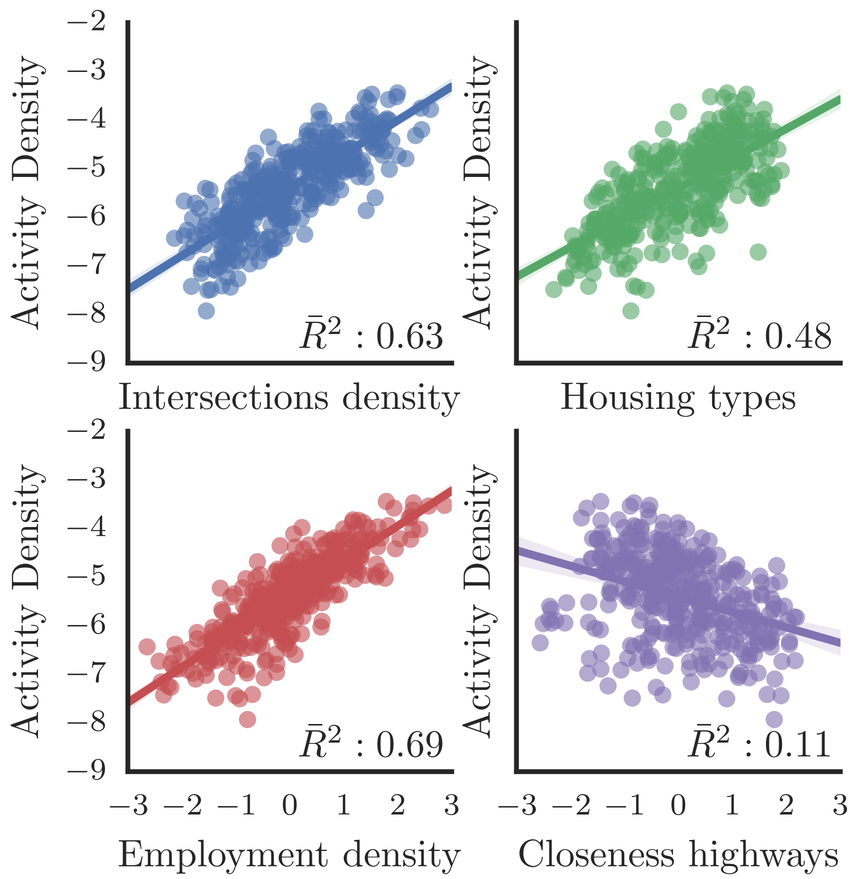


Figure 3: Plots of district activity as a function of intersection density, housing types, employment density, and closeness to highways. A district's activity is strongly associated with the district's static structural features.

urban vitality was being destroyed by the implementation of zoning laws. Our work has focused on urban diversity features, the very features that zoning laws impact. To track the impact of a regulation, one now knows *which features* to contrast before and after the implementation of the regulation.

Place recommendation. How to best place retail stores? Researchers have recently tried to answer that question with semi-automatic mechanisms that identify the amenities that are missing from a neighbourhood. Hidalgo *et al.* [66] analysed the economic diversity of neighbourhoods to recommend new places for them. Similar analyses could profit from the addition of our work's structural features to the already present economic ones.

Theoretical Implications. Some of Jane Jacobs's theories continue to be contro-

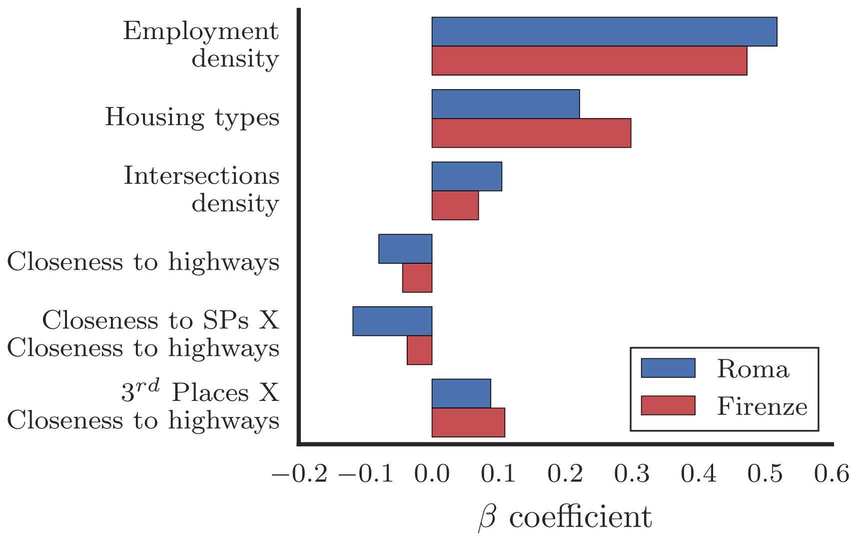


Figure 4: The β coefficients of two city-specific linear regression models for Roma and Firenze.

versial, while others have been criticized for their non-verifiability. This study showed that those theories not only are verifiable but also are still important today and valid in the European context too.

Limitations. Our measure of activity has been assumed to be a good proxy for city's vitality. However, that measure has left out one important factor that might be extracted in the future: temporal dynamics of district activity (e.g., differences between day and night). Also, if one were to have individual calls records instead of (as in our case) aggregate ones, one could well extract additional factors: duration of individual stays, radius of gyration, and identification of particular segments of the population (e.g., pedestrians *vs.* non-pedestrians, residents *vs.* dwellers). Finally, it would be difficult to *fully* replicate this study without call records. Those records were difficult to obtain in the past but have now been increasingly made publicly available by telecoms operators for research purposes [5].

Future work. In this chapter we explored, for the first time, how almost static features of the built environment can predict a dynamic dimension such as human behaviour. However, we did not explore the temporal dimension. Does people behaviour follow the changes of neighbourhoods? What is the latency and the impact of this change? Urban phenomena such as gentrification are profoundly changing neighbourhoods in many cities in the world by offering different point of interests, impacting crime, but also the life of people. Thus, a possible new research direction could be based upon existing preliminary work on early sign of gentrification from urban and big data [113, 114], collect timestamped data of the built environment and people, and predict vitality over time.

However, a number of challenges are also apparent. Collecting timestamped data is very demanding and might not be possible for some variables based on census, which is usually collected every ten years. Moreover, alternative sources of data such as Street-view imagery is not tested as well over time.

Additionally, future work could address the relationship of the built environment with human behaviour in different cities in the world. The Jane Jacobs theory was indeed based on the qualitative observation of a few neighbourhoods in New York. Although this theory is now empirically tested in some cities in Italy and Seoul, there is no evidence the indexes works at the same way in cities that are historically and culturally diverse. Our framework based on Open Data and CDRs is potentially applicable all over the world. This opens wide possibilities to qualitative and quantitative comparative analysis of urban characteristics and influences across countries. We will explore this direction in Chapter 5.

2.6 SUMMARY

For the first time, we verified Jane Jacobs's four conditions necessary for the promotion of urban life in the Italian context. We have done so by operationalizing her concepts in new ways: we used mobile phone records to extract a proxy for urban vitality, and web data to extract structural proxies for urban diversity. As Jacobs envisioned, vitality and diversity are intimately linked.

Until now, the research community has relied on limited data sets that reflect how a group of people experience bits of the city. Here we have found that web data and mobile records offer insights on how most urban dwellers experience entire cities. Web data (especially Open Street Map data) is available around the world, and mobile phone records have been increasingly made publicly available by telecoms operators in search of new business models. As such, our study can be replicated at scale without resorting to lengthy and costly survey-based data collection.

By collecting bread crumbs from cell phones, social media, and participatory platforms, researchers will increasingly rely on data sets orders of magnitude richer than previous urban studies data sets [115] and, consequently, they will be able to test traditional urban theories in fine-grained detail - a totally new way to look at cities.

3

HOW PERCEPTION INFLUENCES VITALITY

Consider a sidewalk. Some litter accumulates.
Soon, more litter accumulates.
Eventually,
people even start leaving bags of trash
from take-out restaurants there
or even break into cars.

— James Q. Wilson and George L. Kelling [76]

Does a neighbourhood's appearance of safety affect how active it is? For decades scholars from a variety of disciplines, but mainly from urban planning, have been exploring the potential connection between a neighbourhood's appearance of safety and its levels of human activity.

As seen in the previous chapter, Jane Jacobs introduced the eyes-on-the-street [1], or natural surveillance hypothesis [116], which suggests that citizens can maintain the safety of their neighbourhoods naturally through continued surveillance. For natural surveillance to take place, however, Jacobs argued that neighbourhoods needed to have certain physical qualities, such as well lit streets and buildings with street facing windows. Jacobs' idea that the physical quality of a neighbourhood can enhance its safety was later expanded by Oscar Newman's *defensible space theory* [74]. Defensible space theory expands on the idea of natural surveillance by suggesting that neighbours will be more likely to protect an area when there are clear physical demarcations separating what is considered public and private property [73, 74]. Examples of architectural markers of defensible space are archways in the entrance of building complexes, or staircases in the entrance of townhouses. These archways and staircases do not only serve an aesthetic purpose, but also, signal the boundary between a city's public space and the private and semi-private spaces that neighbours are expected to watch and defend.

In this chapter, we strengthen the link between Jacobs' and Newman's theories by asking whether safer looking neighbourhoods are more likely to experience more human activity–and hence, experience more natural surveillance. We explore this connection, by combining computer vision methods, that can be used to measure the physical characteristics of neighbourhoods [65, 117–

119], with mobile phone data, which has become a common proxy for human activity [41, 55, 120–122], for two Italian cities (Rome and Milan). The combination of computer vision and mobile phone data helps us test whether safer looking neighbourhoods are more active, and therefore, if neighbourhoods that look physically safer could be experiencing more natural surveillance.

Our data provides support for a connection between appearance and activity. Using spatially filtered multivariate regressions we find that neighbourhoods that are perceived as safer are more active than what is expected from their population density, the density of employees, and their distance to the city centre. Also, we find that the perception of safety appears to modulate the relative population of females and adults, with unsafer looking neighbourhoods experiencing a lower number of female and people over 50 than safer looking neighbourhoods. Conversely, we find that younger populations are disproportionately more active in unsafe looking neighbourhoods. Finally, we use occlusion techniques to identify the areas of an image that trigger a positive or negative evaluation of safety in the Artificial Neural Network, finding that greenery and street facing windows tend to be associated with higher levels of safety, as suggested by Oscar Newman's defensible space theory. These observational results strongly suggest–but don't causally prove–that the appearance of neighbourhoods has an effect on their levels of human activity, and potentially, on a neighbourhood's level of natural surveillance.

The chapter is based upon De Nadai *et al.* [2] and it is organized as follows. Section 3.1 reviews the relevant literature; Section 3.2 presents the sets of data and describes the methods; and in Section 3.3 we provide our results.

3.1 STATE OF THE ART

The connection between urban perception and human activity speaks primarily to two streams of literature. The first one is the stream of literature focused on the environmental factors contributing to crime, which has a long tradition in criminology and urban sociology. While our chapter does not focus on crime per se, the connection between the physical appearance of neighbourhoods and natural surveillance suggested by Jacobs and Newman makes our results relevant to that stream of literature [10, 73, 74]. The second one is the stream of literature using surveys, and more recently, computer vision methods to quantify people's perception of urban environments [65, 123, 124].

3.1.1 *neighbourhood appearance and crime*

Beyond Jacob's and Newman's theories, the most widely known theory suggesting a connection between urban perception and crime is the *broken windows theory* (BWT) of Wilson and Kelling [75, 76]. The BWT is the hypothesis that urban incivilities, such as broken windows and litter, promote criminal activity. The classical mechanism used to justify the theory says that urban incivilities signal lawlessness, and may cause the offenders of small incivilities to scale their criminal behaviour to more predatory forms of crime if they are not reigned in. The policy implications of the BWT, however, vary from community policing—the promotion of ties between police officers and their communities—to zero-tolerance polices, which promote cracking down on all minor offences to deter more serious forms of crime.

Evidence in favour of the broken windows theory has been presented by Kelling and Coles [75], who looked at data and stories from New York to Seattle to argue that community policing is an effective way to deter more serious forms of crime. Kelling and Sousa [77] provide additional evidence by using an extensive dataset on crime, demographics, and economic data from New York City. More recently, Corman and Mocan [125] used New York City data on the policing of misdemeanours (as a proxy for broken windows policing), and on robbery, car theft, and grand larceny, to provide evidence in support of broken windows policing.

The broken windows theory is also supported by a few field experiments, such as those conducted by Keizer *et al.* in The Netherlands [126]. In six experiments, Keizer *et al.* intervened environments by spraying graffiti on walls, or leaving supermarket carts unattended and studied the behaviour of subjects, in both the presence and absence of disorder, to see when people broke norms (such as littering). Their data showed a significant increase in people's norm breaking behaviour when they were in the presence of disorder.

But not all of the evidence collected to test the BWT, and its policy implications, is favourable to it [127]. In a 2006 paper, Bernard Harcourt re-analysed the data presented by Kelling and Sousa [77] and found no evidence of the effectiveness of the broken windows policing [128]. More recently, Harcourt and Ludwig used a dataset of more than fifty thousand marijuana related arrests to provide evidence that community policing is not only ineffective, but that it also unfairly targets minorities [129].

Moreover, the BWT has been criticized by work showing that the social and ethnic context of a neighbourhood may matter more than urban disorder. In Sampson *et al.* [30] and Sampson and Raudenbush [32, 34], community data from Chicago was used to argue that racial and economic context were more predictive of disorderly behaviour than physical disorder. To help bridge their results with the literature, Sampson and Raudenbush [32] proposed an alternative interpretation of the theory, where both neighbourhood disorder and crime, are manifestations of a lack of informal forms of control within disengaged and distrusting communities. The reframed theory, therefore, interprets the link between disorder and crime as a manifestation of the lack of informal forms of social control and organization.

3.1.2 *The social and computational image of the city*

The second literature our chapter speaks to is the literature measuring urban perception and understand its social and economic implications. The original literature on this topic can be traced back to seminal work by the urbanist Kevin Lynch [130], who interviewed people in Boston, Jersey City, and Los Angeles, to understand the large scale image of cities that people made in their heads. This work on a city's imageability was later continued by social psychologists like Stanley Milgram [131], and by urbanists like Jack Nasar [124, 132], who created evaluative maps of cities, also using survey methods.

More recently, however, this literature started leveraging crowdsourcing [65] and computer vision methods [119, 123] to improve the scale, precision, and resolution of the evaluative maps created.

On the data collection side of this literature, Salesses *et al.* [65] created a large crowdsourced visual perception survey to measure people's perception of streetscapes, and to create comparable evaluative maps for New York, Boston, Linz, and Salzburg, which they also used to measure the segregation and inequality of experiences in these cities, and to show that violent crime correlates with the variance of appearance of safety in an area.

These new sources of crowd-sourced data gave rise to studies looking to understand the features of an image that explain how streetscapes are perceived. On a recent study, Quercia *et al.* [118] investigated which visual aspects of London neighbourhoods make them appear beautiful, quiet and/or happy. A related study by Porzi *et al.* [119] identified the visual elements that contributed to an image's perceived level of safety.

But to scale the study of urban perception to multiple cities, and to high spatial resolutions, researchers begun developing computer vision methods to score millions of images [123]. These computer vision approaches build on research predicting human perception from visual data [133–135] and on research analysing visual streetscapes for city understanding [136–141]. The latter of these two lines of research has been fuelled by the widespread availability of geolocated image data, such as Google Street View maps or city snapshots publicly shared on social networks (e.g. Flickr, Instagram) [136, 141, 142]. Using geotagged image data Doersch *et al.* [138] showed that geographically representative visual elements, like architecture styles, can be automatically discovered from Street View Imagery. In a dynamic study, Naik *et al.* [117] used computer vision and images from different time periods to measure urban change, and to study the factors that contribute to neighbourhood improvement. Computer vision methods have also been used to show that a city's visual attributes work as proxy of social and economic characteristics, such as crime rates and proximity to local businesses [137, 140], or census characteristics, such as income and inequality [143].

This new wave of research has benefited from advances in deep learning, which have been used not only to measure appearance, but also for place recognition. Zhou *et al.* [141] introduced Places205 Dataset, a large data collection gathering more than 7 million labeled pictures of scenes. They achieve state-of-the-art classification results by training deep Convolutional Neural Networks (CNNs). More recently, Arandjelović *et al.* [136] introduce NetVLAD, a modified CNN architecture able to address large scale visual place recognition. In this work, we build on top of recent work in place recognition [136, 141] to fine-tune a deep CNN architecture and show experimentally that under scarce training data, sample augmenting helps achieve state-of-the-art results on safety prediction from streescape images.

3.2 OUR APPROACH

This section focuses on investigating the relationship between the appearance of safety and the activity or liveliness of neighbourhoods in Rome and Milan. To achieve this goal we estimate the appearance of safety for different districts of the city by spatially aggregating the safety scores obtained from the images within these areas, and then, observe people's activity using mobile phone data. We densify our maps by collecting additional images and scoring them using computer vision to achieve enough coverage. Finally, we spatially aggregate scores including census tracts for the 2011 Italian census. In the reminder of this section we explain the methodology used to score images and to measure activity.

3.2.1 *Sets of data*

We start by describing the datasets used to estimate urban appearance, and human activity.

Urban appearance data. We use urban appearance data from the Place Pulse dataset[1]. Place Pulse is a large, crowdsourcing project on human perception of cities. The data collection is designed as an online game, where participants are shown two images of streetscapes and are asked to choose one image in response to an evaluative question such as: Which place looks safer? Or: Which place looks more lively? A score is later computed for each image using the TrueSkill [144] algorithm.

Place Pulse begun as Place Pulse 1.0 (PP1), which scored 4,109 images from two US cities (New York City and Boston) and two European cities (Linz and Salzburg), and was launched publicly in 2011. PP1 scored images across three evaluative dimensions: *Safety*, *Upper-Class* and *Unique*.

The current version of Place Pulse (Place Pulse 2.0), launched publicly in July 2013, extended the data collection effort to 56 cities from all continents (except Antartica) and to six evaluative questions: *Safety*, *Wealthy*, *Boring*, *Lively*, *Depressing* and *Beautiful*. For more details, please see Dubey *et al.* [145]

Here we use data from the Place Pulse 2.0 dataset for Milan and Rome (PP2-I). The data includes 3,897 images that received about 25,000 evaluations for their perception of safety[2] - corresponding to an average of 7.6 clicks per image.

Mobile phone activity data. To proxy human activity we use mobile phone billing data, which records the time of communication, and the radio base station that handled it, for various types of communication (e.g. incoming calls, outgoing SMS). The data is provided by TIM and it is aggregated every 60 minutes, and includes both TIM and roaming customers in Milan and Rome from February to June 2015.

Differently from the CDRs used before (see Section 2.3.1), this data only refers to calls and SMS either received or made in Milan and Rome. Moreover, the mobile operator spatially aggregated the data in grid cells, obtained from the underlying coverage area of the radio base stations. The cell size is $\sim 300 \times 300$ m in the city centre of Milan and it increases up to $\sim 2,300 \times 2,300$ m in

1 http://pulse.media.mit.edu
2 We include in this sum only pairwise comparisons involving either two Italian cities or one Italian vs one non-Italian city

the peripheral part of the cities, where few customers are served per unit of area. For each grid cell, we count the number of people who made or received a call on an hourly basis, broken down by gender (number of males/females) and age.

Thus, while we only base our analysis on calls and SMS, the location is more precise than previously used data (Section 2.3.1), and the demographics of people are also available. This allows to break-down the behaviour of people by age and gender. We note that our data cannot distinguish between pedestrians and people using their phones in their homes, so our measures of activity proxy the number of TIM customers in an area, but not necessarily in the street.

3.2.2 *Measures of Urban Appearance*

3.2.2.1 *Safety Perception from Visual Data*

We use deep Convolutional Neural Networks (CNNs) as our model for predicting safety perception from streetscape imagery. We base this choice on the recent success of CNNs in various computer vision problems (especially object classification, object detection and scene recognition [146–148]). We show that fine-tuning a CNN on predicting the level of safety along with data augmentation leads to improved performance when compared to recent work on the same task [123, 149].

Since we have a limited amount of training data (*i.e.* a sample of a few thousands), we retrain CNNs trained on related domains (e.g. images captured in urban environments that are likely to show similar visual content). In particular, we fine-tune the well known AlexNet CNN [146], trained on Places205 Dataset [141], also known as *Places205-AlexNet*. This model was trained on 205 categories of scenes (many of which capture different areas from cities such as office buildings, churches, residential neighbourhoods, shops, etc.) summing up around 2.5 million images. By retraining a CNN previously trained in a similar domain we transfer some of the knowledge contained in this network. In our case, we find that the resulting network can accurately predict the appearance of safety, as we show in the validation section (see section 4.1.2).

To increase the generalization ability of the trained CNN, we adopt data augmentation by cropping all the images used during training and testing. We find this particularly suitable for our scenario where no constraints regarding image alignment are imposed. Specifically, for every image in the training set, we generate n crops by randomly assigning values to the coordinates of the top left and bottom right cutting points, respectively. We control the size of the crops by bounding the coordinates of the cutting points to ranges proportional with the image size. Formally, given an image I of size $W \times H \times 3$, we generate points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ such that the quantities: $x_1/W, y_1/H, 1 - x_2/W$ and $1 - y_2/H$ are bounded by k_1 and k_2, respectively. We empirically set k_1 to 0.05, k_2 to 0.2 and n to 30 in all our experiments. Additional constraints to control diversity of the crops could be implemented, such as monitoring the *intersection over the reunion (IoU)* between pairs of crops and/or original image. During testing, we average all predictions belonging to the crops of the same test sample. The safety scores obtained at image level are then aggregated into city's districts.

Part of the standard pre-processing steps, all the images are subject to scaling to 227×227 pixels and mean image subtraction. The task is modeled as a regression problem, where the goal is to minimize the l_2 loss between the sample labels and the model predictions. As labels, we refer to the Trueskill scores computed in [123], also used in [149]. We fine-tune with *Caffe* [150] for $10,000$ iterations using a base learning rate of $1e^{-4}$. We note no further decrease in the training loss beyond this limit.

3.2.2.2 *Validation of computer vision models*

We validate the fine-tuned CNN on the two US cities from PP1 (New York and Boston), following training and evaluation protocols from [123] and [149], respectively. In the first case, we perform a 5-fold cross-validation on the 2920 US images and report the average R^2 measure, after scaling all the scores to the interval $[0, 10]$. We obtain an average R^2 of 62.2% a 9.5% relative improvement over the best result reported in [123] and 4.8% over the case of not using data augmentation during testing. In the second case, the source (train) and target (test) domains are populated by alternating US cities (e.g. NY - Boston, NY - NY, Boston - Boston, Boston - NY). As performance measure, the authors report Pearson correlation between the predicted regression values and the label scores. Table 6 shows the comparison results, where we consistently outperform [149] in all four combinations. We also observe how the difficulty of the task (for different pairs) modulates the predicting performance of both sets of models in the same way (e.g. training and testing on Boston seems to be the easiest for both our models and the ones from [149]).

3.2.2.3 *Densification of Appearance data using Computer Vision*

Since the distribution of the annotated images from PP2-I is sparse, on average 6.7 images/km^2, we retrieve additional images for the cities of Rome and Milan by densely sampling geo-referenced images from Google Street View. Data

Model type	Best from [149]	Our result
NY - NY	0.687	0.718
Boston - Boston	0.718	0.744
NY - Boston	0.701	0.734
Boston - NY	0.636	0.693

Table 6: Performance on the estimated level of safety (comparison with [149]). For all cases, we report Pearson correlations with $\mathbf{p} < \mathbf{0.001}$.

densification for the analysis of urban landscape has been used in previous works for generating high resolution maps of city perception [123, 149]. First, we generate a grid of points inside the area we want to cover. The granularity we choose is 100 points/km^2. To better represent the safety perception of the location, we retrieve four images from each location, which have 90 degrees horizontal field of view and different headings (north, east, south and west). By doing this we cover 360 degrees from each location, thus getting less biased safety perception of an area by averaging predicted safety scores of these four images. We developed the script which iterates through all the points and used Google Street View API to obtain four images for each location. The script discards locations where no images are available. Using this method we obtained 83,203 images for Rome and 74,815 images for Milan.

3.2.2.4 *Validation of Densification*

Table 7 reports the performances on predicting safety perception for the city of Rome and Milan. We evaluate the correlation between the aggregated original scores and (i) the predictions over the images from PP2-I and (ii) the predictions over the images from the densified dataset. In general, a slight decrease in performance is observed in the second case. We can attribute this loss in performance to the fact that images from PP2-I were retrieved in 2010, thus a variation in the urban appearance may have occurred within this time lapse. For example, the Expo Milano 2015 - a Universal Exposition - was held from May to October 2015, and Milan underwent a profound (visual) change in its north-west area to prepare for this event.

Predicting safety on Milan and Rome: We are interested in finding a good candidate model for labeling the densely sampled Street View images from Milan and Rome. We experimented with several model choices (including training on PP2-I) and discovered that, surprisingly, using PP1 for training yields the best correlation value for the two Italian cities. We attribute this result to

City	PP2-I	Densified
Milan	0.621	0.488
Rome	0.635	0.548

Table 7: Performance on the estimated safety perception level for each city. All values are statistically significant, with **p** < **0.001**.

the much inferior average number of votes per image (around 7.6 for PP2-I, compared to around 90 for PP1). For the rest of the experiments, we use only the model trained on PP1 for safety prediction. In Figure 5 we visually report the spatial distribution of the safety prediction for Milan and Rome.

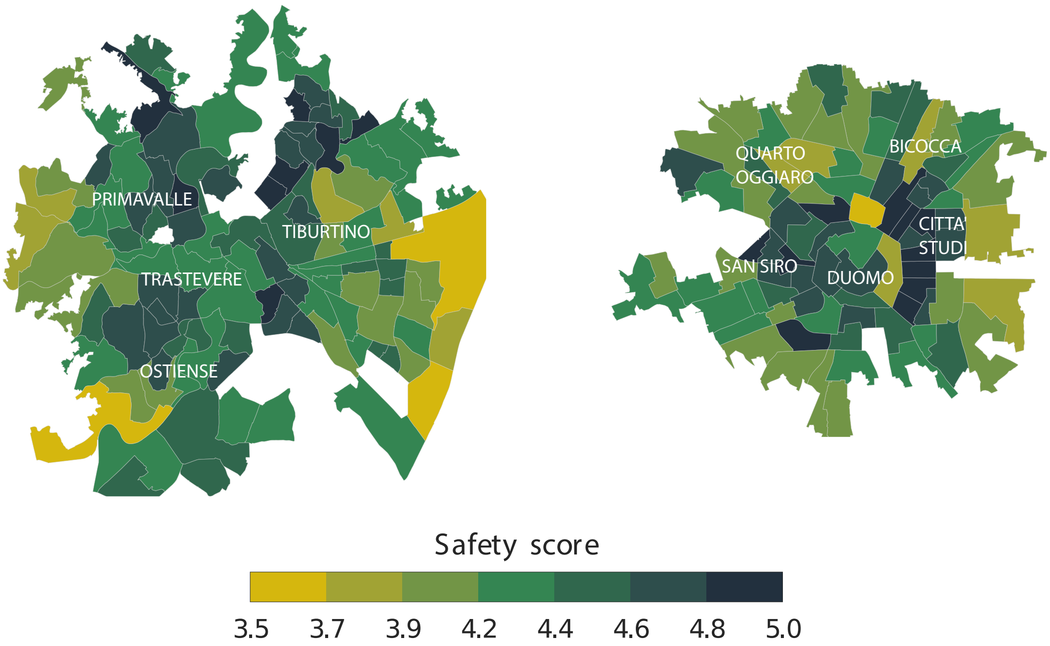


Figure 5: Spatial distribution of perceived safety in each district of Rome and Milan.

3.2.3 *Metrics for Urban Liveliness or Activity*

Next, we define the metrics we use as proxies for an area's level of activity, or liveliness. Here we study only in urban areas where more than 50% of the surface is not composed by farmlands or forests. We measure activity for four populations, all people, females, people younger than 30, and people older than 50.

Formally, we measure the density of all people in district i as:

$$R_p(i, 24h) = \frac{|people_{i,24h}|}{area_i} \tag{29}$$

Next, we measure the fraction of females in a district i as:

$$R_f(i, 24h) = \frac{|females_{i,24h}|}{|people_{i,24h}|} \tag{30}$$

Additionally, we measure the population of people below 30 and above 50 as:

$$R_{<30}(i, 24h) = \frac{|people(<30)_{i,24h}|}{|people_{i,24h}|} \tag{31}$$

$$R_{>50}(i, 24h) = \frac{|people(>50)_{i,24h}|}{|people_{i,24h}|} \tag{32}$$

.

3.2.4 *Spatial Regression*

We test the connection between urban appearance of safety and activity using spatially corrected Ordinary Least Squares (OLS) regressions. Since we are dealing with spatial variables, OLS residuals are assumed not to be spatially auto-correlated; otherwise the regression model is said to be misspecified. Thus, we use the Griffith filtering [151] which extracts a set of orthogonal and uncorrelated eigenvectors from the expression:

$$(I - \frac{11^T}{n})W(I - \frac{11^T}{n}) \tag{33}$$

derived from the spatial auto-correlation Moran's I numerator, where I is a $(n \times n)$ identity matrix, 1 is a $n \times 1$ vector containing only 1's and W is a $(n \times n)$ spatial weight matrix based on topological adjacency, so-called Queen criterion: if two areas share a boundary or a vertex, the entity of the spatial weight matrix is coded as 1, and otherwise, 0. The eigenvectors obtained can be employed in a multivariate regression to account for spatial auto-correlation. However, it is clear that employing all n eigenvectors in a regression framework is not desirable for reasons of model parsimony. Thus, a subset of eigenvectors are selected in a step-wise fashion so as to minimize the sequential residual spatial correlation (Moran's I) values [151]. The final subset of candidate eigenvectors represents the *spatial filter* for the variable analysed. Before applying the regression model, the data were Z-score scaled.

3.3 RESULTS

After describing our methods to measure urban appearance and neighbourhood activity we test whether the appearance of safety and the activity of neighbourhoods is correlated. To test for this correlation we merge our data with information from the Italian census so we can control for other sources that intuitively correlate with neighbourhood activity: population density (residential), density of employment (which should also proxy pedestrian density during the day), distance to city centre, and a deprivation index (to control for poverty in the neighbourhood).

We begin by looking simply at the correlation between the number of people per unit of area observed in our mobile phone dataset and the appearance of safety in a neighbourhood while controlling for population density, employee density, deprivation, and distance to centre.

Table 8 shows the result of a spatially corrected multivariate regression with the number of people per unit of area measured using the density of all people as the dependent variable. Not surprisingly, the strongest correlate of the number of people present per unit of area is employee density, and the number of people per unit area decreases with distance to the center. Yet, despite the strong effect of the other control variables, the appearance of safety is significantly and positively correlated with the number of people present in a neighbourhood per unit area.

Presence of people (29)[l]	
Population density[l]	0.155**
Employees density[l]	0.328**
Deprivation	-0.022
Distance centre	-0.257**
Safety appearence	0.105**
Spatial Eigenvectors	11
Adj-R^2	0.91
Moran's I (p-value)	0.07 (0.08)

[l] log transformed variable.

Table 8: OLS regression model between presence of people and safety perception. The β coefficients are reported in the table. *$\mathbf{p < 0.01}$, **$\mathbf{p < 0.001}$.

Next, we look at the fraction of females present in an area. Looking at the population of females separately is motivated by empirical research showing that women are twice as likely as men to report feeling unsafe [152], even though they have a much smaller risk of being victimized [153, 154]. This suggest that the presence of women in a neighbourhood should be more strongly affected by its appearance of safety than the presence of men. In fact, Felson and Clarke [155] suggest that a high ratio of women in the street is a positive sign towards urban safety, as they act as “crime detractors," in agreement with Jacobs’ natural surveillance hypothesis. These theories would suggest that the ratio of females is lower in places perceived as unsafe.

Table 9 looks at the ratio of females in the population observed in our cell phone data as the dependent variable, finding that the appearance of safety is highly significant and positive. In fact, the coefficient is roughly twice that observed for the general population.

Presence of women (30)	
% of women (residents)^s	0.001
Deprivation	-0.005
Distance centre	-0.003
Safety perception	0.020**
Spatial Eigenvectors	12
Adj-R^2	0.65
Moran's I (p-value)	0.06 (0.11)

^{s} cube-root transformed variable.

Table 9: OLS regression model between presence of women and safety perception. The β coefficients are reported in the table. *$p < 0.01$, **$p < 0.001$.

We also look at the proportion of people younger than 30 and older than 50 in an area. According to Felson and Clarke [155], a younger population is a predictor of criminal incidents in an area, as they show a higher aggression potential compared to older populations. Nevertheless, younger people, especially men, show less fear of crime [156, 157].

Table 10 looks at the ratio of people younger than 30 in the population observed in our cell phone data as the dependent variable. The variable which contributes the most to the correlation is the distance from the city centre, followed by appearance of safety which is highly significant and negative. Contrarily, when we look at the ratio of people older than 50 (Table 11), we find that the appearance of safety is highly significant and positive. This is in agreement with the theory, showing that older people are more likely to be present in places that appear safe.

Presence of people younger than 30 (31)[l]	
% of younger residents[l]	-0.001
Deprivation	0.032**
Distance centre	-0.150**
Safety perception	-0.048**
Spatial Eigenvectors	16
Adj-R^2	0.66
Moran's I (p-value)	0.07 (0.09)

[l] log transformed variable.

Table 10: OLS regression model between presence of younger people and safety perception. The β coefficients are reported in the table. *$\mathbf{p < 0.01}$, **$\mathbf{p < 0.001}$.

Presence of elderly people (32)	
% of elderly residents[s]	0.006**
Deprivation	-0.006**
Distance centre	0.006**
Safety perception	0.017**
Spatial Eigenvectors	14
Adj-R^2	0.64
Moran's I (p-value)	0.07 (0.09)

[s] cube-root transformed variable.

Table 11: OLS regression model between presence of elderly people and safety perception. The β coefficients are reported in the table. *$\mathbf{p < 0.01}$, **$\mathbf{p < 0.001}$.

3.3.1 *Visual Attributes Determining Safety*

Finally, we explore the visual attributes of the images that contribute positively, or negatively, to their appearance of safety. To identify these attributes we set up an occlusion sensitivity experiment. In this experiment, inspired by [158], we randomly generate occluding patches in images and replace them with the average pixel value. For every such altered image, we monitor the effect at the output of the predictor (did the image score higher or lower in its appearance of safety). This allows us to identify patches in an image that contribute positively or negatively to their appearance of safety.

Figure 6 shows some examples, with the original image on the (top row), followed by the areas that contribute to a low appearance of safety (middle row)

Figure 6: (Top) Sample images associated to a (left) low and (right) high level of safety and corresponding activation masks: highlighted areas correspond to the ones that mostly contribute to the perception of (center) unsafety and (bottom) safety.

and a high appearance of safety (bottom row). The images are sorted from an overall low appearance of safety to a high one. The examples, while illustrative instead of comprehensive, show that street facing windows and greenery tend to to contribute positively to an streetscapes appearance of safety. The positive effect of street facing windows is in agreement with the natural surveillance hypothesis of Jane Jacobs.

3.4 SUMMARY

In this chapter we explored the question: "*Are safer looking neighbourhoods more lively?*" in the context of two Italian cities: Milan and Rome. Our findings suggest that perceived safety modulates the active population in an area, with effects that depend on age and gender. The overall effect of the safety appearance in activity appears to be positive, even after controlling for population, employment density, and distance to the city centre. Yet, the effect does not appear to be universal, and depends on the demographic with the population, with females and people older than 50 appearing to have a stronger preference for the appearance of safety.

Our results, however, do not provide a causal explanation of the observed effects. For instance, our data cannot distinguish between the hypotheses that people over than 50 prefer safer looking places, or that they modify their homes and shops to make the places they live and work at look safer. Nevertheless, they provide preliminary evidence suggesting a connection between the appearance of safety and levels of human activity that is strong enough to manifest itself at the city scale. Future directions could explore the temporal aspects of this relationship and find some instrumental variables to define the causality link between appearance and human behaviour. Until this time we can only highlight to policy makers this strong connection.

These methods, which could be readily applied to other cities if the data were available, can help improve modern efforts to use computational methods for urban recommendations. Some recent literature has focused on developing algorithms to recommend places for new business by using data on the presence of amenities in neighbourhoods [66] and other urban features [1].

Part III

THE SPATIAL OUTCOMES

In the book "A Brief History of Time", Stephen Hawking states that it might be impossible have a complete solution by investigating parts of the problem in isolation, especially if everything is related to everything else [159]. In urban systems, this conjecture might be true as it ever was. Yet, existing literature often favoured very simple models that neglected this complexity.

In this part our focus is on models that account for a large number of factors to contrast and compare competing hypothesis. First, we analyse the complex nature of the housing market and show, through a predictive model, the effect of neighbourhood features on the housing value. Thus, we consider both the objective and subjective characteristics that we examined in the previous chapters. Finally, we show how crime can be explained by two tightly coupled and alternative hypothesis coming from urban planning and criminology. Using the model in different cities, we also highlight the limitations on transferring what we learn from one city to another.

4

THE ECONOMIC VALUE OF NEIGHBOURHOODS

If everything in the universe depends on
everything else in a fundamental way,
it might be impossible to get close to a full solution
by investigating parts of the problem in isolation.

— Stephen Hawking [159]

Real estate appraisal is a challenging multi-dimensional problem that involves estimating many facets of a property, its neighbourhood, and its city. In the real estate market, professional appraisers use Multiple Listing Services (MLS) to estimate prices from similar recently sold properties within the same market area [79]. The recent years have seen the advent of new sources of data and new methods, mainly coming from machine learning [160] and the computer vision community [78, 80]. For example, the availability of historical listings, Open Data, and online markets have allowed companies like Zillow[1] and Trulia[2] to emerge, and estimate property values from millions of historical listings. Though, these proprietary systems are heavily dependent on the availability of timeless sale transaction data. In addition, the role of neighbourhood characteristics on housing value remains an open question.

Instead, the neighbourhood's dependency on home prices is very well documented. Property infrastructures [81], traffic [82], neighbourhood popularity and reviews [83] are found to influence the real estate market. Pleasant and *walkable* neighbourhoods have a direct translation on higher housing prices [84, 85]. Security perception and neighbourhood environmental physical characteristics impact on economic activity [161], vitality [1] and liveliness [2], which in turn increase housing prices [162]. These studies often rely on a limited number of factors, thus neglecting other facets and the complex intricacies inside neighbourhoods. This is not to mention the common linearity assumptions between variables under investigation.

In this chapter, we study how neighbourhood's features influence housing values, relating the characteristics of the property together with the environmental, physical, and perceptual characteristics in the surroundings. To do so,

1 https://www.zillow.com/
2 https://www.trulia.com/

we analyze more than 70,000 online advertisements of the largest Italian real estate website and provide a multi-modal analysis of the price drivers in play. Our model does not require timeless historical listings, and it allows predictions to be responsive to urban changes, thanks to the up-to-date geographical data we use.

Our experiments show that the neighbourhood characteristics seen in Chapter 2 and Chapter 3 have a significant economic impact on housing price. Moreover, the use of this information helps the automatic appraisal of houses, reducing the prediction error of the model by 60%.

The chapter is based upon De Nadai *et al.* [3] and it is organized as follows. In the next section we discuss previous work about housing value description and prediction. In Section 4.2.1 we describe the data we used. Section 4.2 formalizes how we approach the problem and how we conduct the experiments. Section 4.3 shows our results. Finally, in Section 4.4 we discuss the results and present some implications of our work for citizens, local governments, and real estate companies and investors.

4.1 STATE OF THE ART

The task of automatically estimate the market value of houses can be seen as a regression problem, where the price (or the price per square meter) is the dependent variable, while the independent one is the available information that could help to determine correctly the price. Hence, the task can be based on a weighted regression of house features [79], historical [163] and neighbourhood prices [164], but also pictures [78, 165]. For example, You *et al.* [78] created a Recurrent Neural Network (RNN) using the images of sold houses in the neighbourhood. Liu *et al.* [165] combined textual features and external pictures of the sold house to rank and predict the price. Fu *et al.* [166] ranked houses through point of interests, their popularity and reviews. This recent line of research is also of paramount interest for the real estate industry. Let us mention as examples Zillow and Redfin[3]: these companies collect past sales prices, mortgage records attached to those sales, and prior tax assessments. Then, they relate these variables to the physical features of the property. This allows them to have good estimates for both private users and industry. However, their approach neglect completely the characteristics of the neighbourhood. Besides, Zillow and Redfin do not disclose nor release the used data and methods, thus making difficult a clear understanding of their performances.

3 http://www.redfin.com

Researchers also have analyzed some socioeconomic drivers and environmental characteristics that influence the price of houses. Cortright *et al.* [84] found a positive correlation between *walkability* and housing prices in almost all the analyzed US cities. People, indeed, prefer to live in places full of opportunities, and reachable without depending on cars. This could also be related to the presence of a public transportation system and low traffic [82]. Other researchers found that intangible qualities of neighbourhoods like culture [162], perception [167] and design [80] can be related to houses' price. For example, Hristova *et al.* [162] found that Flickr tags related to culture are positively associated to urban development and housing prices in a neighbourhood, capturing some aspects of the role played by gentrification. Buonanno *et al.* [167] combined data from a victimization survey and data from the housing market to estimate the effect of crime perception on the housing prices in Barcelona. Their results show that in districts perceived as less safe than the average, houses are highly discounted. Poursaeed *et al.* [80] found that luxury level and design qualities detected from houses' images are found to impact prices of sold houses, after controlling for the offered price. Boys *et al.* [168] analyzed six British cities to find that land use, urban form, design, and diversity matter. Moreover, they found that some features such as greenery are not always a positive thing. Taken all together, however, these works focus on a very limited number of factors per time, neglecting the role played by the others. Moreover, they often assume a linear relationship between the analyzed variables, thus failing in providing a complete and conclusive picture on how the neighbourhood influences the housing prices.

For this reason, our work explores many different facets (e.g. *security perception*, the proximity of greenery) of neighbourhoods at the same time analyzing geographical data and online advertisements in a non-linear fashion. We also release a scalable Open framework that can be employed by researchers, individuals, and companies to estimate prices without the need of timeless historical data.

4.2 OUR APPROACH

In this chapter, we explore the influence played by neighbourhood's characteristics on the houses' value. To this end, we leverage multimodal data to obtain (i) security perception scores from Google Street View images, (ii) socioeconomic characteristics from census data, and (iii) the built environment characteristics from geographical Open Data. First of all, each advertised house is geo-referenced to its containing census block. Then, we compute the features of its neighbourhood (e.g., *walkability*, security perception), which are concatenated to the textual characteristics of the property (see Figure 7). We use these data as input to a predictive model that now-casts the value of houses in 8 Italian cities.

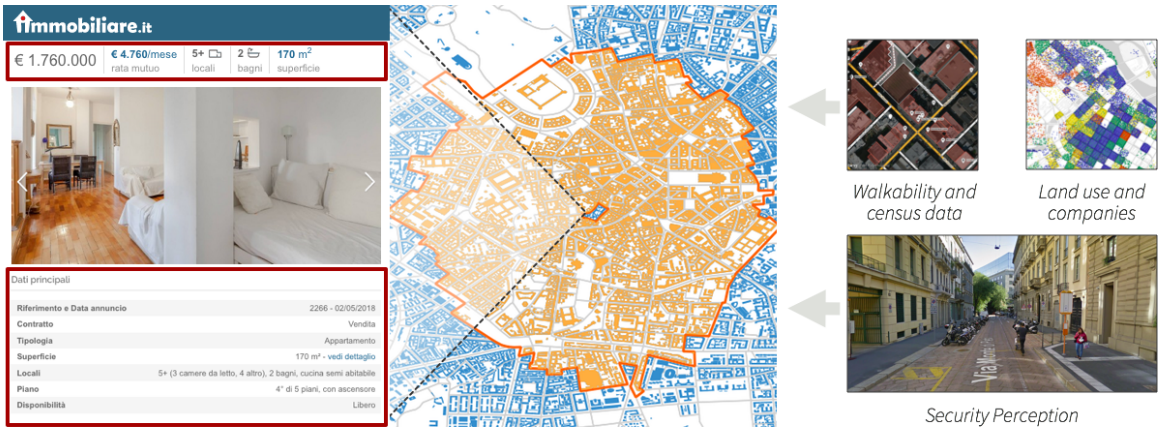


Figure 7: left) An example of housing sales advertisement on Immobiliare.it. We highlighted the information we extract from it; middle) The housing unit is georeferenced and the neighbourhood is built with the census blocks within a 1 km buffer (in orange); right) The model uses the security perception, *walkability*, the data about companies and census of the neighbourhood.

4.2.1 *Sets of data*

We collected several sources of data for the 8 largest Italian cities: Turin, Milan, Genoa, Bologna, Florence, Rome, Naples, and Palermo. Here we discuss the collected sources of data.

Home listings. We collected online advertisements of the largest Italian real estate website Immobiliare.it[4]. In this website, real estate agencies and private people can upload ads specifying the type and the location of the property they want to sell. Moreover, sellers describe the apartment through a brief description of the dwelling, adding some pictures, and a list of characteristics such as the property type and the square meters, the heating system, the floor number and the maintenance status.

Our data consists of a snapshot of online advertisements temporarily collected on May, 10th 2018 from the website.

We focus on apartments, attics, detached and semi-detached houses, loft and open spaces that have both geographical coordinates and asked price. We

4 http://www.immobiliare.it

also excluded foreclosure auctions, buildings under construction, and advertisements that are older than one year. The filtered dataset has 73,383 houses advertised from May, 10th 2017 to May, 10th 2018.

Geographical information: OpenStreetMap and Urban Atlas. We use data from OpenStreetMap[5], downloading a full snapshot of Italy in March 2018. From it, we focus on road networks and amenities. Differently from previous chapters, we here use amenities from OpenStreetMap without relying on Foursquare data. This allows to use Open Data available to the public, with a small trade-off in the precision for the Italian territory.

Similarly to Section 3.2, we also make use of the Urban Atlas 2012, a pan-European project that specifies through 20 classes (e.g., continuous urban fabric, discontinuous low-density urban fabric, green areas, airports) the use of the land.

Google Street View images. We use the urban appearance data temporarily downloaded from Google Street View API[6]. The images were obtained with the methodology aforementioned in Section 3.2.2.1. First, for each city, we generated a grid of points distant from each other 100 meters. Then, for each location, we downloaded four 90-degrees images (north, east, south, and west) to have a full panoramic view of the place. The final dataset includes 154,147 images.

Census data and other sources. Similarly to Chapter 2, the data about population, buildings, and industries are downloaded from the most recent Italian census. The data are publicly provided by the Italian National Institute for Statistics (ISTAT)[7]. We extracted the sizes of businesses from the recent Big Data Challenge 2015[8] that released information about companies in 5 Italian cities.

Property taxes data. We collected property taxes on houses from the disclosed information in advertisements of Immobiliare.it.

5 http://www.openstreetmap.org
6 https://developers.google.com/maps/documentation/streetview
7 https://www.istat.it/it/archivio/104317
8 http://www.telecomitalia.com/tit/it/bigdatachallenge.html

4.2.2 *The neighbourhood*

neighbourhoods are the fundamental geographical unit where individuals' activities and social interactions happen the most. Thus, they have always been investigated by social scientists, urban scientists, and criminologists to study human behaviour [169]. However, defining the boundaries of a neighbourhood is a critical challenge. Usually, they are defined as non-overlapping units through the administrative boundaries defined by the census, but it is unlikely that people move and live obeying to these artificial boundaries. Hence, in our work we define overlapping neighbourhoods resorting to the *egohoods* [170] of census blocks.

Starting from a census block, which we will name *ego-place*, we consider as *egohood* all the census areas within a circular buffer of 1 km. The features of the *egohood* are then computed from this set of blocks (see Figure 7). Thus, we build a spatial binary contiguity matrix W with $W_{i,i} = 0$, where:

$$W_{i,j} = \begin{cases} 1, & \text{if } distance(i,j) < 1\,\text{km} \\ & \text{and } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

Then, the matrix is row-normalized to have $\sum_j^n W_{i,j} = 1, \quad i = 1, 2, .., n.$

The *egohood* features E are then computed as the dot product between the spatial matrix and the characteristics of the blocks F:

$$\underbrace{\begin{bmatrix} x_{0,0} & x_{0,1} & \dots & x_{0,c} \\ x_{1,0} & x_{1,1} & \dots & x_{1,c} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,0} & x_{n,1} & \dots & x_{n,c} \end{bmatrix}}_{\textbf{place-features}} \underbrace{\begin{bmatrix} 0 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 \end{bmatrix}}_{\textbf{W}} = F \cdot W = E$$

We will explain in then next sections some precautions to ensure the validity of the hold-out dataset and the cross-validation.

We assume that a house in block A has a value that depends not only by the characteristics of the property and the block but also by the *egohood* that surrounds it. As we previously mentioned, the study of the neighbourhood effects on human behaviour has a long tradition in social science, economics, urban science and criminology [169]. However, the recent emergence of new sources of data has allowed to empirically connect neighbourhood's characteristics with people behaviour at a large scale. In the next sections, we describe the neighbourhood characteristics (e.g., *walkability*, *urban fabric*, *cultural capital* and *presence of industries*, *perceived security*, and *living/socioeconomic conditions*) captured by our approach.

4.2.2.1 *Walkability*

Empirical studies in the US found that neighbourhood's *walkability* has a substantial impact on housing prices [84]. In this report, researchers explored this connection by using the Walk Score index, a proprietary algorithm that scores the walking distance of some typical consumers' destinations. This score ranges from 0 to 100, where the first value represents a car-dependent place, while the latter describes a place where all the typical amenities are reachable by foot. Walk Score accounts for nine different categories of amenities (e.g., restaurants and bars, libraries, grocery stores), and it measures the walking distance through an exponential decay function that reaches 0 at 1 mile distance. This score is very convenient for researchers; however, it is proprietary and available only in the US at the present moment. Thus, we here create a similar score based on OpenStreetMap data.

OpenStreetMap contains all the ingredients - the road network, points of interest, and the geographical areas (e.g., parks) - that can be used to create an Open version of Walk Score. Thus, for each *egohood* we compute the *walkability* score for the following categories: coffee places, entertainment, shopping, restaurants, schools, grocery, library, and parks. Similarly to Walk Score, the *walkability* score range is in $(0, 1)$. It is computed from a distance decay function that equals 1 when amenities are less than 500 meters far, then decays very fast until values are close to a maximum distance M, where it starts decaying slowly. Amenities at distances higher than M do not contribute to the score. Thus, the score of block i to amenity j is computed as:

$$s_{i,j} = e^{-5(d(i,j)/M)^5} \tag{34}$$

where $d(i, j)$ is the number of meters of the shortest path between i and j, along the road network. M is the maximum walking distance considered, which is set to 1.

Restaurants, bars, shops, and parks are among the most common destinations reachable by walking. In these categories, a variety of options are found to be significant. Thus, the *walkability* score of restaurants is averaged over the ten nearest destinations; this limit is set to five for the shopping destinations, and to two for the park categories. Similar to Walk Score, restaurants and bars are merged into one category due to their overlapping nature.

Public transports represent an efficient and sustainable way to move around the city. For this reason, we also measure the accessibility to the nearest metro and railway stations, the distance to the nearest airport, and the number of bus stops in the neighbourhood.

4.2.2.2 *Urban fabric*

The *vitality* of neighbourhoods is believed to depend on how the city is organized and built. As aforementioned, urban activist Jane Jacobs stated that neighbourhoods should have four essential conditions to be vital [10]: I) two or more primary uses to have people flocking for different reasons all day long; II) small blocks to boost face-to-face interactions; III) a mix of old and new buildings to mix big and small enterprises but also people from different income brackets; and IV) a sufficient concentration of people and enterprises to have a reason to live in the neighbourhood. In this Section we define some indexes similar to Chapter 2.

From the urban ATLAS dataset we extract the areas dedicated to urban, commercial, and green uses of land. Then, we also use this information to compute the land use mix [1] as:

$$\mathrm{LUM}_i = -\sum_{j=1}^{n} \frac{P_{i,j}\log(P_{i,j})}{\log(n)} \tag{35}$$

where $P_{i,j}$ is the probability to have land use j in the *ego-place* i. $n = 3$ is the number of land uses considered: (i) urban, (ii) commercial, and (iii) green areas. We also account for the number of residential, commercial and total buildings from census data. Differently from urban ATLAS, these numbers are tight to the census, and thus updated only every ten years. On the contrary, satellite data can be updated very often.

To account for the presence of small blocks, we compute for the average square meters of census blocks. The census institute (ISTAT) delineates these blocks as the smallest enclosed area surrounded by roads or water.

The third Jane Jacobs' condition is about the mixture of companies, old buildings, and people's income. Thus, we extract from census data the "average" size of companies, and the number of buildings for each year-bracket. We also compute the average and standard deviation of the construction years of the buildings.

The fourth and last condition is about residential and commercial density. Thus, we extract the number of companies, shops, and population for each census block.

4.2.2.3 *Cultural capital and heavy industries*

Cultural capital has been found to influence housing prices and people's behaviour in cities [162]. Thus, we extract the data of companies with ATECO codes falling in the following macro-categories: 58) Publishing; 59) Film, TV, video, radio, photography; 62-63) IT software and computer services; 71) Architecture; 73) Marketing; 74) Design: product, graphics and fashion; 90) Music, performing, and visual arts; 91) Libraries and museums. Then we add the count of cultural-related companies to the list of features.

Industries can also influence the health and the likeability of neighbourhoods. We account for the presence of heavy industries through the number of heavy industries from census data, and the distance from industrial areas, computed by using OpenStreetMap.

4.2.2.4 *Security perception*

Aesthetics is a crucial element to evaluate neighbourhood livability [118], vitality [2] but also social characteristics such as gentrification [64]. Moreover, crime perception negatively affects housing prices [167]. Thus, we estimate the security perception of places through the pre-trained *Safety Perception* Convolutional Neural Network (CNN) model we presented in Chapter 3. The *Safety Perception* CNN was trained using a crowd-sourced project where people were asked to choose the safer looking place between two randomly chosen Google Street View images. Then, it was fine-tuned with a subset of images from Italy, achieving a $R^2 = 0.62$ over the ground truth labels. Given an image, this model predicts a score in the range $(0, 10)$ where 0 means that a place is perceived unsafe, while 10 is considered safe. We predict the score of each image from our collected Google Street View dataset, and then we average the scores within the *ego-place*.

4.2.2.5 *Living conditions*

The cost of housing is very dependent on the living conditions of people. In a city, people's earnings, economic stress, and unemployment can modify the availability of houses, and their price. We use the average amount of property taxes as a proxy for the living conditions, to establish a spatial baseline of prices.

4.2.3 *The property*

The advertisement of the property is provided with a collection of textual features that contain a brief description, and the list of characteristics (e.g., number of rooms, garage, garden). We pre-process these data, and we use a subset of 25 features that are one-hot encoded or transformed to numbers. The considered properties' attributes are: square meters, built year, energy certification, monthly expenses (condominium), floor number, heating type, type of fixtures, garden, furnished, terrace, sun exposition, kitchen type, spa, cellar, garage, fireplace, place type, property class and type, property taxes, condition, and number of rooms, bathrooms, bedrooms. If the advertisement does not mention it, we assume that some features such as cellar, spa, fireplace are absent.

The property is geo-referenced through latitude and longitude information. Thus, we assign the census block that contains the house as its *ego-place*.

4.2.4 *Model and experimental setting*

In this work, we propose to explore the economic impact of neighbourhoods on housing values through a predictive model based on XGBoost [171]. It is a widely-used model based on Gradient Boosted Trees, able to scale very well even for high-dimensional data such as the one we analyse. Moreover, XGBoost allows to "open" its black box, thus explaining features' importance and negative/positive contributions of each feature to the predicted values. This answers to the intriguing question "Why did the model predicted this home value?", a very important answer that helps real estate companies and local governments make important decisions.

The model is defined as nowcasting the value of houses Y from the property's characteristics P, and the neighbourhood's features E as $P \cdot E \approx Y$.

We train our model using a K-fold cross-validation schema to ensure that the model is robust to unseen neighbourhoods and houses. We divide the original dataset into five folds, assigning three to the training set, one to the validation and one to the test set. Then, we iterate the process shifting the folds. *Nowcasting* spatial data poses some challenges on the creation of cross-validation independent folds. Thus, given the train, validation and hold-out sets:

(i) For each house x in a neighbourhood N_x in the training set, it does not exist a house y in the validation (or hold-out) set that is in the same neighbourhood of x;

(ii) For each house z in a neighbourhood N_z in the validation set, it does not exist a house u in the hold-out set that is in the same neighbourhood of z.

The model is trained with the following parameters: learning rate of 0.001, $\lambda = 5$, $\alpha = 1$, 3 minimum child weight, max depth of 20, and 4,000 estimators, stopping the training when the validation error does not decrease for 50 rounds.

We evaluate the results computing the errors on all cross-validation's hold-out sets through the standard Mean Absolute Error (MAE) and the Median Absolute Percentage Error (MdAPE):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$p_i = \left| \frac{y_i - x_i}{y_i} \right|$$

$$\mathrm{MdAPE} = \mathrm{median}(\{p_1, p_2, \ldots, p_n\}) \cdot 100$$

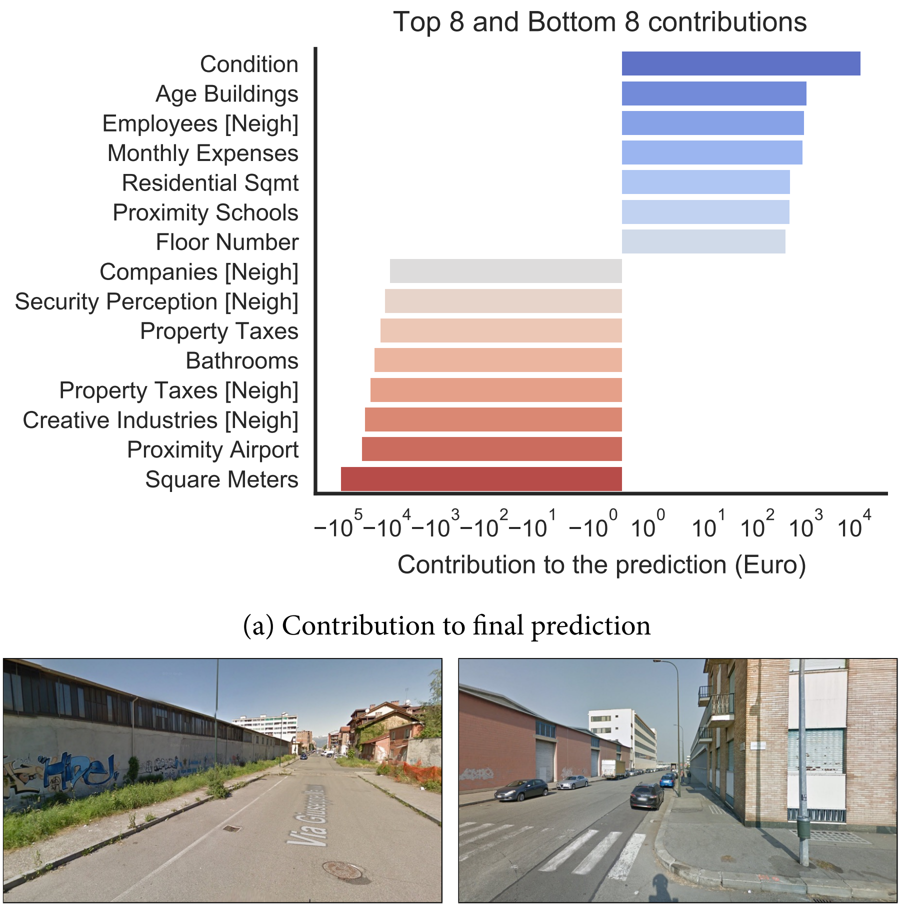
4.2.5 *Open license model*

To the best of our knowledge, there are no released systems that allow to simply deploy, analyze and *nowcast* the real-estate housing market. In this chapter, we release a framework able to download, process and predict properties' market prices from textual information and neighbourhood characteristics. The data used by our Open model comes from heterogeneous sources released with an Open license. This potentially allows end-users from universities and research centers, companies, and governments to replicate and update this study at any time. This model does not make use of relevant data such as Google Street View images and the property's tax information. For this reason we evaluate it separately in Table 13.

4.3 RESULTS

While property's characteristics and market's network effects were widely used to nowcast housing market prices [79], to the best of our knowledge this is the first study that connects the value of houses with its surroundings.

We evaluated our framework for two settings: (i) it uses only textual features of the property, (ii) it uses the characteristics of both the property and the neighbourhood to infer the housing price. Table 13 shows that, among these models, the use of neighbourhood features significantly improves the results, dropping the percentage error by 60%. The MAE across all cross-validation's



(a) Contribution to final prediction

(b) Visual inspection of security perception

Figure 8: Housing price prediction for a house in Corso Grosseto, Turin. (a) Top positive (negative) contributions of each feature to the final predicted value. Here, the low number of square meters of this house drives the final prediction down. (b) Qualitative valuation of security perception in the *egohood* of the house.

Feature name	**Effect on house's value**	**Value**
Condition	↗	Excellent
Bathrooms	↙	1
Security perception	↙	3.82
Creative industries	↙	1

Table 12: Housing price prediction for a house in Corso Grosseto, Turin. (c) Textual features that were very important for this prediction.

folds is around €104,586, which is very promising accounting that housing prices range from €20,000 to €20million. This confirms the important economic role played by the neighbourhood, which previously was only hypothesized.

We also tested an *Open* version of the model II, that uses only data released with an Open license. This model has an error of 18.02%, which is a bit higher than the non-Open model. This is caused by the lack of the *security perception* information and the property taxes of the neighbourhood.

Model	**MAE**	**MdAPE**
I) Property	148,109	23.78%
II) Property + neighbourhood	**104,586**	**15.44%**
III) Property + neighbourhood (Open)	138,929	18.02%

Table 13: The prediction error of real-estate housing prices for three different models. Property uses only textual features of the house (e.g., the number of rooms, floor number). The second model uses both the textual and the *ego-place* features. Its Open version uses only contextual data with an Open license.

The XGBoost model allows us to inspect the most important features among all the predictions. Figure 9 shows that square meters, monthly expenses and the age of the building seem to be the primary drivers of price, which is intuitively reasonable. Other notable possible drivers of property's characteristics are its taxes, the floor number and the condition of the apartment. However, nine of the 15 top-features are related to the neighbourhood. Among them, the number of employees, the population and the proximity of amenities confirms previous studies on the importance of build environment [1]. The presence of creative industries in the *egohood* also seems to have an impact on price, confirming the preliminary results of Hristova *et al.* [162] in London and New York. It is worth noting that, despite the use of many urban co-variates, visual appearance is one of the most important predictive factors for housing value, as similarly proven for other outcomes such as crime [64], happiness [118] and presence of people [2].

4.3.1 *Qualitative/Local Results*

Error metrics and variable importance across all predictions are often not enough to *trust* a model [172]. To make decisions, especially in policy-making and business settings, end-users have to be confident on how the model reaches a given

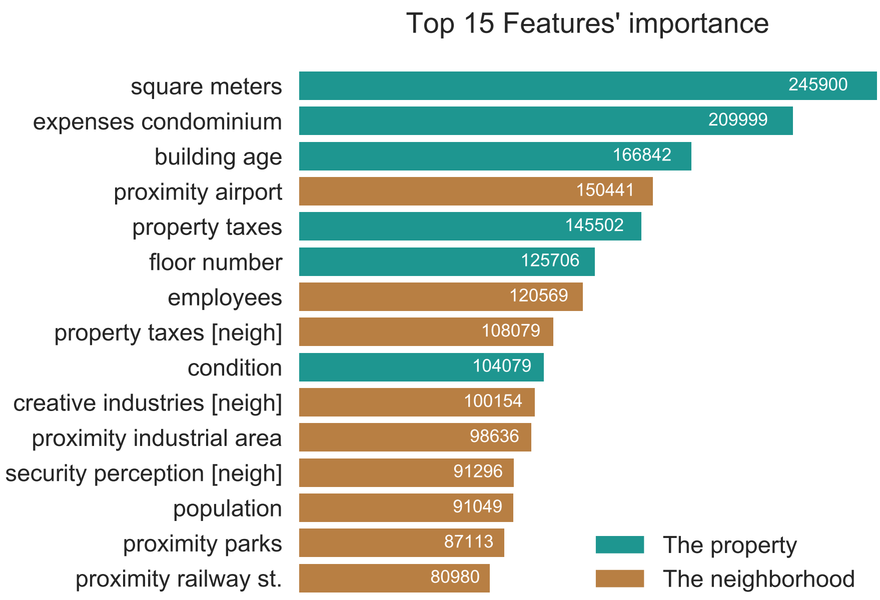


Figure 9: The 15 most important features to predict real-estate housing prices.

decision (e.g., the price assigned to the assessed property). Often, this may be achieved by using simpler models, with a small number of features, which make people more comfortable with interpreting results. Here, we show how individual predictions are achieved, by breaking down the impact of each feature to the final value of the assessed property.

Tree-based models allow to follow the decision tree to understand *how* each prediction is made. Similarly, we follow each of the XGBoost's boosted tree and then sum the contribution of each variable through the decision path. The final prediction can be the interpreted as the sum of the average bias term and the contribution of each feature:

$$y_i = \frac{1}{J}\sum_{j=1}^{J} c_{j_{full}} + \sum_{k=1}^{K}(\frac{1}{J}\sum_{j=1}^{J} contrib_j(x,k))$$

where J is the number of trees.

Thanks to this, we now analyse the prediction of a house in corso Grosseto, Turin, from the hold-out set. Figure 8 (a) and Table 12 show the 8 most and least contributing variables in the prediction. The excellent condition of the house increases the price by 12,572 €, while having just one bathroom penalizes the property. However, the house is placed in a peripheral and working-class neighbourhood of Turin. This emerges from our model. Low-security perception score (3.82), the absence of creative industries and low average property taxes significantly decrease the value. For example, it shrinks by 11,713 €, because of the insecure perception of the *egohood*. A visual inspection of Figure 8 (b) confirms that the lower perception could be related to the presence of graffiti and some disorder typical of the peripheral part of cities.

Figure 10: Alluvial plot showing that the features related to neighbourhood's characteristics are the most important to predict the housing value.

4.4 DISCUSSION AND IMPLICATIONS

Taken together, our results clearly show the economic effect of neighbourhood on housing value. As Figure 10 details, the most important groups of predictors belong to neighbourhood metrics, while property's predictors account less. Our work is not free of limitations. For example, our collected dataset is prone to noise, as we used advertised houses instead of sold ones. Moreover, we did not observe any temporal trend of the market within cities. This influences how the model learns from one city and test into another. To the best of our knowledge there are no other papers with code and released datasets. Since we did not have neither one nor the other, we could not apply our work in different countries and compare the diverse baselines we aforementioned. We praise whoever will release geo-located and detailed housing prices, possibly following the guidelines of Loberto *et al.* [173]. Additionally, we note that in this project we did not use historical prices to nowcast housing price. This choice is motivated by the high cost of obtaining historical sale transaction data, which is proprietary and not publicly available. We propose a method that is applicable where it is impractical to obtain it. Moreover, we stress that the best forecast predictor makes use of previous sale prices and that neighbourhood data might have an impact where few (or no) data is available, and where neighbourhood changed over time. However, these urban changes are usually slow. Although we do not expect a huge improvement on short-term predictions, we see the studied variables very important in medium- and long-term predictions especially in areas undergoing profound changes such as gentrification.

However, we believe our work is the first that quantify the possible economic impact of neighbourhood's factors to this extent. Moreover, our machine learning approach relies on geographical and socio-economic characteristics without the need of historical trades.

From our analyses, we may also draw several implications for citizens, local authorities and urban planners, and for real-estate investors:

For citizens. Buying a house is, without any doubt, one of the biggest investments of people life. A right decision can lead to a satisfactory daily living condition without long commuting time to work, with better health outcomes due to higher *walkability* and less traffic, and even higher profits in the future if the house's value increases. An informed judgment on the value of the property is, thus, necessary. People should consider property's characteristics, but also what surrounds the house to buy. Our results found that people should care about the proximity of amenities (e.g., parks, railway stations) but even intangible elements such as the perception of security, which is one of the most important factors. After their investment, people should avoid the vicious loop of degrading, caring about political decisions such as large-scale urban renewal actions, fighting public incivilities such as litter and graffiti, defending neighbours' habits and characteristics. All these factors have an impact on neighbourhood's liveliness [10] and crime levels [75], and thus affect the house prices.

For local authorities and urban planners. Houses are not islands unto themselves, as they are embedded in complex intricacies of factors in play. Thus, planners have to discourage the creation of oasis in the deserts, where the effect of the neighbourhood is demised. For example, Figure 8 and Figure 9 show that proximity to schools, the presence of amenities and services such as public transport, and absence of creative industries influence housing price. Moreover, changes of the built environment (e.g. the *walkability* and amenities) and spill-over effects of changes in perception can be early signs of gentrification [64]. Data and new computational tools represent invaluable tools to analyze cities empirically.

For real-estate investors. Developers and real-estate investors often think about houses as *units*. This definition should be replaced in favour of *place* and *surroundings*. Investors, especially those working with foreclosures, should examine properties where neighbourhoods have high *vitality*, and a virtuous loop of services and improvements over time. Figure 8 shows that people might prefer to avoid houses near to airports, without creative industries, in neighbourhoods where people can not socialize safely. As Marc Augé argued [174], there are places of every day passing (e.g., malls, airports) where people experience alienation as they cannot live, socialize and have a real identity in the place. He called them *Non-places*. Similarly, neighbourhood without the essential social and *vital* qualities can potentially alienate people; hence, they cannot be considered as an adequate investment in this sense.

4.5 SUMMARY

This chapter applies the notions discussed in Chapter 2 and Chapter 3 to study the economic impact of neighbourhood characteristics on housing values. We formalized this problem through a multi-modal now-casting prediction where we infer the price of houses in unseen neighbourhoods' conditions. Thus, we collected a large dataset of real estate properties and computed neighbourhood features based on overlapping geographical boundaries called *egohoods*.

We trained a Gradient Boosting algorithm on 8 Italian cities, and we found that neighbourhood characteristics seem to drive more than 20% of the house's advertised price. Our results also show that the use of this information in the model lowers the prediction error by 60%. The qualitative analysis clearly demonstrates the soundness of the proposed solution. We also released the code and dataset to foster research in this novel direction.

Indeed, our work suggests several open questions for future work. Can we design experiments and housing policies that can be responsive to neighbourhood's changes? Is there a trade-off between urban well-being, the economic success of cities and affordable housing? Can the neighbourhood's physical environment predict gentrification? How can we deploy a pipeline able to react to multiple conditions and work adequately in different countries? This last question will be instrumental in the analysis of the next Chapter.

5 THE INTRICATED FACTORS OF CRIME

Behind every crime is a story of sadness.

— Enrique Pena Nieto

In criminology, social cohesion among neighbours has been linked to their willingness to cooperate in order to solve common problems and reduce violence [35, 175, 176]. Cooperation, as opposed to disorganization, of neighbours is indeed believed to create the mechanisms by which residents themselves achieve guardianship and public order [35]. This mechanism, also called natural surveillance, finds its roots in urban planning, where the relationship between specific aspects of urban architecture [74] and urban physical characteristics [10] are related to security. However, neighbourhoods are not to be considered islands unto themselves, as they are embedded in a city-wide system of social interactions. On a daily basis, the people's routine exposes residents to different conditions, possibilities [86], and it may favor crime [177]. Yet, mainstream studies focus on just a subset of static factors at a time, often in a single city (e.g., Chicago or New York), thus neglecting the complex urban interplay between crime, people, places, culture and human mobility. Hence, our understanding of the factors influencing crime across cultures and cities is very limited.

In this chapter, we seek to shed light on the diverse set of factors at play with urban crime exploring how this is related, at the same time, to social disorganization, the built environment, and mobility of people. Moreover, we address the need of a comprehensive study that explores crime theories across multiple cities of the world, analysing Bogotá, Boston, Los Angeles and Chicago. We present a Bayesian hierarchical model that uses massive and ubiquitous data sources such as mobile phone records and OpenStreetMap. The resulting framework is potentially replicable at scale, to analyze crime patterns and use the generated insights to make recommendations for policies and initiatives that could be the most effective in improving urban citizen security. From our data, it emerges that previous theories are often conflicting when analysed across cities. However, we show that accounting for the place, mobility, and socio-economic conditions together can give a more reliable picture of crime patterns.

The chapter is based upon De Nadai *et al.* [4] and it is organized as follows. Section 5.1 describes the state of the art; Section 5.2 describes the methods and the data sources. In Section 5.4 we discuss the results of our analysis.

5.1 STATE OF THE ART

Urban sociologists, urban planners, and criminologists have long been interested in studying the conditions that influence crime in neighbourhoods [169, 176, 178, 179]. The primary focus of research in criminology has been on individuals and the environmental conditions that increase the likelihood of crime. Thus, scholars have first sought to understand why certain individuals, as opposed to others, become criminals [180], how daily routine activities shape criminality [177], or how neighbourhood-level variables influence the level of violence [28, 181]. At the neighbourhood level, crime has often been associated to the presence of an offender, its suitable target, and the absence of any deterrence system such as police or even ordinary citizens [177, 182]. The concept of natural surveillance by ordinary citizens spans from criminology to sociology and urban planning. Notably, urban activist Jane Jacobs [10] argued that "a well-used city street is apt to be a safe street and a deserted city street is apt to be unsafe." She states that four conditions have to be valid at the same time to ensure a virtuous loop between the presence of people and "eyes on the street." First, a district should serve at least two or more functions to have streets continuously used by residents and strangers. Second, street blocks should be small and short to have both high *walkability* and a mixture of people that frequently meet each other at street intersections. Third, diverse buildings make it possible to have low- and high-rent spaces, and thus a mixture of people and enterprises. The fourth condition is about dense concentration, which ensures a sufficient presence of people and enterprises to attract dwellers from different neighbourhoods continuously.

Daily commuting has been found to influence the attractiveness of a place or person to predatory crime, as people mobility and presence influence the supervision and surveillance in a place [177]. Previous studies have also shown that poorly connected neighbourhoods connect people to a smaller pool of diverse social interactions, services, and jobs [86], hence weakening their cooperation [183, 184]. Others have even observed that mobility can predict crime [185]. Altogether, these studies confirm the tight, and unexplored, relation of mobility, cooperation, and place.

The goal of this research goes beyond crime prediction. Offenses are concentrated in a small number of places [27] and they are tightly coupled with places, stable over time [186]. Thus, the easiest way to predict crime is modelling those few places with the highest number of crimes, also known as *hotspots* [59, 187]. On the contrary, we seek to shed light on the diverse set of factors at play with urban crime. To the best of our knowledge, there is no comprehensive study exploring how crime is related, at the same time, to social disorganization, physical characteristics of the city and mobility of people. Mainstream studies focus on just one city (usually Chicago or New York), thus limiting the impact of the conclusions drawn, and neglecting the physical characteristics, or the built environment from criminology research. Altogether these limitations result in a fragmented and incomplete picture of how the numerous factors influence crime in the urban context.

5.2 OUR APPROACH

In this chapter we want to study crime in cities through a multi-variate regression that takes into account, at the same time, the proxy variables for the Social disorganization (SD) theory, the Jane Jacobs theory, and the mobility of people. Taking this into account, we extract SD variables from census data, the build environment (BE) from both census and geographical data, while mobility is extracted from mobile phone traces. Here, we explain our approach in collecting the data, and modelling the problem.

5.2.1 *Spatial aggregation*

As we want to study the neighbourhood effect on crime, it is necessary to define what a neighbourhood is. Criminology literature often relied on either Chicago's community areas [1], or census tracts. The former were defined by the University of Chicago for statistical and planning purposes and often used by the Chicago school of criminology to analyse Social Disorganization. The latter are defined for census reasons to be homogeneous units with respect to population characteristics and socio-economic conditions, with an average population of 4,000[2]. Census tracts, however, are not defined to account for the perceived boundaries and behaviour of people who live them.

---

1 Here it is possible to have more information about community areas https://www.chicago.gov/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Community_Areas/Community_Areas_W_Numbers.pdf

2 As specified from the definition at https://factfinder.census.gov/help/en/census_tract.htm

In this chapter, we thus rely on neighbourhoods definitions drawn by the municipality. In Chicago, we rely on the community areas, according to the literature. In Boston and Los Angeles, we use the definition draw by each city, in collaboration with its citizens. In Bogotà we rely on the Unidad de Planeamiento Zonal (UPZ), drawn for urban planning reasons. All the metrics and data computed at higher spatial granularity are aggregated to these geographical units of analysis.

5.2.2 *Sets of data*

Crime data. The measurement of crime has not to be taken for granted. In criminology, there are usually two ways to assess crime: crime counts and rates. The former are typically defined in papers where hotspots of crime are being detected, while the latter can be found in papers where the population at risk is of interest. These rates are commonly calculated by dividing the crime counts with the population at risk, which is usually the residential one. Recently, some scholars have discussed the costs and benefits of alternative denominators such as the ambient population [188]. Thus, they computed the average number of people in a place through satellite imagery [188], census data [189] and Foursquare check-ins [190]. However, neither it is clear whether using residential rates, ambient rates, nor the policy implications of using one or the other. Moreover, the bias of ambient rates can be potentially lead to a misleading interpretation of crime.

In this research, we thus prefer to describe crime counts adjusting for residential and ambient population. Hence, we describe their relative role through the β coefficients of the regression.

Data collection mechanisms and crime categories can vary from country to country, or even at the city level. In the US, the Uniform Crime Reporting (UCR) organizes crimes into two main groups: Part 1 and Part 2 offences. The former is composed by violent crimes (aggravated assault, forcible rape, robbery, and murder), and property crimes (larceny-theft, motor vehicle theft, burglary, and arson), while the latter are considered less serious and they include offenses such as simple assaults and nuisance crimes. To be consistent with this categorization, we mapped crime categories in Bogotá to UCR Part 1 and Part 2. This mapping is available in Appendix A.3.

For each city, we thus collect the geo-referenced data of committed crimes, and we filter out those crimes not belonging to Part 1 of UCR, similarly to most of the criminology literature.

Mobile phone data. Our work is not the first one using mobility data to understand crime. Previous literature has indeed used employment [183], taxi [185] flows, which suffers from limitations including sampling and reporting bias.

Here, we use mobile phone data, as it proved to be an invaluable way to overcome the high cost of gathering travel surveys [191]. The data is collected for billing purposes by two mobile operators in Bogotá, Boston, and Los Angeles. The time-frame of the mobile phone activity is 12-01, 2013 to 05-31, 2014 (Bogotá), six weeks in 2010 (Boston), and 10-15, 2012 to 11-24, 2012 (Los Angeles).

Spatial and census data. Census blocks, population, employment and poverty for US cities were drawn from the American Community Survey (ACS)[3]. For US cities we also used some city-specific datasets that are described in the Appendix A.1 and Appendix A.1. The census data of Bogotá was obtained by the Departmento Administrativo Nacional de Estadística (DANE), which organized the 2005 general census for the city[4]. The poverty data of Bogotá was extracted from the Multipurpose Survey (EM) of 2014. The detailed list of datasets and URLs are listed in the Appendix A.

5.2.3 *Built environment*

To study the urban planning conditions argued by Jane Jacobs, we use some of

We describe the four Jane Jacobs conditions through the operationalized metrics previously defined in Chapter 2. The analysis, in order to be effective, has however to be at easy interpretation to further understand the causes of crime and prevent it. For this reason, we used only a subset of the aforementioned metrics. Specifically, we employ one metric for each condition, namely Land Use Mix, Small blocks, Building mix, Building density and Walkscore. The use of less indexes to describe the built environment allows also the collection of less amount of data across the cities, which might have different types of data released.

The Land Use Mix, previously defined in Equation (1), is computed as the average entropy among land uses:

$$\mathrm{LUM}_{L,i} = -\sum_{j \in L} \frac{P_{i,j} \log(P_{i,j})}{\log(|L|)}$$

where $P_{i,j}$ is the percentage of square meters having land use j in unit i, and $L = \{\text{residential, commercial and institutional, park and recreational}\}$ represents

3 https://www.census.gov/programs-surveys/acs
4 http://www.dane.gov.co

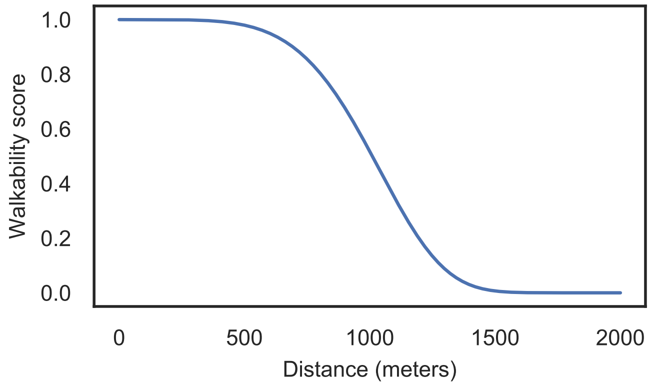


Figure 11: Polynomial distance decay function for the Walkability score.

sents the considered land uses in the metric. The LUM ranges between 0, wherein the unit is composed by only one land use (e.g., residential), and 1, wherein developed area is equally shared among the n land-uses.

To further characterize the diversity of land uses, we determine the walkability of a neighbourhood through its accessibility to the nearest point of interests (*e.g.*, convenience stores, restaurants, sport facilities). As previously described in Section 4.2.2.1, we formalize the walkability score as the weighted sum of the walk distances $s_{i,j}$ from a block i to all the amenities j, which belong to a category (e.g. Restaurants). We define the walk distance as:

$$s_{i,j} = e^{-5(d(i,j)/M)^5} \tag{36}$$

where $d(i,j)$ is the number of meters of the shortest path between i and j, along the road network (see Figure 12). M is the maximum walking distance considered, which is set to 1. This score assigns the maximum score to amenities ~ 500 meters far from the starting point, then the score decays quickly until 1500 meters, where it first slows down then it goes to zero. The function is shown in Figure 11.

All the amenities are downloaded from Foursquare. We refer to Section 4.2.2.1 for further details of this method.

As previously defined in Equation (10), we compute the average block area among the set B_i of blocks in unit i as:

$$\text{Blocks area}_i = \frac{1}{|B_i|} \sum_{b \in B_i} \text{area}(b)$$

In the previous chapters we defined the building diversity through metrics that measured the mean and standard deviation of buildings age (see Equation (16) and Equation (14)). These indexes were motivated through the importance that Jane Jacobs demonstrated towards the diversity of buildings age.

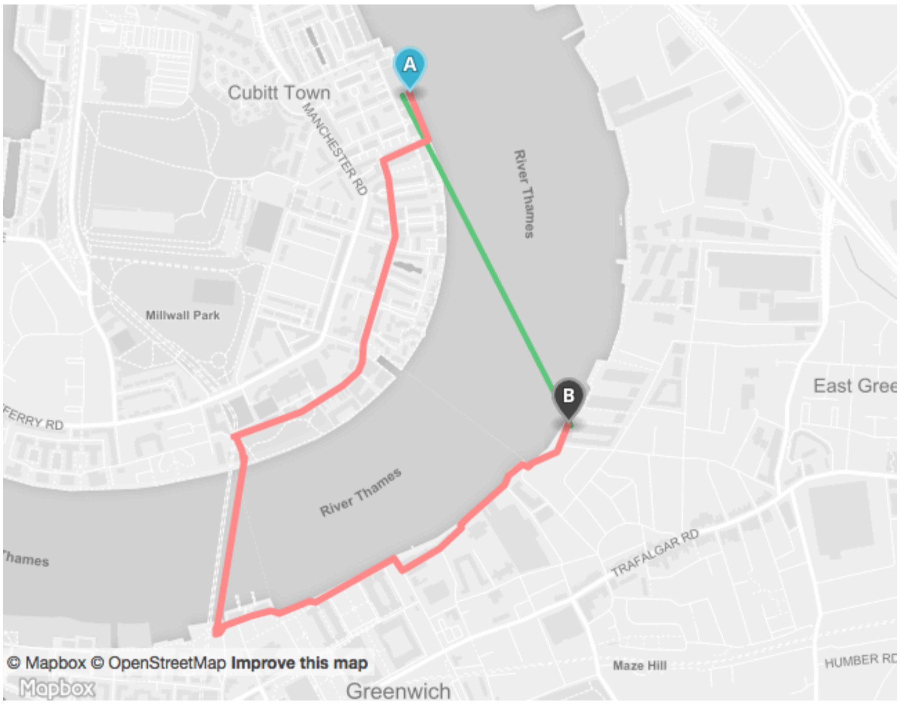


Figure 12: The geometric distance (green) and shortest path walking distance (red) between two points.

A mix of old and new buildings is indeed supposed to thrive local economy by making available to companies and people high- and low-rent places.

Thus, we here compute the building diversity as the weighted standard deviation of the property values in a neighbourhood, weighted on the population of each block. In Boston and Los Angeles we first get all the parcels' values and population for each block group, then we compute the weighted standard deviation for each neighbourhood. For Bogotá, we calculate the property values through the stratum, which classifies buildings in different regimes of tax payments for utilities and rents in Colombia. Stratum is based on the external physical characteristics of the building, and it reflects the quality of life of residents with a six-level classification from 1 (low) to 6 (high).

Jacobs' fourth and final condition is about concentration of both buildings and people. Similar to Section 2.3.2.4, we determine population density by counting the number of people by net acre (the area of the land zoned for housing). In a neighbourhood with small blocks the land covered by streets and their intersection is higher, thus this lowers the density measures that use the total area of the neighbourhood:

$$\mathrm{Pd_i} = \frac{|\mathrm{Population}_i|}{\mathrm{netarea}_i}$$

Then we compute the square meters density per each unit i as:

$$\text{Buildings density}_i = \frac{1}{|E_i|} \sum_{b \in E_i} \text{sqft}(b)$$

where E_i is the set of buildings in neighbourhood i and sqft(b) is the sum of square meters of building b.

5.2.4 *Social-disorganization*

The Social disorganization theory can be traced back to the Chicago schools of sociology, which defined it as the inability of communication and cooperation to realize common goals and achieve effective social control. The original theory was operationalized mainly through three variables: disadvantage, ethnic heterogeneity, and residential stability. Although it was later revitalized with stringer definitions of social capital and control, the original theory has still the advantage to be defined from census data at scale. Thus, we rely on the original definition.

Disadvantage is defined by the percentage of people living below the poverty line and by the percentage of unemployment. Since these variables are highly intra-correlated, we obtain this index through their first eigenvector from Factor Analysis. Previous studies in Chicago also incorporated in this index the percentage of people receiving public assistance, the percentage of female-headed families with children and the percentage of black people. We do not include race-specific variables other than diversity because they might be present only in some places and not in others (e.g. native Americans in Bogotá), and to avoid any race-specific bias.

Ethnic diversity represents the difficulties of a community to communicate and collaborate for a common goal. Accordingly to the literature, it is computed as the Hirschman-Herfindahl diversity index of six population groups:

$$H = 1 - \sum_{i=1}^{N} s_i^2$$

where s_i is the proportion of people belonging to the race i, and N is the number of races. Consistently with the literature we include for US cities: Hispanics, non-Hispanic Blacks, Whites, Asians, Native Hawaiians - Pacific Islanders and others. For Bogotá we include: Indigenous, Rom, Islanders (San Andrés), Palenquero, Black and others.

Similarly to disadvantage, residential stability is calculated from the first eigenvector from Factor Analysis of the percentage of house owners and the percentage of residents who lived in the same house one year earlier for US cities, and five years before for Bogotá. This controls for the rapid population turnover in neighbourhoods, which is supposed to lower the people's attachment to the district.

5.2.5 *Mobility*

Differently from previous chapters, instead of using the raw data, we use a model that describes the mobility and presence of people in the city. This allows to describe the movements of people from each area of the city to all the others with is more similar to the ground truth. First, we fit the data through the TimeGeo framework [191], state of the art for mobility modelling. Then, we simulate individual trajectories that are close to the real behaviour of people. Finally, we aggregate flows by origin and destination to generate a Origin-Destination (OD) matrix, and we aggregate by source and destination to generate the average amount of people in a neighbourhood, which we call ambient population [192].

We note that we use the TimeGeo framework to simulate data similar to the human mobility through the year, and we selected all other sources of data (e.g crime data) at the nearest time interval data available.

5.2.6 *Bayesian model*

Let y_i be the discrete number of crimes for a set of spatial regions $i = \ldots, N$. We approximate this behaviour with a Poisson regression, which is appropriate to model the non-negative nature of the crime-counts in a city. The Poisson regression can be written as:

$$y_i \sim \text{Poisson}(E_i \exp(X_i \beta)) \tag{37}$$

where E_i is the expected number of offences in absence of covariates effects (also called offset variable), β are the coefficients and X_i are the region-specific covariates that could be correlated with the dependent variable y_i. The restrictive assumption of the Poisson distribution to have identical mean and variance is violated in many real-world situations. Crime is not exception. Figure 13 shows that is indeed highly clustered in a small number of places and that there is high variation from one area to another.

Thus, the model in Equation (37) can be expanded to account additional random effects that accommodate overdispersion. Formal tests on the presence of overdispersion have confirmed this assumption (see Appendix A.2).

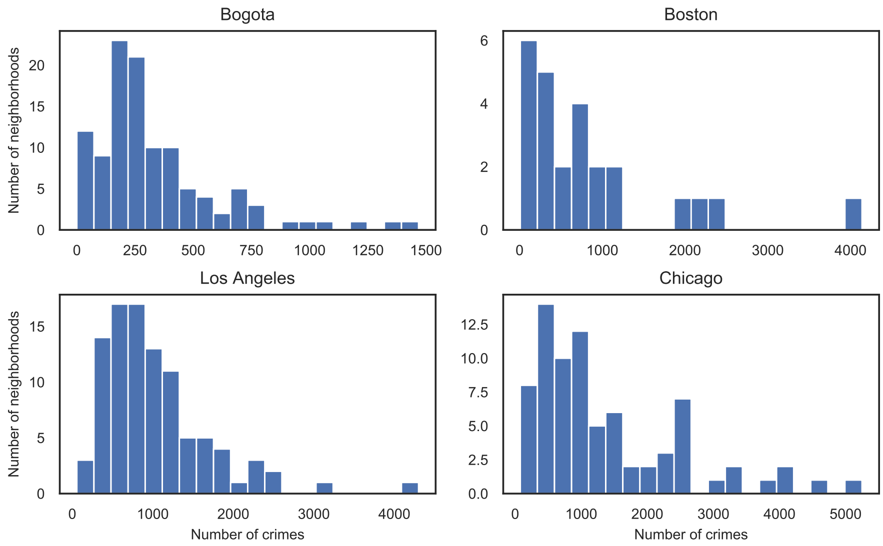


Figure 13: Distribution of the number of total committed crimes for each neighbourhood.

In Bayesian models the spatial-autocorrelation and random effects are usually dealt with a $\boldsymbol{b}$ term in the Poisson equation:

$$log(E(Y_i)) = \sum_{k=1}^{n} X_k \beta_k + \boldsymbol{b} \tag{38}$$

where $\boldsymbol{b} = (b1, \ldots, b_N)^t$ represent the residual spatial effect. In this particular application we define the prior of this effect with a Gaussian Markov Random (MRF) field prior:

$$\boldsymbol{b} \sim \mathrm{MVN}(\mathbf{0}, \Sigma(\boldsymbol{\theta})) \tag{39}$$

where $\boldsymbol{\theta}$ represents a vector of variance component parameters. Although many MRF models are discussed in literature, we here refer to the Leroux model [193] where $\boldsymbol{b}$ can be represented as a conditional autoregressive (CAR) model with the conditional distribution of b_j being:

$$b_j \sim \mathrm{N}\left(\frac{\lambda_1}{1 - \lambda_1 + \# W_j^{(1)}} \sum_{j \sim k} b_k, \frac{\sigma^2}{1 - \lambda_1 + \# W_j^{(1)}}\right), j = 1, \ldots, N \tag{40}$$

where $0 \leq \lambda < 1$ is the spatial auto-correlation term, $\# W_j^{(1)}$ denotes the number of neighbours for area j and $j \sim k$ indicates that area j is neighbour of area k. Noticeably, when $\lambda = 0$ there is Gaussian independence among the areas.

As we have no prior knowledge about the model hyperparameters $(\boldsymbol{\beta}, \sigma^2, \lambda)$, these parameters are set with non-informative vague hyperpriors. We set the priors related to the spatial random effect as:

$$\tau = \sigma^{-2} \sim C^+(0,1)$$
$$\lambda_1 \sim \text{Uniform}(0,1)$$

where $C^+(0,1)$ is a standard half-Cauchy distribution on the positive reals, chosen as a scale parameter to have a better convergence [194] than the previously used Inverse gamma prior [193].

To account for mobility of each neighbourhood, we normalize the number of outgoing people with the population level at the home neighbourhood. Then, we binarize the matrix by setting to 1 all cells with mobility rate higher than a threshold, 0 otherwise. Differently from previous approaches [183], this threshold is estimated in a Bayesian way with a Uniform prior. Thanks to this matrix $W^{(2)}$ we account for mobility in a similar way to spatial auto-correlation. Thus, we extend the CAR [195] model with:

$$E(b_i|b_{-i}) = \frac{\lambda_1 \sum_{j \in W_i^{(1)}} b_k + \lambda_2 \sum_{l \in W_i^{(2)}} b_l}{1 - (\lambda_1 + \lambda_2) + (\lambda_1 \# W_i^{(1)} + \lambda_2 \# W_i^2)}$$
$$var(b_i|b_{-i}) = \frac{\sigma^2}{1 - (\lambda_1 + \lambda_2) + (\lambda_1 \# W_i^{(1)} + \lambda_2 \# W_i^{(2)})}$$

where λ_2 has the same vague prior of λ_1.

As we have no prior knowledge about the model hyperparameters, but we have to account for collinearity and outliers, we choose to have a robust estimate having Student-t priors for $\boldsymbol{\beta}$.

The model was calibrated by means of Markov Chain Monte Carlo (MCMC) approach. Convergence was assured by the Gelman-Rubin convergence statistics, and discarding the first 25,000 iterations and running the model over 15,000 iterations. The model was evaluated through the DIC [196] to assess both the model complexity and fit. Moreover, we also computed the Bayesian predictive information criterion (BPIC) [197] that supposedly overcomes the over-fitting problems of DIC. Since it is not a common used metric as DIC, we prefer to show it in the SI. As a rule of thumb, differences of more than 10 of DIC rule out the model with higher DIC, while differences between 5 and 10 are still considered substantial. However, when the difference is less than 5 the model can be considered comparable.

5.3 RESULTS

While the effects of urban environment's characteristics, socio-economic conditions, and mobility have been empirically tested separately [1, 176, 183, 198, 199], to the best of our knowledge, this is the first study to support with large-scale data the association of crime with socio-economic conditions, the built environment, and the mobility.

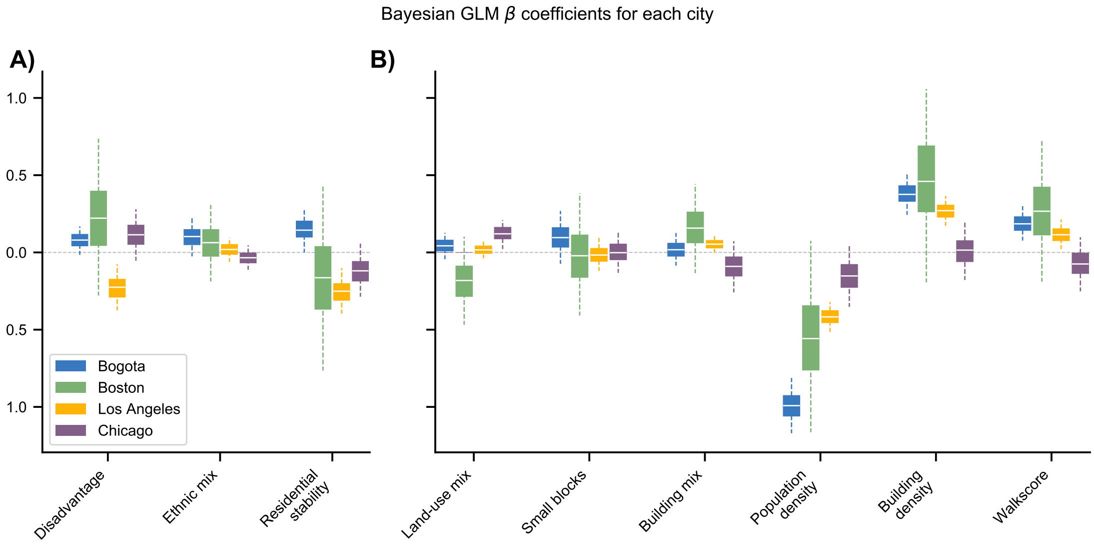


Figure 14: Generalized Linear Model's β coefficients showing that Social disorganization (A) and Built environment (BE) variables do not act with the same strength and direction in all cities, and only some of them are clearly contributing to the fit. All together this figure shows that there is not a universal explanation of crime from previous theories.

We begin by presenting the results of the Social-disorganization (SD) model, which describes crime only through socio-economic variables. neighbourhoods with higher level of disadvantage, ethnic diversity and residential instability are posited to have more social disorganization and, hence, crime [176]. However, Figure 14 A) shows that SD's variables do not act with the same strength and direction in all cities, and only some of them are clearly contributing to the fit. For example, disadvantage is associated with higher crime in Bogotá and Chicago, while the opposite is true in Los Angeles. Similarly, places with low crime have residential stability in Los Angeles and Chicago, while in Bogotá residential stability is associated with higher crime. Ethnic diversity is, in agreement with the Social-disorganization theory, positively correlated with crime in Los Angeles, Bogotá, and Boston, even though the credible intervals sometimes comprehend zero. In Chicago the contrary seems true. Taken all together, these disagreements on the sign of the correlations suggest that conventional metrics based on the original Social-disorganization theory [176] might not be sufficient to explain crime in the same way in different cities around the world.

The built environment affects trust and the sense of community and participation [10, 198]. This emphasizes the relationship of social disorganization with the neighbourhood's built environment. Table 14 (row IV) shows that the Median Absolute Percentage Error (MdAPE) and the Deviance Information Criterion (DIC) decreases if we add the characteristics of the neighbourhood to the SD model, at the price of an increased model complexity (p_D). We also see from Table 14 (row II) that the performance of the Built environment (BE) model is very similar to the SD one, which represent the model using the classic variables of criminology. This means that the built environment is tightly coupled with the people living it [186], and that the use of this information makes possible to achieve a good description of crime, even disregarding socio-economic characteristics related to people.

However, urban characteristics from Jane Jacobs' theory do not play the same role in all cities, as Figure 14 B) shows. In Bogotá, Boston and Los Angeles places with high *walkability* seem to have higher crime, while Chicago do not. Higher land-use mix and low building mix in Chicago relate to higher crime, while in Boston high land-use mix and high building mix are linked to low crime.

The difference on the results of each city once again suggests that crime correlates differently with space and people. The urban environment is used differently in each city. Small blocks in U.S., for example, are often those *walkable* and well-to-do places with low-rise buildings. On the contrary, in Bogotá they are located in low-income and distressed neighbourhoods (see Figure 15). Thus, these coefficients often go against Jane Jacobs theory. The *informal surveillance* of the theory is based on the presence, at the same time, of land-use mix, small blocks, building diversity, high population and building density. Here, we found no city in a complete agreement with this expectation. In particular, crime is positively correlated with the mixture of functions, which supposedly brings people continuously to the neighbourhood and, hence, "eyes on the street".

People are everyday exposed to different conditions as they typically travel distances that exceeds the size of neighbourhoods. In line with routine activity theory [177], work and mobility also influence the probability of being targets of crime. Thus, we also consider the mobility of people in and to neighbourhoods of the same city as important factor associated with crime. We compute the Origin-Destination (OD) matrix and the number of people that are usually present in a place (ambient population) from mobile phone data, pas-

Model		**Bogotá**		**Boston**	
		DIC (p_D)	MdAPE	DIC (p_D)	MdAPE
I	Social-disorganization (SD)	1235 (115)	44%	330 (30)	43%
II	Built environment (BE)	1164 (119)	24%	326 (33)	31%
III	Mobility (M)	1112 (137)	37%	335 (35)	40%
IV	SD+BE	1159 (120)	22%	326 (37)	21%
V	SD+M	1193 (223)	33%	335 (56)	51%
VI	BE+M	1092 (228)	23%	334 (60)	**32%**
VII	SD+BE+M (Full)	1080 (237)	**19%**	315 (65)	38%

		Los Angeles		**Chicago**	
		DIC (p_D)	MdAPE	DIC (p_D)	MdAPE
I	Social-disorganization (SD)	1111 (103)	22%	890 (82)	34%
II	Built environment (BE)	1056 (105)	22%	907 (85)	41%
III	Mobility (M)	894 (191)	21%	-	-
IV	SD+BE	1052 (109)	19%	890 (89)	**27%**
V	SD+M	911 (105)	18%	-	-
VI	BE+M	952 (205)	17%	-	-
VII	SD+BE+M (Full)	937 (208)	**15%**	-	-

Table 14: Results of the crime count models applied for each city. DIC is the Deviance Information Criterion, p_D is the deviance derived from the expected value of the parameters (aka complexity of the model), MdAPE is the Median Absolute Percentage Error between the estimated number of crimes and the real one. Chicago is shown only for comparison purposes with previous literature, as we do not have access to its mobility information.

sively collected for billing purpose. Table 14 (row VII) shows that the jointly use of SD, BE and mobility (M) variables has the lowest MdAPE and DIC, with the only exception of Boston, which suffers from the small number of neighbourhoods. This result means that the best model accounts for all the aspects of urban life: Social-disorganization of people, the physical characteristics of the neighbourhoods, and the mobility. Particularly, the ambient population is one of the most important variables in the model as it seems to better assess the number of people at risk in a place as suggested by previous works on aggregated mobility [200], satellite imagery [192], Twitter [201] and census data [189]. Mobility flows further improve the fit of the Ambient population model, as shown in SI. Table 14 (row III) shows that also the use of mobility information has a very competitive fit. However, this model only explains crime through the presence of people (ambient population) and the interactions between places. Thus, there is little point in using only this model to describe ecological factors related to crime.

From Figure 16 we can see the improvement of the fit through the models in Bogotá comparing each map with the ground truth (D). From left to right, SD variables over-estimate crime in the eastern part of the city, where people are more disadvantaged (see Fig. S8 G)). The results of the Built environment (BE) model are very similar to the ground truth but it over-estimates the northwestern part of the city, where residents are more wealthy (see Fig. S8 G)). On the contrary, the SD model in Boston over-estimates crime in well-to-do areas, while the BE model over-estimates crimes in single land use areas such as the airport (see Fig. S9 G)).

Finally, the map of the Full model is almost equal to the ground truth map and its residuals are showed in the right-most map. In the supplementary ma-

(a) Carrera 91c #2-55, Bogotá, Colombia (b) 23 Melrose St, Boston, Massachusetts

Figure 15: Very different characteristics of places in neighbourhoods with the lowest highest Small blocks index in Bogotá (a) and Boston (b).

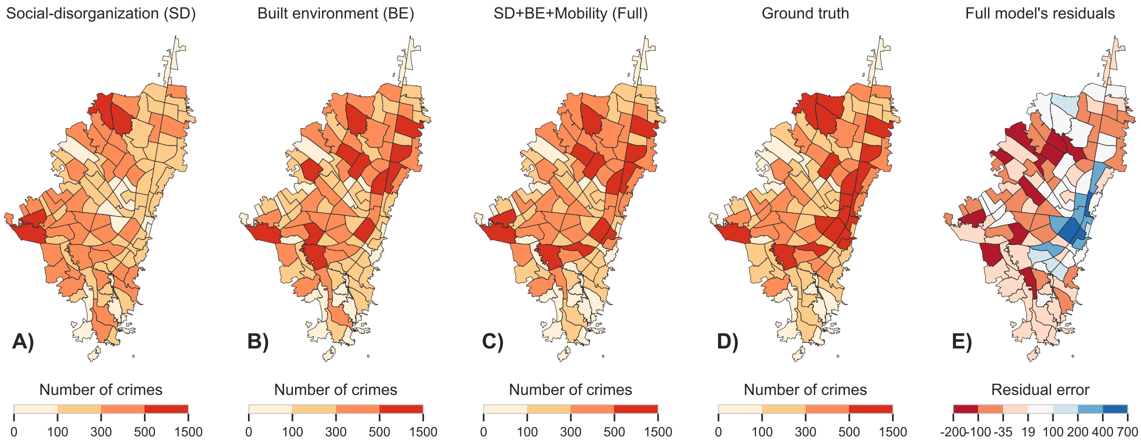


Figure 16: Maps of the estimated number of crime for each neighbourhood in Bogotá for the A) Social-disorganization, B) Built environment, C) Full model. D) shows the ground truth crime count. E) shows the spatial pattern of the Full model's residuals. The colors of the maps are chosen through the Natural Breaks algorithm.

terials (SI), we contrast the relationship of all the factors with violent and property crime composing the total misbehaviour.

5.4 DISCUSSION

In this chapter we analysed four different cities, very different in cultural, economical, historical and geographical aspects. The great variability of the dynamics and history of each city poses a challenge to the existence of a model that "fits it all" able to learn from one city and to predict on another one. Our cumulative results also confirm the lack of a unified framework that describes how the socio-economical characteristics of the city are correlated with crime. Instead, we present a model able to describe and disentangle the role of diverse factors in urban crime and draw some theoretical and practical implications.

The finding that the built environment well describes crime supports the theory of the strong relationship of urban design [74] and planning [10] with the phenomenon. The place is thus not only logically required but its characteristics influence the likelihood of crime [186]. Despite this, Jane Jacobs' core assumptions have to be considered in context. High diversity and density, small blocks, land-use-mix, and vacuums are not characteristics to be widely appraised. While small blocks are linked to beautiful, low-rise houses of Greenwich Village, they could also be linked to slums in Latin America. This is because her theories were mainly based on observations of the mid-20th century New York, and they can fit diversely in different contexts. However, the study of the place should be taken into consideration because it can provide insights to policymakers to reduce crime. Moreover, the finding that built environment describes the crime as well (or better) as socio-economical conditions is particularly relevant for those cities where census data is not available. Thus, crime can also be reliably described through geographic data available in the public domain.

The importance of mobility in our results confirms the hypothesis that accounting for routine exposure to multiple neighbourhoods is important for crime [183]. Small flows relate to higher spatial mismatch and social isolation [86, 202], while higher external connectivity improves trust and other dimensions of internal social order [28]. However, the strong effect of the mobility flow's coefficient in the models suggest that crime propagates between neighbourhoods, as flows create hyperlinks that connect non-adjacent regions and can thus spread people, diversity, opportunities, but also the crime [185]. In fact, there might be social processes at work that make the level of crime change with actions and activities occurring in other neighbourhoods, depending on the social distance between them [203].

Mobility flows and ambient population strongly relate to crime likelihood in cities. As previously observed with social interactions [204], we show that behavioural observations from mobile phone data are more informative than static, census information. Mobile phone activity is useful in estimating poverty [205], well-being [206], unemployment [207] and, thus, most of the Social-disorganization variables in real-time. Therefore, it is crucial to consider continuously updated geographical information (i.e. OpenStreetMap) and mobile phone data to achieve societal goals and develop effective crime prevention strategies.

Our results demonstrate that the combination of diverse sources of information such as urban environment, socio-economical characteristics, and mobility better describe the crime that any of these factors alone. Moreover, we show that considering only a subset of these variables can lead to discrepancies and incomplete pictures of the dynamics of the city. If crime control is the goal, targeting crime through interventions on socio-economic problems only [208] might not be as effective as considering the place and people's mobility as well.

Our results do not come without limitations. We describe crime through the social disorganization theory and its connection to the informal surveillance theory. To do so, we use several proxies for the built environment and the socio-economical characteristics of residents, which might not correspond to the real collaboration of people. Other variables, based on the friendship network, education and family might better approximate the social-disorganization of a community. However, these variables are usually complicated to collect, especially at scale, for different cities. We leave to further studies how to extract proxies of social-disorganization from alternative sources such as mobile phone data. We note, moreover, that we do not focus on forecasting. We stress that to predict future crime researchers should focus on dynamic information such as the mobility of people (e.g. [209]) and base future prediction on previous crimes as well. Here, we are instead interested in the possible underlying mechanisms to identify causes or suitable interventions to ameliorate and prevent crime.

Most of the traditional criminology analysis rely on frequentist p-values and confidence intervals that are based on hypothetical unobserved data, and thus do not allow an interpretation of the underlying phenomena. On the contrary, we present a Bayesian analysis that provides the highest density intervals (HDI), which are based only on observed data, and describe the distribution of the most credible values of each parameter. Moreover, our framework can shrink down to zero the variables that are non-necessary to the description of the crime. In cities many variables are collinear to the others, and this was often neglected in previous approaches. Thus, literature often describes crime with complex models without regularization, that could lead to false interpretation of the coefficients and possibly wrong results for policymakers.

5.5 SUMMARY

In summary, we have shown that building environment has an effect on crime, and this can be instrumental especially in cities where census variables are difficult to collect. Moreover, we compared two alternative and almost static definitions of neighbourhood effect, namely the Social disorganization theory and the Jane Jacobs theory. Then, we modelled them together with the dynamics of people extracted from mobile phone data and shown that the best description of crime requires to model the socio-economic conditions, built-environment and mobility together.

Our work seeks to make headway on the previous limitation of a single site of study origin. Analysing multiple cities together exposed criminology theories to discrepancies and differences. Transferring what we learn from one city to another remains an open question.

Descriptive modelling can help policymakers to understand the use of urban space and deploy future investments and resources thoughtfully. Moreover, from the scientific perspective, descriptive modelling can provide insights for strong predictors, and potentially for explanatory variables, to be further explored by explanatory modelling and experiments [210]. Thus, we hope that additional research keeps exploring multi-dimensional aspects related to crime, to clarify potential crime causes and design better cities.

Part IV

CONCLUSION

Self-explanatory title.

6 CONCLUSION

To understand is to perceive patterns.

— Isaiah Berlin [211]

Globalization and the Internet make us believe that we live in a flat world, where there are no places nor geographical distances. However, local patterns do exist and influence our perception and social relations, with consequences that span beyond the local boundaries. Any investigation aiming at understanding cities must take into account this complex relationship between places and human behaviour. For this reason, we here studied cities and people as two sides of the same coin.

Our dissertation adopted a data mining approach, where we relied on passively collected data to understand the city. We developed new models that aim at extracting predictive information from geographical data, high-level human perception from images, and tools to analyse cities at scale. Previous literature has often relied on limited data sets that reflect how a group of people experience bits of the city. In this thesis, we have found that web, Open and mobile data offer insights on how most urban dwellers experience entire cities.

In Chapter 2 we verified, for the first time, the four Jane Jacobs's conditions necessary for the promotion of urban life in seven Italian cities. We have done so by *operationalizing* her concepts in new ways: we used mobile phone records to extract a proxy for urban vitality, and web data to extract structural proxies for urban diversity. We showed that the structural and static features of urban diversity explain as much as 77% of the variability of district activity. Thus, as Jacobs envisioned, vitality and diversity are intimately connected.

Then, we moved to human perception. In Chapter 3 we explored the question: "Are safer looking neighbourhoods more lively?" in the context of two Italian cities: Milan and Rome. Our findings suggest that perceived safety modulates the active population in an area, with effects that depend on age and gender. The overall effect of the safety appearance in activity appears to be positive, even after controlling for population, employment density, and distance to the city centre. Yet, the effect does not appear to be universal, and depends on the demographic with the population, with females and people older than 50 appearing to have a stronger preference for the appearance of safety. By randomly occluding images in the deep neural network model, we identified patches of images that contribute positively and negatively to safety perception. We showed that street-facing windows and greenery tend to contribute positively to an appearance of safety, while lit streets contribute negatively to it.

In Chapter 4 we analysed the economic impact of neighbourhood characteristics on housing values. We formulated this analysis as a multi-modal nowcasting problem where the price of unseen neighbourhoods' conditions was predicted. Thus, we collected a large dataset of real estate properties neighbourhood features based on overlapping geographical boundaries called *egohoods*. We found that neighbourhood characteristics seem to drive more than 20% house 's advertised price. Our results also show that the use of this information in the model lowers the prediction error by 60%.

Finally, in Chapter 5 we analysed four cities, very different in cultural, economic, historical and geographical aspects. We presented a Bayesian model able to describe and disentangle the role of the diverse factors in urban crime. We showed that the built environment well describes the urban crime, and it is very correlated to the socio-economical characteristics of neighbourhoods. Despite this, Jane Jacobs' core assumptions have to be considered in context. High diversity and density, small blocks, land-use-mix, and vacuums are not characteristics to be widely appraised. While small blocks are linked to beautiful, low-rise houses of Greenwich Village, they could also be linked to slums in Latin America. However, the study of the place should not be ignored as it is instrumental for those cities where census data is not available. Our results also demonstrate that the combination of diverse sources of information such as urban environment, socio-economical characteristics, and mobility better describe the crime that any of these factors alone. Thus, considering only a subset of these variables can lead to discrepancies and incomplete pictures of the dynamics of the city.

Altogether, our results highlight the neighbourhood effect importance on the description and prediction of emergent urban patterns. We did not, however, provide any causal explanation of the observed effects. For instance, we cannot distinguish between the hypothesis that the built environment impede crime, or that criminals with their actions start some processes that modify the urban physical environment. Nevertheless, our results provide preliminary evidence suggesting a connection between the physical characteristics, the appearance and the levels of human activity that is strong enough to manifest itself at the city scale. Since we rely on new sources of data automatically sensed from the environment, our methods can be readily applied to other cities and complement traditional methods to improve urban spaces.

6.1 THE LIMITS

Despite the advancements presented in this thesis, many problems still hinder the scientific understanding of cities through the lens of multi-modal data.

First, urban data does not come without a cost. Census and geographical data are produced by municipalities and scientific committees at a slow pace. Thus, although we demonstrated that the built environment and socio-economic conditions affect human behaviour and other urban phenomena, our methods rely on data produced to a large extent by humans. Understanding how to describe urban environments from passively and automatically collected data alone might be the key to understand cities at scale, over time.

Second, our analysis does not consider how neighbourhoods change over time. This is mainly due to lack of data. However, this limits the applicability of our methods in real-world scenarios, from predicting and simulating urban outcomes, to inform policymakers and stakeholders in urban planning. Including temporal aspects in our methodology might drive future research and the development of new models.

Finally, alongside the scientific limitations of our work, one of the obstacles to designing impactful research on cities come from the problematic collaboration between scientists and policymakers. It is still indeed a problem for scientists to speak to a different community to rely both on the new streams of data revealing how the city changes, and on the experience of decision makers. Tighter collaboration between these two scientific fields might help understanding cities, and avoid wrong urban choices (e.g. Le Corbusier style failures) that worsen human life in cities.

6.2 IF I HAD TO WRITE A SECOND THESIS (FUTURE DIRECTIONS)

These three years of the Ph.D. program were instrumental for investigating multiple research questions that, inevitably, raised new questions in both urban studies and computer science. Here we briefly describe the topics we are currently addressing, and those we plan to investigate in the future.

As we have seen, the life of a city is deeply influenced by its urban structure and appearance. However, it is still unclear, even for urban planners, how to design and refine a neighbourhood in order to accommodate for higher population, increase safety, or decrease congestion. In ongoing work, we propose a novel Generative Adversarial Network (GAN) architecture that serves to this purpose. Given an aerial image of a neighbourhood and its related attribute (e.g. population of the area), our framework simulates the changes that would happen to the area at a higher (lower) population. This work follows recent literature in computer vision where GANs have been successfully applied for transferring style from an image to another [212] or to learn the mapping between different visual domains [213]. However, existing models usually allow mapping an image from one visual domain to another, but they cannot generate an image where a given attribute satisfies a specific value. Here we plan to create a GAN model that uses both land use constraints and satellite imagery to generate realistic conditioned urban images. We thus separate the image in information (e.g., buildings, parks) and content (e.g. street-network) and generate a new image that modifies the information conditioned to some semi-supervised constrains. Qualitative evaluation will be done by urban experts, while we will quantitatively evaluate generate images in a predictive task were we predict with state of the art models the desired target property from generated images. Not only does this tool help policymakers to anticipate urban changes, but it also suggests to urban planners the urban transformation which would induce the desired change to that specific attribute (e.g., to lower traffic). Thus, our contribution here is going to be multi-disciplinary as we will introduce a novel GAN scheme, and provide a reliable tool for urban planners.

Then, we plan to study and automatically define *what* is a neighbourhood. As surprising as it may seem, although literature frequently uses the word "neighbourhood", we still lack a clear definition of its boundaries and meaning. Previous research often relied on boundaries defined through surveys, administrative areas or census tracts, which delineate homogeneous units with respect to population characteristics and socio-economic conditions. For example, in Chapter 2, Chapter 3 and Chapter 5 we used administrative boundaries that were close to the somewhat unclear definition of Jane Jacobs. Instead, in Chapter 4, we relied on automatically collected *egohoods* composed by a static buffer area surrounding a house. Static definitions, however, do not reflect the urban life. As human perception, social interactions and movements are forever changing, scholars have increasingly relied on automatic methods to define neighbourhoods from social media, location-based social networks [52, 214] and crowd-sourced methods [215]. Yet, we still lack an unambiguous and theoretically grounded definition of neighbourhood. Thus, following what we learned in Chapter 5, can we define neighbourhoods from mobility, social interactions, and the built environment? What does it happen if we consider only one of these dimensions? Are they different neighbourhoods? Do we have activity-related neighbourhoods (e.g. work- and home-neighbourhoods)? Many research fields would benefit from a reliable neighbourhood definition, from sociology and criminology, to urban planning and urban health.

One could easily fill more chapters with research ideas, but, to the delight of the reader, my thesis must end somewhere.

A

APPENDIX TO CHAPTER 4

A.1 DATA SOURCES

Table 15: Data sources.

Topics	Bogotá	Boston
Neighborhoods	IDECA[B1]	Analyze Boston[BO1]
Blocks	DANE[*]	US Census Tiger[BO3]
Census Blocks	IDECA[B1]	US Census Tiger - 2014[BO3]
Boundaries	DANE	Analyze Boston[BO2]
Buildings	IDECA[B1]	Analyze Boston[BO4]
Crime	Local police[*]	Analyze Boston[BO5]
Employment	DANE[*]	ACS 2014 5-years[ACS]
Ethnic mix	DANE[*]	ACS 2014 5-years[ACS]
Land Use	IDECA[B1]	Analyze Boston[BO6]
Mobile phone data	12-01, 2013 to 05-31, 2014[*]	6 weeks in 2010
POIs	Foursquare[F]	Foursquare[F]
Population	DANE[*]	ACS 2014 5-years[ACS]
Poverty	Multipurpose Survey (EM)[*]	ACS 2014 5-years[ACS]
Residential stability	DANE[*]	ACS 2014 5-years[ACS]
Streets	OSM	OSM

[*] Privately shared
[ACS] https://factfinder.census.gov
[C] https://www.census.gov/geo/maps-data/data/cbf/cbf_blkgrp.html
[B1] https://www.ideca.gov.co/es/servicios/mapa-de-referencia/tabla-mapa-referencia?tid_1=All&title=&submit-b=Filtrar
[BO1] https://data.boston.gov/dataset/boston-neighborhoods
[BO2] http://bostonopendata-boston.opendata.arcgis.com/datasets/142500a77e2a4dbeb94a86f7e0b568bc_0
[BO3] ftp://ftp2.census.gov/geo/tiger/TIGER2014/TABBLOCK/
[BO4] https://data.boston.gov/dataset/buildings
[BO5] https://data.boston.gov/dataset/crime-incident-reports-july-2012-august-2015-source-legacy-system
[BO6] https://data.boston.gov/dataset/parcels-2016-data-full
[F] Foursquare API https://developer.foursquare.com/places-api

Table 16: Data sources.

Topics	**Los Angeles**	**Chicago**
Neighborhoods	LAcity[LAT]	Chicago data portal[CH1]
Blocks	US Census Tiger[LA5]	US Census Tiger
Census Blocks	LA Blocks[LA1]	Census Bureau[C]
Boundaries	Chicago data portal[CH4]	
Buildings	Countywide Building Outlines 2014[LA2]	Chicago data portal[CH4]
Crime	LAPD 2014[LA3]	Chicago 2014[CH5]
Employment	ACS 2013 5-years[ACS]	ACS 2014 5-years[ACS]
Ethnic mix	ACS 2013 5-years[ACS]	ACS 2014 5-years[ACS]
Land Use	Assessor Parcels – 2014[LA4]	Cook County Government[CH6]
Mobile phone data	10-15, 2012 to 11-24, 2012[*]	-
POIs	Foursquare[F]	Foursquare[F]
Population	ACS 2013 5-years[ACS]	ACS 2014 5-years[ACS]
Poverty	ACS 2013 5-years[ACS]	ACS 2014 5-years[ACS]
Residential stability	ACS 2013 5-years[ACS]	ACS 2014 5-years[ACS]
Streets	OSM	OSM

[*] Privately shared
[ACS] https://factfinder.census.gov
[C] https://www.census.gov/geo/maps-data/data/cbf/cbf_blkgrp.html
[LA1] http://egis3.lacounty.gov/dataportal/2016/01/26/census_blocks/
[LA2] https://egis3.lacounty.gov/dataportal/2016/11/03/countywide-building-outlines-2014-update-public-domain-release/
[LA3] https://data.lacity.org/A-Safe-City/Crimes-2012-2015/s9rj-h3s6
[LA4] https://egis3.lacounty.gov/dataportal/2015/03/10/assessor-parcel/
[LA5] https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Blocks
[LAT] https://data.lacity.org/A-Well-Run-City/Neighborhoods/ykhe-zspy/data
[CH1] https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
[CH4] https://data.cityofchicago.org/Buildings/Building-Footprints-deprecated-August-2015-/qv97-3bvb
[CH5] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
[CH6] https://datacatalog.cookcountyil.gov/GIS-Maps/ccgisdata-Parcel-2014/2m9h-cq6j
[F] Foursquare API https://developer.foursquare.com/places-api

A.2 TEST FOR OVERDISPERSION

We test whether crime has some extra-Poisson variability that has to be accounted for. The Potthoff-Whittinghill index of dispersion test [216] confirms the hypothesis of no overdispersion is rejected. It is defined as:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2/\bar{y} \tag{41}$$

that is approximately a chi-square distribution with $k-1$ degrees of freedom. We also apply the Lagrange multiplier test, defined as:

$$\frac{(\sum_{i=1}^{n}\mu_i^2 - n\bar{y})^2}{2\sum_{i=1}^{n}\mu_i^2} \tag{42}$$

With one degree of freedom, the test appears to be significant – the hypothesis of no overdispersion is again rejected.

A.3 CRIME MAPPING

Table 17 show the map of each crime category created by the police of Bogota and the UCR category described by the FBI in the US.

A.4 ADDITIONAL MAPS

In this section we show some additional results that might be useful to qualitatively interpret the predictive model from the maps. Thus, we show Bogotá in Figure 17, Boston in Figure 18, Los Angeles in Figure 19, and Chicago in Figure 20.

Modality	**Number of crimes**	**UCR part**	**Subsection**
Factor De Oportunidad	13300	1	Larceny-theft
Atraco	10957	1	Robbery
Raponazo	2958	1	Larceny-theft
Cosquilleo	2486	1	Larceny-theft
Mechero	2247	1	Arson
Llaves Maestras	1502	1	Larceny-theft
Homicidios	1335	1	Criminal homicide
Violación De Cerraduras	1302	1	Larceny-theft
Ventosa	994	1	Larceny-theft
Rompimiento Vidrio	504	1	Larceny-theft
Escopolaminado	458	1	Robbery
Supl. Dijin-Sijin	168	1	Larceny-theft
Rompimiento Pared	49	1	Larceny-theft
Violacion Caja Fuerte	10	1	Larceny-theft
Auto Robo	2	1	Motor vehicle theft
Palancas	1	1	Larceny-theft

Table 17: Mapping between the crime categories in Bogota and the Uniform Crime Report (UCR) of the US. Here we show just the UCR part 1 categories.

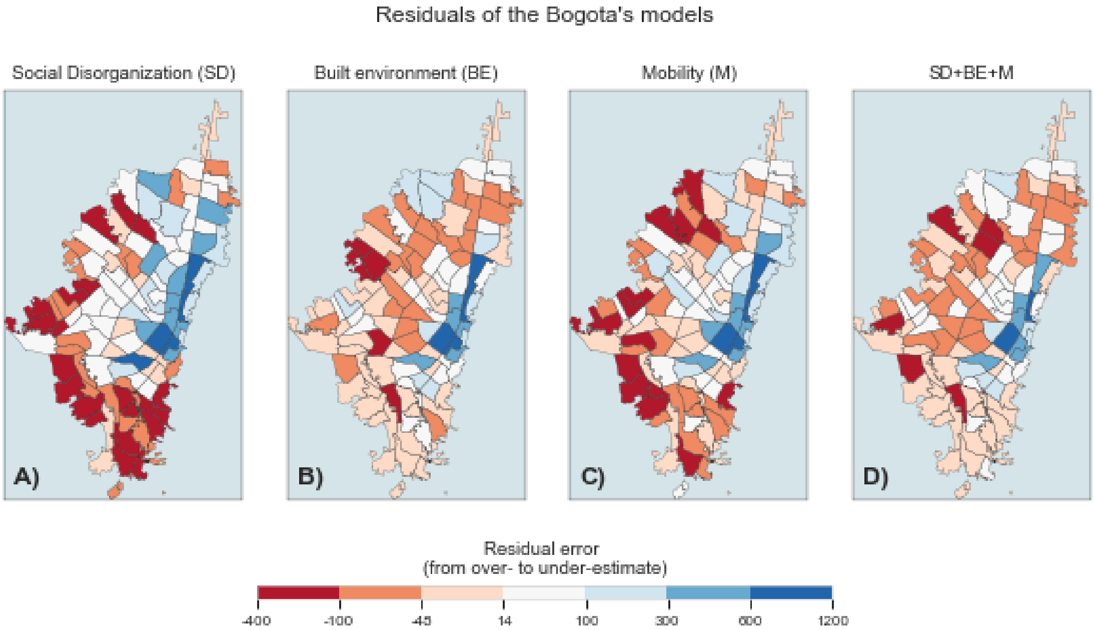


Figure 17: The residuals in Bogotá of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; D) The full model.

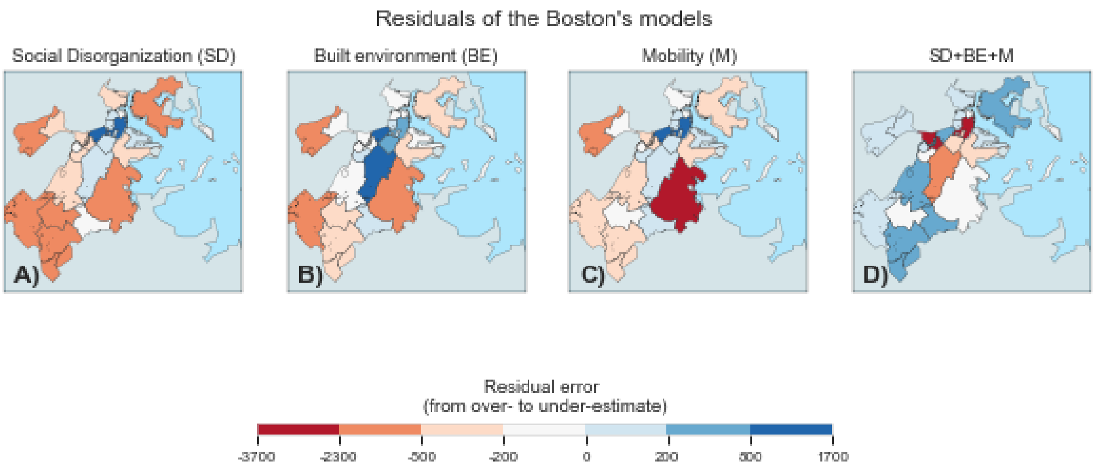


Figure 18: The residuals in Boston of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; D) The full model.

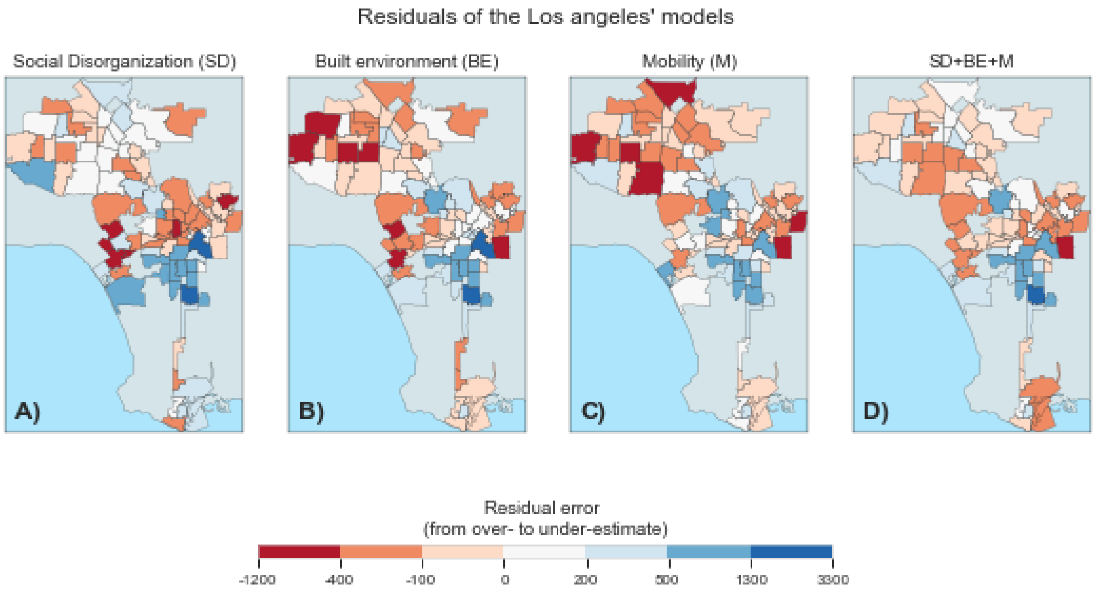


Figure 19: The residuals in Los Angeles of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; D) The full model.

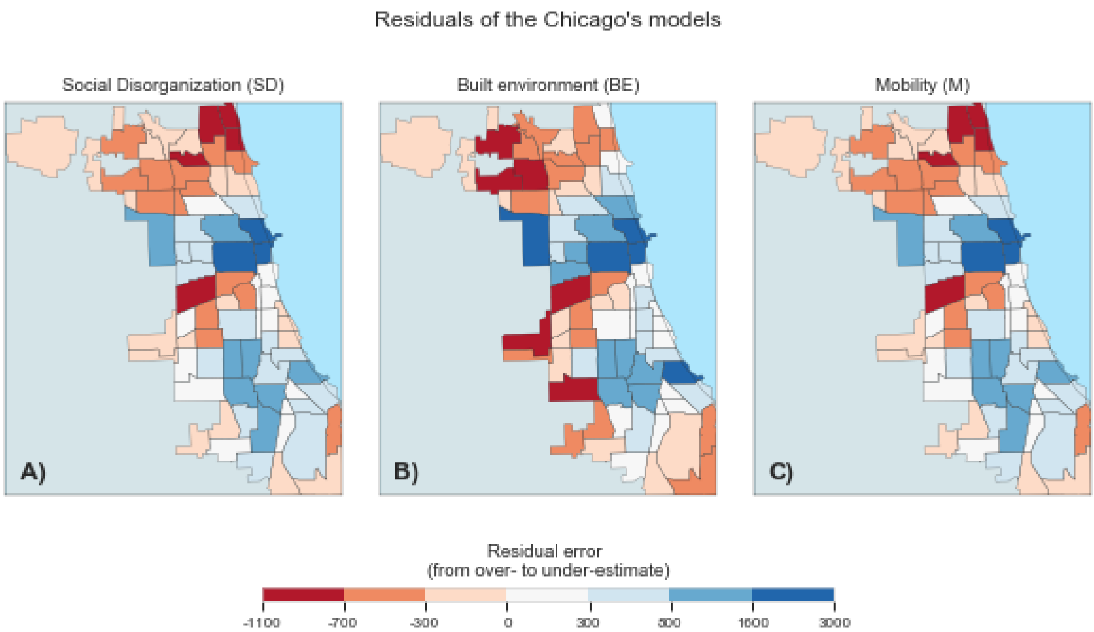


Figure 20: The residuals in Chicago of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; D) The full model.

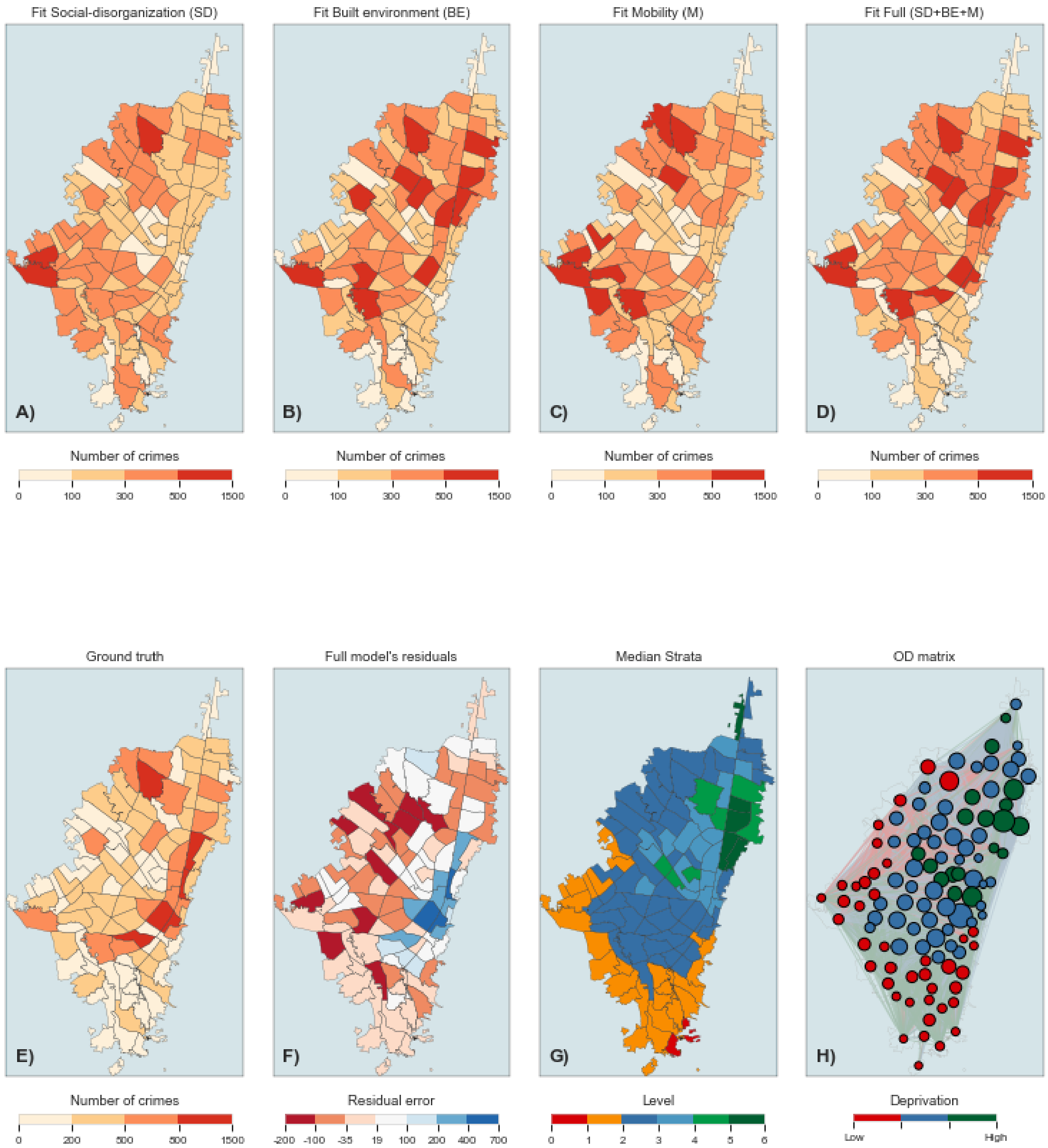


Figure 21: The fit in Bogotá of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; and D) the Full model. E) Shows the ground truth data of crime counts. F) The residual error of the Full model. G) the social disorganization index distribution; H) The mobility of people. Origin nodes are colored by the social disorganization index grouped in three bins, while the sizes represent the total number of commuting trips.

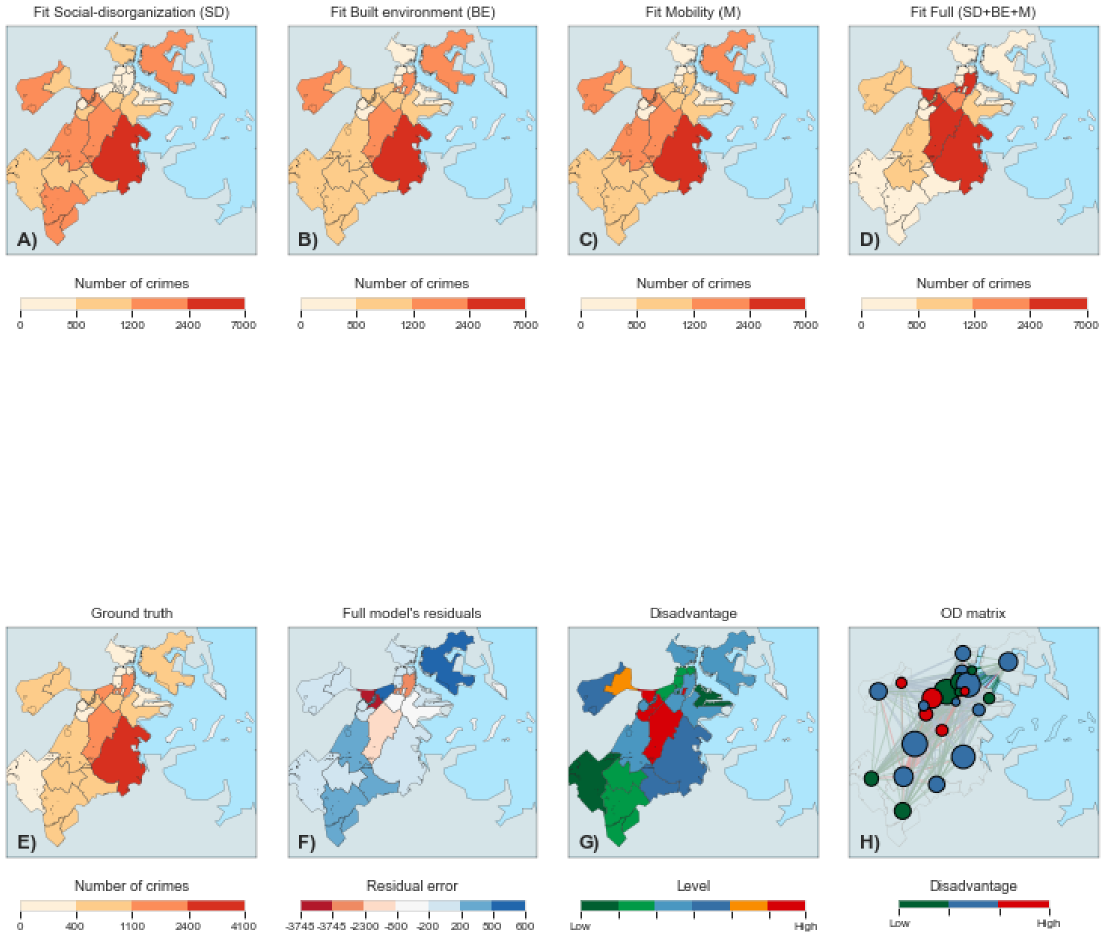


Figure 22: The fit in Boston of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; and D) the Full model. E) Shows the ground truth data of crime counts. F) The residual error of the Full model. G) the social disorganization index distribution; H) The mobility of people. Origin nodes are colored by the social disorganization index grouped in three bins, while the sizes represent the total number of commuting trips.

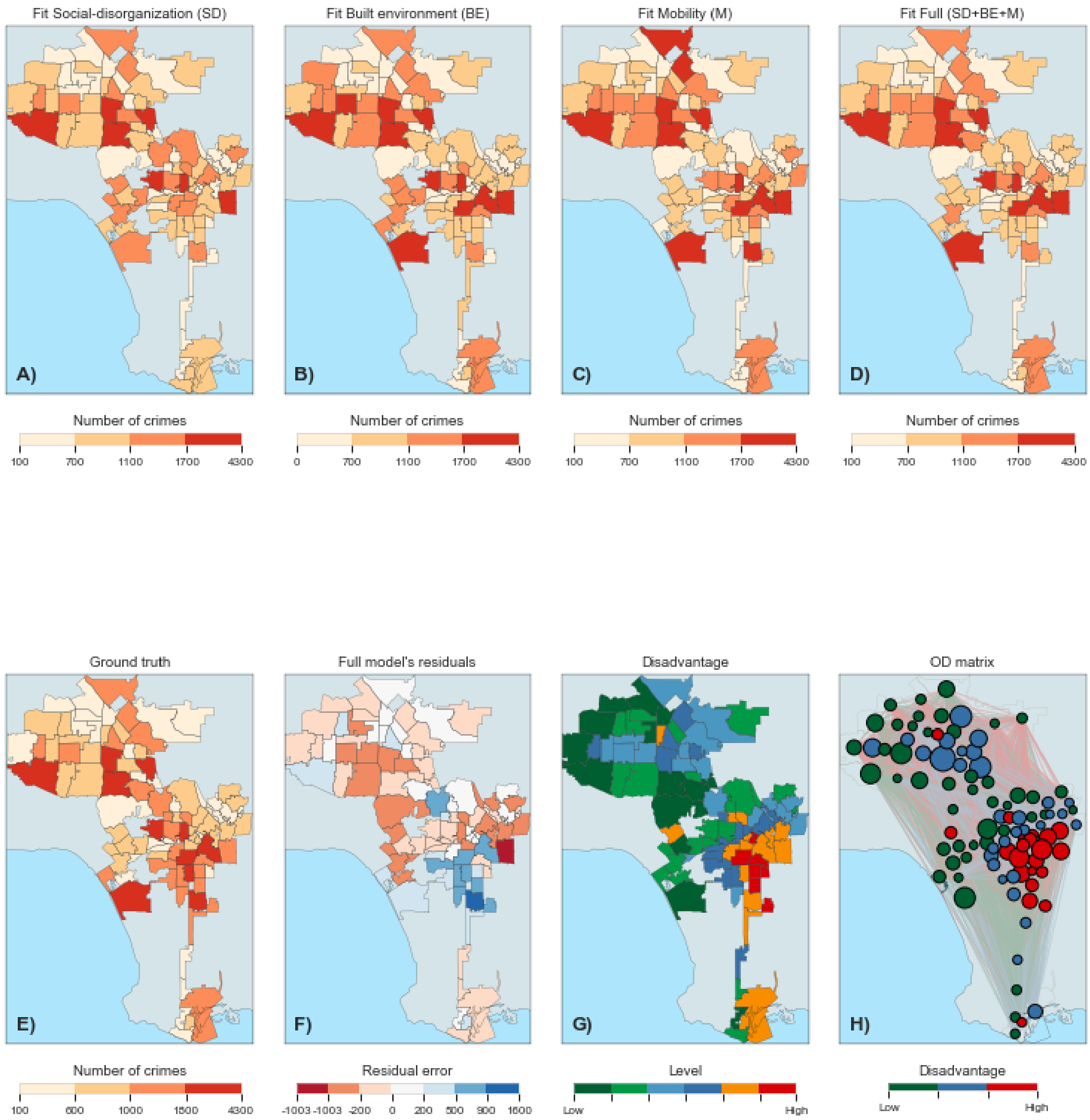


Figure 23: The fit in Los Angeles of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the Mobility (M) model; and D) the Full model. E) Shows the ground truth data of crime counts. F) The residual error of the Full model. G) the social disorganization index distribution; H) The mobility of people. Origin nodes are colored by the social disorganization index grouped in three bins, while the sizes represent the total number of commuting trips.

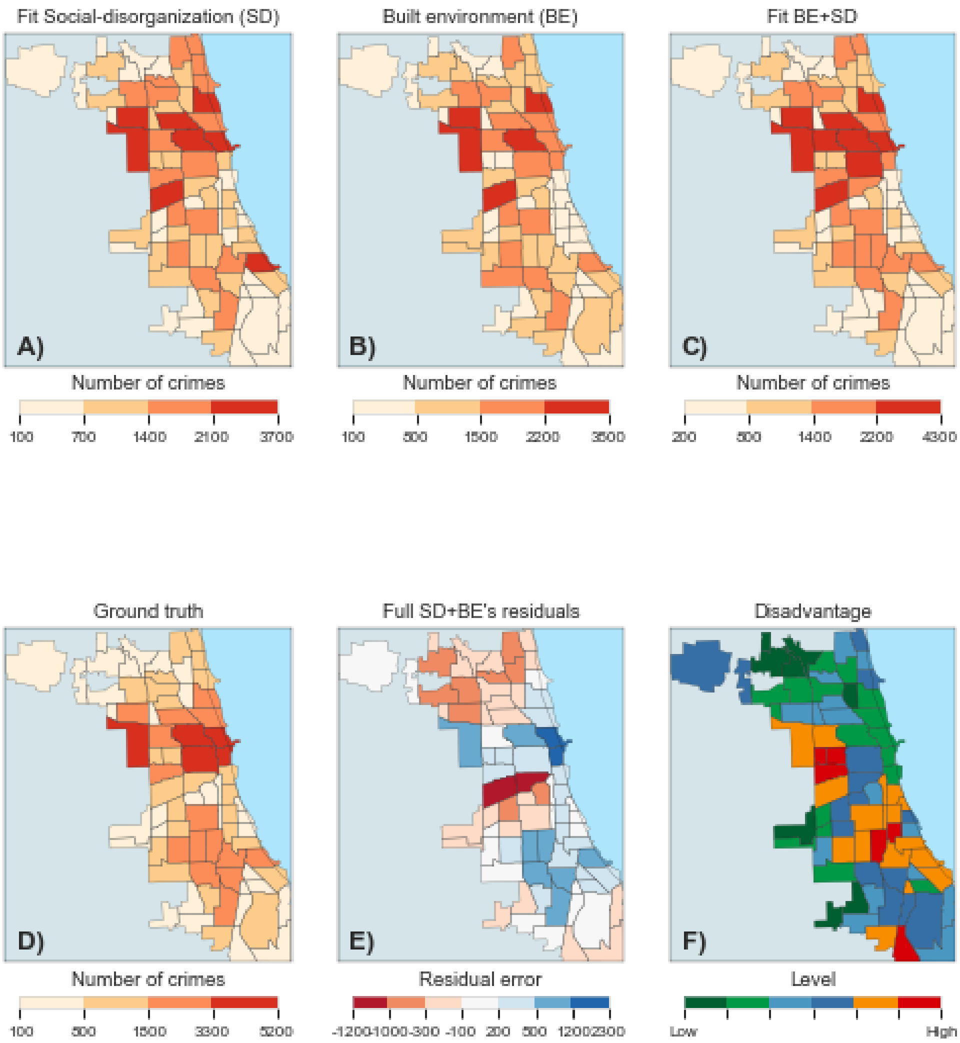


Figure 24: The fit in Chicago of: A) the Social-Disorganization (SD) model; B) the Built environment (BE) model; C) the SD+BE model. D) The ground truth data of crime counts. E) The residual error of the SD+BE model. F) the disorganization index distribution.

A.5 RESULTS BROKEN DOWN BY CRIME TYPE

Results might be influenced by the type of crime. For example, the presence of small blocks could have an higher effect on property crimes than violent ones. For this reason, we here compare the β coefficients of the models with all the crimes (Total crime), the ones using only property crimes, and finally the models using violent crimes only.

We show Bogotá models in Table 18, Table 21, Table 25, and Table 29. Boston models are then listed in Table 19, Table 22, Table 26, and Table 30, while Los Angeles in Table 20, Table 23, Table 27, and Table 31. Finally, Chicago is shown in Table 24, and Table 28.

Table 18: Results of the crime count full model in Bogotá. HDI is the Highest Density Interval (2.5%, 97.5%).

Bogotá, Full model	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	0.05	(-0.04, 0.13)	0.04	(-0.05, 0.13)	0.10	(0.01, 0.19)
Small blocks	-0.02	(-0.21, 0.17)	0.08	(-0.12, 0.29)	-0.22	(-0.42, -0.01)
Building diversity	0.13	(-0.01, 0.26)	0.18	(0.03, 0.34)	0.08	(-0.07, 0.22)
Population density	-0.67	(-0.88, -0.46)	-0.78	(-1.03, -0.53)	-0.47	(-0.71, -0.24)
Building density	0.26	(0.11, 0.40)	0.30	(0.14, 0.46)	0.17	(0.02, 0.34)
Walkscore	0.20	(0.08, 0.31)	0.28	(0.15, 0.40)	0.12	(0.01, 0.24)
Social-disorganization						
Disadvantage index	0.07	(-0.04, 0.17)	0.01	(-0.11, 0.14)	0.09	(-0.01, 0.21)
Ethnic diversity	0.03	(-0.06, 0.12)	0.04	(-0.05, 0.14)	0.03	(-0.07, 0.13)
Residential stability	0.09	(-0.00, 0.19)	0.11	(0.00, 0.21)	0.04	(-0.07, 0.14)
Mobility						
Ambient population	0.33	(0.16, 0.53)	0.27	(0.11, 0.44)	0.35	(0.16, 0.55)
Mobility autocorr.	0.57	(0.13, 1.00)	0.48	(0.06, 0.91)	0.54	(0.13, 1.00)
Others						
Population.	0.76	(0.48, 1.02)	0.90	(0.61, 1.21)	0.55	(0.26, 0.83)
Spatial autocorr.	0.98	(0.96, 1.00)	0.98	(0.96, 1.00)	0.98	(0.94, 1.00)

Table 19: Results of the crime count full model in Boston.

Boston, Full model	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	-0.08	(-0.34, 0.17)	-0.05	(-0.30, 0.20)	-0.27	(-0.64, 0.10)
Small blocks	0.44	(-0.12, 0.94)	0.43	(-0.07, 0.95)	0.26	(-0.43, 0.94)
Building diversity	0.18	(-0.13, 0.50)	0.19	(-0.11, 0.50)	0.04	(-0.40, 0.46)
Population density	-0.34	(-0.94, 0.29)	-0.35	(-0.96, 0.26)	-0.18	(-0.96, 0.61)
Building density	0.34	(-0.55, 1.25)	0.34	(-0.58, 1.20)	0.09	(-1.02, 1.23)
Walkscore	0.61	(0.17, 1.03)	0.63	(0.21, 1.03)	0.24	(-0.35, 0.82)
Social-disorganization						
Disadvantage index	1.02	(0.41, 1.61)	0.93	(0.36, 1.53)	1.16	(0.37, 1.99)
Ethnic diversity	-0.12	(-0.32, 0.08)	-0.13	(-0.33, 0.07)	0.08	(-0.21, 0.36)
Residential stability	-0.68	(-1.25, -0.08)	-0.68	(-1.24, -0.12)	-0.42	(-1.23, 0.39)
Mobility						
Ambient population	0.61	(-0.26, 1.47)	0.65	(-0.20, 1.51)	0.71	(-0.33, 1.78)
Mobility autocorr.	0.45	(0.06, 0.85)	0.45	(0.07, 0.85)	0.46	(0.07, 0.86)
Others						
Population.	0.65	(-0.19, 1.54)	0.63	(-0.24, 1.45)	0.54	(-0.52, 1.58)
Spatial autocorr.	0.97	(0.93, 1.00)	0.97	(0.93, 1.00)	0.91	(0.81, 0.99)

Table 20: Results of the crime count full model in Los Angeles.

Los Angeles, Full model	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	-0.00	(-0.06, 0.06)	-0.01	(-0.07, 0.05)	0.02	(-0.07, 0.12)
Small blocks	-0.07	(-0.19, 0.06)	-0.09	(-0.21, 0.04)	-0.11	(-0.30, 0.09)
Building diversity	0.03	(-0.04, 0.09)	0.04	(-0.03, 0.10)	-0.05	(-0.15, 0.05)
Population density	-0.29	(-0.46, -0.12)	-0.35	(-0.51, -0.18)	-0.20	(-0.47, 0.07)
Building density	0.14	(0.02, 0.27)	0.16	(0.04, 0.28)	0.06	(-0.14, 0.26)
Walkscore	0.03	(-0.07, 0.13)	0.01	(-0.09, 0.11)	0.14	(-0.03, 0.31)
Social-disorganization						
Disadvantage index	-0.01	(-0.13, 0.10)	-0.07	(-0.18, 0.04)	0.36	(0.18, 0.55)
Ethnic diversity	-0.05	(-0.12, 0.01)	-0.04	(-0.11, 0.02)	-0.08	(-0.18, 0.02)
Residential stability	-0.18	(-0.31, -0.05)	-0.22	(-0.34, -0.09)	-0.09	(-0.29, 0.12)
Mobility						
Ambient population	0.37	(0.11, 0.61)	0.31	(0.07, 0.55)	0.41	(0.02, 0.82)
Mobility autocorr.	0.46	(0.06, 0.86)	0.46	(0.07, 0.87)	0.45	(0.06, 0.84)
Others						
Population.	0.18	(-0.08, 0.43)	0.24	(0.00, 0.49)	0.10	(-0.29, 0.51)
Spatial autocorr.	0.99	(0.98, 1.00)	0.99	(0.98, 1.00)	0.98	(0.95, 1.00)

Table 21: Results of the crime count Social-Disorganization (SD) model in Bogotá.

Bogotá	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Social-disorganization						
Disadvantage index	0.08	(-0.02, 0.17)	0.02	(-0.10, 0.13)	0.13	(0.04, 0.22)
Ethnic diversity	0.10	(-0.03, 0.23)	0.12	(-0.04, 0.25)	0.08	(-0.05, 0.20)
Residential stability	0.14	(-0.01, 0.28)	0.19	(0.04, 0.35)	0.06	(-0.08, 0.20)
Others						
Population.	0.72	(0.58, 0.85)	0.74	(0.60, 0.89)	0.75	(0.61, 0.87)
Spatial autocorr.	0.90	(0.82, 0.97)	0.89	(0.79, 0.97)	0.89	(0.81, 0.98)

Table 22: Results of the crime count Social-Disorganization (SD) model in Boston.

Boston	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Social-disorganization						
Disadvantage index	0.22	(-0.29, 0.74)	0.31	(-0.42, 1.05)	0.77	(0.02, 1.58)
Ethnic diversity	0.06	(-0.19, 0.31)	0.05	(-0.21, 0.30)	0.19	(-0.09, 0.47)
Residential stability	-0.16	(-0.77, 0.43)	-0.15	(-0.79, 0.47)	-0.26	(-0.94, 0.41)
Others						
Population.	1.18	(0.79, 1.58)	1.40	(0.94, 1.87)	1.21	(0.72, 1.71)
Spatial autocorr.	0.89	(0.79, 0.99)	0.89	(0.79, 0.98)	0.82	(0.68, 0.96)

Table 23: Results of the crime count Social-Disorganization (SD) model in Los Angeles.

Los Angeles	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Social-disorganization						
Disadvantage index	-0.23	(-0.38, -0.07)	-0.31	(-0.46, -0.14)	0.17	(-0.00, 0.34)
Ethnic diversity	0.02	(-0.07, 0.10)	0.02	(-0.06, 0.11)	-0.01	(-0.10, 0.09)
Residential stability	-0.25	(-0.40, -0.10)	-0.27	(-0.42, -0.12)	-0.26	(-0.44, -0.10)
Others						
Population.	0.51	(0.44, 0.58)	0.52	(0.45, 0.59)	0.45	(0.37, 0.54)
Spatial autocorr.	0.97	(0.94, 1.00)	0.97	(0.93, 0.99)	0.95	(0.90, 0.99)

Table 24: Results of the crime count Social-Disorganization (SD) model in Chicago.

Chicago	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Social-disorganization						
Disadvantage index	0.11	(-0.06, 0.29)	0.02	(-0.15, 0.20)	0.48	(0.27, 0.69)
Ethnic diversity	-0.03	(-0.12, 0.05)	-0.02	(-0.12, 0.06)	-0.06	(-0.16, 0.04)
Residential stability	-0.12	(-0.29, 0.06)	-0.14	(-0.32, 0.04)	-0.02	(-0.24, 0.20)
Others						
Population.	0.69	(0.61, 0.77)	0.67	(0.59, 0.76)	0.79	(0.69, 0.89)
Spatial autocorr.	0.96	(0.93, 1.00)	0.96	(0.93, 0.99)	0.94	(0.89, 0.99)

Table 25: Results of the crime count Built environment (BE) model in Bogotá.

Bogotá	Total crime		Property crime		Violent crime	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	0.04	(-0.05, 0.13)	0.03	(-0.06, 0.13)	0.10	(-0.01, 0.19)
Small blocks	0.10	(-0.08, 0.27)	0.22	(0.02, 0.40)	-0.16	(-0.35, 0.03)
Building diversity	0.02	(-0.09, 0.13)	0.18	(0.05, 0.32)	-0.06	(-0.18, 0.05)
Population density	-0.99	(-1.18, -0.81)	-1.08	(-1.28, -0.89)	-0.87	(-1.07, -0.67)
Building density	0.38	(0.24, 0.51)	0.47	(0.32, 0.62)	0.26	(0.11, 0.40)
Walkscore	0.19	(0.07, 0.31)	0.22	(0.09, 0.35)	0.13	(0.00, 0.26)
Others						
Population.	1.25	(1.05, 1.45)	1.35	(1.13, 1.57)	1.10	(0.89, 1.31)
Spatial autocorr.	0.94	(0.89, 0.99)	0.94	(0.88, 0.99)	0.93	(0.87, 0.99)

Table 26: Results of the crime count Built environment (BE) model in Boston.

Boston	Total crime		Property crime		Violent crime	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	-0.18	(-0.47, 0.11)	-0.17	(-0.45, 0.10)	-0.35	(-0.78, 0.11)
Small blocks	-0.02	(-0.41, 0.39)	0.08	(-0.41, 0.56)	0.16	(-0.58, 0.90)
Building diversity	0.16	(-0.14, 0.45)	0.15	(-0.11, 0.43)	0.17	(-0.26, 0.63)
Population density	-0.56	(-1.17, 0.08)	-0.70	(-1.37, -0.02)	-0.27	(-1.11, 0.62)
Building density	0.46	(-0.20, 1.06)	0.88	(-0.14, 1.88)	0.29	(-0.95, 1.55)
Walkscore	0.27	(-0.20, 0.73)	0.31	(-0.08, 0.73)	0.15	(-0.47, 0.79)
Others						
Population.	1.28	(0.96, 1.58)	1.52	(1.16, 1.86)	1.34	(0.78, 1.90)
Spatial autocorr.	0.91	(0.83, 0.99)	0.92	(0.84, 0.99)	0.81	(0.66, 0.96)

Table 27: Results of the crime count Built environment (BE) model in Los Angeles.

Los Angeles	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	0.02	(-0.04, 0.08)	0.01	(-0.05, 0.07)	0.06	(-0.04, 0.17)
Small blocks	-0.02	(-0.13, 0.10)	-0.04	(-0.16, 0.07)	0.07	(-0.12, 0.26)
Building diversity	0.05	(-0.01, 0.11)	0.06	(0.00, 0.13)	-0.02	(-0.13, 0.08)
Population density	-0.42	(-0.52, -0.32)	-0.46	(-0.56, -0.36)	-0.18	(-0.35, -0.01)
Building density	0.27	(0.17, 0.37)	0.28	(0.18, 0.38)	0.21	(0.04, 0.40)
Walkscore	0.11	(0.02, 0.22)	0.09	(-0.01, 0.19)	0.27	(0.10, 0.45)
Others						
Population.	0.55	(0.50, 0.61)	0.56	(0.51, 0.62)	0.50	(0.40, 0.60)
Spatial autocorr.	0.98	(0.96, 1.00)	0.98	(0.95, 1.00)	0.95	(0.90, 0.99)

Table 28: Results of the crime count Built environment (BE) model in Chicago.

Chicago	**Total crime**		**Property crime**		**Violent crime**	
	β	HDI	β	HDI	β	HDI
Built environment						
Land-use-mix	0.12	(0.02, 0.22)	0.13	(0.03, 0.22)	0.08	(-0.06, 0.22)
Small blocks	-0.00	(-0.14, 0.14)	0.00	(-0.13, 0.13)	-0.09	(-0.28, 0.11)
Building diversity	-0.09	(-0.26, 0.08)	-0.03	(-0.19, 0.13)	-0.36	(-0.61, -0.13)
Population density	-0.15	(-0.36, 0.06)	-0.25	(-0.44, -0.05)	0.22	(-0.08, 0.51)
Building density	0.01	(-0.18, 0.20)	0.08	(-0.10, 0.26)	-0.25	(-0.51, 0.03)
Walkscore	-0.07	(-0.26, 0.10)	-0.03	(-0.20, 0.14)	-0.23	(-0.50, 0.02)
Others						
Population.	0.76	(0.66, 0.86)	0.76	(0.66, 0.84)	0.82	(0.69, 0.96)
Spatial autocorr.	0.96	(0.93, 1.00)	0.97	(0.93, 1.00)	0.94	(0.88, 0.99)

Table 29: Results of the crime count Mobility (M) model in Bogotá.

Bogotá	**Property crime**		**Violent crime**		**Total crime**	
	β	HDI	β	HDI	β	HDI
Mobility						
Ambient population	0.62	(0.45, 0.87)	0.61	(0.43, 0.81)	0.58	(0.44, 0.72)
Mobility autocorr.	0.52	(0.12, 1.00)	0.50	(0.12, 1.00)	0.50	(0.12, 1.00)
Others						
Population.	0.21	(0.03, 0.36)	0.21	(0.03, 0.38)	0.29	(0.16, 0.42)
Spatial autocorr.	0.96	(0.92, 1.00)	0.96	(0.92, 1.00)	0.96	(0.91, 1.00)

Table 30: Results of the crime count Mobility (M) model in Boston.

Boston	**Property crime**		**Violent crime**		**Total crime**	
	β	HDI	β	HDI	β	HDI
Mobility						
Ambient population	1.10	(0.34, 1.83)	1.17	(0.45, 1.88)	0.61	(-0.39, 1.61)
Mobility autocorr.	0.47	(0.07, 0.88)	0.47	(0.06, 0.88)	0.48	(0.08, 0.88)
Others						
Population.	0.49	(-0.16, 1.20)	0.43	(-0.23, 1.05)	0.86	(-0.04, 1.73)
Spatial autocorr.	0.97	(0.92, 1.00)	0.97	(0.93, 1.00)	0.93	(0.83, 1.00)

Table 31: Results of the crime count Mobility (M) model in Los Angeles.

Los Angeles	**Property crime**		**Violent crime**		**Total crime**	
	β	HDI	β	HDI	β	HDI
Mobility						
Ambient population	0.72	(0.52, 0.91)	0.74	(0.54, 0.94)	0.46	(0.12, 0.79)
Mobility autocorr.	0.46	(0.06, 0.86)	0.46	(0.05, 0.85)	0.45	(0.06, 0.85)
Others						
Population.	-0.14	(-0.34, 0.05)	-0.17	(-0.37, 0.03)	0.14	(-0.20, 0.47)
Spatial autocorr.	0.99	(0.98, 1.00)	0.99	(0.99, 1.00)	0.97	(0.94, 1.00)

BIBLIOGRAPHY

[1] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. In *WWW '16*, 2016.

[2] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *MM '16*. ACM, 2016.

[3] Marco De Nadai and Bruno Lepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment. In *DSAA '18*, 2018.

[4] Marco De Nadai, Yanyan Xu, Emmanuel Letouzé, Marta C. González, and Bruno Lepri. Socio-economic, built environment, and mobility conditions associated with crime: A study of multiple cities. *Under submission to Nature Human Behaviour*, 2019.

[5] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific Data*, 2:150055 EP, 10 2015.

[6] Simone Centellegher, Marco De Nadai, Michele Caraviello, Chiara Leonardi, Michele Vescovi, Yusi Ramadian, Nuria Oliver, Fabio Pianesi, Alex Pentland, Fabrizio Antonelli, and Bruno Lepri. The mobile territorial lab: A multilayered and dynamic view on parents' daily lives. *EPJ Data Science*, 5(3), 2016.

[7] Marco Mamei, Francesca Pancotto, Marco De Nadai, Bruno Lepri, Michele Vescovi, Franco Zambonelli, and Alex Pentland. Is social capital associated with synchronization in human communication? an analysis of italian call records and measures of civic engagement. *EPJ Data Science*, 7(1):25, Jul 2018.

[8] Marco De Nadai, Angelo Cardoso, Antonio Lima, Bruno Lepri, and Nuria Oliver. Apps, places and people: strategies, limitations and trade-offs in

the physical and digital worlds. *Under submission to Scientific Reports*, 2019.

[9] Emanuele Strano, Filippo Simini, Marco De Nadai, Thomas Esch, and Mattia Marconcini. Precise mapping, density and spatial structure of all human settlements on earth. *Under submission to Nature Communications*, 2019.

[10] Jane Jacobs. *The death and life of American cities*. Random House, 1961.

[11] George Gmelch, Robert V. Kemper, and Walter P. Zenner. *Urban Life: Readings in the Anthropology of the City*. Waveland Pr Inc, 2009.

[12] Lewis Mumford. *The City in History: Its Origins, Its Transformations, and Its Prospects*. Mariner Books, 1968.

[13] Giles Duranton and Diego Puga. Micro-foundations of urban agglomeration economies. 2003.

[14] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

[15] Masahisa Fujita. A monopolistic competition model of spatial agglomeration: Differentiated product approach. *Regional science and urban economics*, 18(1):87–124, 1988.

[16] E.L. Glaeser. *Triumph of the city: How our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin Press HC, 2011.

[17] Peter Taylor, Ben Derudder, Pieter Saey, and Frank Witlox. *Cities in Globalization: Practices, policies and theories*. Routledge, 2006.

[18] Marc Barthelemy. *The structure and dynamics of cities*. Cambridge University Press, 2016.

[19] Michael Batty. *The new science of cities*. Mit Press, 2013.

[20] Luís M. A. Bettencourt. The origins of scaling in cities. *Science*, 340(6139):1438–1441, 2013.

[21] Luis Bettencourt and Geoffrey West. A unified theory of urban living. *Nature*, 467(7318):912–913, 2010.

[22] John Iceland, Daniel Weinberg, and Lauren Hughes. The residential segregation of detailed hispanic and asian groups in the united states: 1980-2010. *Demographic Research*, 31:593, 2014.

[23] John R Logan, Brian J Stults, and Reynolds Farley. Segregation of minorities in the metropolis: Two decades of change. *Demography*, 41(1):1–22, 2004.

[24] Tim Cassiers and Christian Kesteloot. Socio-spatial inequalities and social cohesion in european cities. *Urban Studies*, 49(9):1909–1924, 2012.

[25] William Goldsmith and Edward Blakely. *Separate societies: Poverty and inequality in US cities*. Temple University Press, 2010.

[26] Sako Musterd, Szymon Marcińczak, Maarten Van Ham, and Tiit Tammaru. Socioeconomic segregation in european capital cities. increasing separation between poor and rich. *Urban Geography*, 38(7):1062–1083, 2017.

[27] YongJei Lee, John E. Eck, SooHyun O, and Natalie N. Martinez. How concentrated is crime at places? a systematic review from 1970 to 2015. *Crime Science*, 6(1):6, May 2017.

[28] Robert J Sampson. *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press, 2012.

[29] Ju¨ rgen Friedrichs, George Galster, and Sako Musterd. Neighbourhood effects on social opportunities: the european and american research and policy context. *Housing studies*, 18(6):797–806, 2003.

[30] Robert J Sampson, Stephen W Raudenbush, and Felton Earls. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328):918–924, 1997.

[31] Robert J Sampson and Stephen W Raudenbush. Systematic social observation of public spaces: A new look at disorder in urban Neighborhoods 1. *American journal of sociology*, 105(3):603–651, 1999.

[32] Robert J Sampson and Stephen W Raudenbush. *Disorder in urban neighborhoods: Does it lead to crime*. US Department of Justice, Office of Justice Programs, National Institute of Justice, 2001.

[33] Robert J Sampson, Jeffrey D Morenoff, and Thomas Gannon-Rowley. Assessing "neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, pages 443–478, 2002.

[34] Robert J Sampson and Stephen W Raudenbush. Seeing disorder: Neighborhood stigma and the social construction of “broken windows”. *Social psychology quarterly*, 67(4):319–342, 2004.

[35] Robert J. Sampson and W. Byron Groves. Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology*, 94(4):774–802, 1989.

[36] Ricardo Araya, Frank Dunstan, Rebecca Playle, Hollie Thomas, Stephen Palmer, and Glyn Lewis. Perceptions of social capital and the built environment and mental health. *Social science & medicine*, 62(12):3072–3083, 2006.

[37] Kevin M Leyden. Social capital and the built environment: the importance of walkable neighborhoods. *American journal of public health*, 93(9):1546–1551, 2003.

[38] Marta Elliott. The stress process in neighborhood context. *Health & place*, 6(4):287–299, 2000.

[39] William Julius Wilson. *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press, 2012.

[40] Vincent Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(10):1–55, 2015.

[41] M. Gonzalez, C. Hidalgo, and L. Barabasi. Understanding individual mobility patterns. *Nature*, 453(7196):779–782, 2008.

[42] Tim Schwanen, Mei-Po Kwan, and Fang Ren. How fixed is fixed? gendered rigidity of space–time constraints and geographies of everyday activities. *Geoforum*, 39(6):2109–2121, 2008.

[43] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[44] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6:8166, 2015.

[45] C. Song, Z. Qu, N. Blumm, and A Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[46] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491, 2018.

[47] Laura Alessandretti, Sune Lehmann, and Andrea Baronchelli. Understanding the interplay between social and spatial behaviour. *EPJ Data Sci.*, 7(1):36, 2018.

[48] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider, and Marta C González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.

[49] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4, 2014.

[50] Fabio Miranda, Harish Doraiswamy, Marcos Lage, Kai Zhao, Bruno Gonçalves, Luc Wilson, Mondrian Hsieh, and Cláudio T Silva. Urban pulse: Capturing the rhythm of cities. *IEEE transactions on visualization and computer graphics*, 23(1):791–800, 2017.

[51] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[52] Amy X Zhang, Anastasios Noulas, Salvatore Scellato, and Cecilia Mascolo. Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *Social Computing (SocialCom), 2013 International Conference on*, pages 69–74. IEEE, 2013.

[53] Chloë Brown, Anastasios Noulas, Cecilia Mascolo, and Vincent Blondel. A place-focused model for social networks in cities. In *Social Computing (SocialCom), 2013 International Conference on*, pages 75–80. IEEE, 2013.

[54] Anastasios Noulas, Blake Shaw, Renaud Lambiotte, and Cecilia Mascolo. Topological properties and temporal dynamics of place networks in urban environments. In *Proceedings of the 24th International Conference on World Wide Web*, pages 431–441. ACM, 2015.

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

[56] Petko Georgiev, Anastasios Noulas, and Cecilia Mascolo. Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. 2014.

[57] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *ACM conference on Human factors in computing systems*, pages 511–520, 2014.

[58] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690, 2017.

[59] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. In *ICMI*, pages 427–434. ACM, 2014.

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

[61] Christopher Barrington-Leigh and Adam Millard-Ball. The world's user-generated road map is more than 80% complete. *PloS one*, 12(8):e0180698, 2017.

[62] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM conference on Computer supported cooperative work & social computing*, pages 945–955, 2014.

[63] Martin Traunmueller, Paul Marshall, and Licia Capra. Crowdsourcing Safety Perceptions of People: Opportunities and Limitations. In *Social Informatics*, pages 120–135. 2015.

[64] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo. Computer vision uncovers predictors of physical urban change. *PNAS*, 114(29):7571–7576, 2017.

[65] Philip Salesses, Katja Schechtner, and César A Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013.

[66] César A Hidalgo and Elisa E Castañer. Do we need another coffee house? The amenity space and the evolution of neighborhoods. *arXiv preprint arXiv:1509.02868*, 2015.

[67] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 254–264. ACM, 2015.

[68] Stanislav Sobolevsky, Iva Bojic, Alexander Belyi, Izabela Sitko, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. In *big data (BigData congress), 2015 IEEE international congress on*, pages 600–607. IEEE, 2015.

[69] Daniele Quercia. Playful cities: Crowdsourcing urban happiness with web games. *Built Environment*, 42(3):430–440, 2016.

[70] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. Smelly Maps: The Digital Life of Urban Smellscapes. In *AAAI*, 2015.

[71] Yu Zheng, Licia Capra, Ouri Wolfson, and Yang Hai. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5, 2014.

[72] Hyungun Sung, Sugie Lee, and SangHyun Cheon. Operationalizing jane jacobs's urban design theory: Empirical verification from the great city of seoul, korea. *Journal of Planning Education and Research*, 35(2):117–130, 2015.

[73] Jane M. Jacobs and Lees Loretta. Defensible space on the move: revisiting the urban geography of alice coleman. In *International Journal of Urban and Regional Research 37.5 (2013): 1559-1583*, 2013.

[74] Oscar Newman. *Defensible space*. Macmillan New York, 1972.

[75] George L Kelling and Catherine M Coles. *Fixing broken windows: Restoring order and reducing crime in our communities*. Simon and Schuster, 1997.

[76] James Q Wilson and George L Kelling. Broken windows. *Critical issues in policing: Contemporary readings*, pages 395–407, 1982.

[77] George L Kelling and William H Sousa. *Do Police Matter?: An Analysis of the Impact of New York City's Police Reforms*. CCI Center for Civic Innovation at the Manhattan Institute, 2001.

[78] Quanzeng You, Ran Pang, Liangliang Cao, and Jiebo Luo. Image-Based Appraisal of Real Estate Properties. *IEEE Transactions on Multimedia*, 19(12):2751–2759, December 2017.

[79] Elli Pagourtzi, Vassilis Assimakopoulos, Thomas Hatzichristos, and Nick French. Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21(4):383–401, 2003.

[80] Omid Poursaeed, Tomáš Matera, and Serge Belongie. Vision-based real estate price estimation. *Machine Vision and Applications*, 29(4):667–676, May 2018.

[81] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In *ACM SIGKDD*, pages 1047–1056. ACM, 2014.

[82] Keith Wardrip. Public transit's impact on housing costs: a review of the literature. *Center for Housing Policy*, 2011.

[83] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors. In *ICDM*, pages 120–129. IEEE, 2014.

[84] Joe Cortright. Walking the walk: How walkability raises home values in us cities. Technical report, CEOs for Cities, 2009.

[85] Emily Washington and Eli Dourado. The premium for walkable development under land use regulations. Technical report, George Mason University, 2018.

[86] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. Urban mobility and neighborhood isolation in america's 50 largest cities. *PNAS*, 115(30):7735–7740, 2018.

[87] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[88] D. Quercia, J.P. Pesce, V. Almeida, and J. Crowcroft. Psychological maps 2.0: A web gamification enterprise starting in london. In *Proceedings of ACM Conference on World Wide Web (WWW)*, 2013.

[89] Philip Salesses, Katja Schechtner, and César A Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013.

[90] Kevin Lynch. *The image of the city*, volume 11. MIT press, 1960.

[91] L. Porzi, S. Rota Bulo, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the ACM MultiMedia*. ACM, 2015.

[92] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. The digital life of walkable streets. In *International Conference on World Wide Web*, pages 875–884, 2015.

[93] K. Kung, K. Greco, S. Sobolevsky, and C. Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9((6):e96180), 2014.

[94] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational Approaches for Urban Environments*, pages 363–387. Springer, 2015.

[95] Maxime Lenormand, Miguel Picornell, Oliva G Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J Ramasco. Comparing and modeling land use organization in cities. *arXiv preprint arXiv:1503.06152*, 2015.

[96] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Pervasive computing*, pages 133–151. Springer, 2011.

[97] M. Traunmueller, G. Quattrone, and L. Capra. Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics*, pages 396–411, 2014.

[98] Susan S Fainstein. New directions in planning theory. *Urban affairs review*, 35(4):451–478, 2000.

[99] Peter Gordon, Sanford Ikeda, et al. Does density matter? *Chapters*, 2011.

[100] Katherine King. Jane Jacobs and "the need for aged buildings': Neighbourhood historical development pace and community social relations. *Urban Studies*, 2013.

[101] Robert Cervero. Land-use mixing and suburban mobility. *University of California Transportation Center*, 1989.

[102] Kevin Manaugh and Tyler Kreider. What is mixed use? presenting an interaction method for measuring land use mix. *Journal of Transport and Land Use*, 6(1):63–72, 2013.

[103] Ray Oldenburg. *The Great Good Place: Café, Coffee Shops, Community Centers, Beauty Parlors, General Stores, Bars, Hangouts, and How They Get You through the Day*. Paragon House Publishers, 1989.

[104] Ramon Oldenburg and Dennis Brissett. The third place. *Qualitative sociology*, 5(4):265–284, 1982.

[105] Leo W Jeffres, Cheryl C Bracken, Guowei Jian, and Mary F Casey. The impact of third places on community quality of life. *Applied Research in Quality of Life*, 4(4):333–345, 2009.

[106] Robert D Putnam. *Bowling alone: The collapse and revival of American community*. Simon and Schuster, 2001.

[107] Rémi Louf and Marc Barthelemy. A typology of street patterns. *Journal of The Royal Society Interface*, 11(101), 2014.

[108] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[109] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[110] P. Calthorpe. *The Next American Metropolis: Ecology, Community, and the American Dream*. Princeton Architectural Press, 1993.

[111] K. Lynch. *What Time is this Place?* MIT Press, 1972.

[112] C. Alexander. A city is not a tree. *Architectural Forum*, 1965.

[113] Karen Chapple and Miriam Zuk. Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape*, 18(3):109–130, 2016.

[114] Edward L Glaeser, Hyunjin Kim, and Michael Luca. Nowcasting gentrification: using yelp data to quantify neighborhood change. In *AEA Papers and Proceedings*, volume 108, pages 77–82, 2018.

[115] Michael Batty. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274–279, 2013.

[116] Hein Doeksen. Reducing crime and the fear of crime by reclaiming New Zealand's suburban street. *Landscape and urban planning*, 39(2):243–252, 1997.

[117] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Do people shape cities, or do cities shape people? The co-evolution of physical, social, and economic change in five major US cities. Technical report, National Bureau of Economic Research, 2015.

[118] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. Aesthetic capital: what makes London look beautiful, quiet, and happy? In *ACM CSCW*, 2014.

[119] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. Predicting and Understanding Urban Perception with Convolutional Neural Networks. In *ACM Multimedia*, 2015.

[120] Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.

[121] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 19–24. ACM, 2010.

[122] Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J. Ramasco. Comparing and modelling land use organization in cities. *Royal Society Open Science*, 2(12), 2015.

[123] Naren Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore–predicting the perceived safety of one million streetscapes. In *CVPRW*, 2014.

[124] Jack L Nasar. *The evaluative image of the city*. Sage Publications Thousand Oaks, CA, 1998.

[125] Hope Corman and Naci Mocan. Carrots, sticks, and broken windows. *Journal of Law and Economics*, 48(1):235–266, 2005.

[126] Kees Keizer, Siegwart Lindenberg, and Linda Steg. The spreading of disorder. *Science*, 322(5908):1681–1685, 2008.

[127] Bernard E Harcourt. *Illusion of order: The false promise of broken windows policing*. Harvard University Press, 2009.

[128] Bernard E Harcourt and Jens Ludwig. Broken windows: New evidence from new york city and a five-city social experiment. *The University of Chicago Law Review*, pages 271–320, 2006.

[129] Bernard E Harcourt and Jens Ludwig. Reefer madness: Broken windows policing and misdemeanor marijuana arrests in new york city, 1989-2000. *Criminology and Public Policy*, 2007.

[130] Kevin Lynch. *The image of the city*, volume 11. MIT press, 1960.

[131] S Milgram. The experience of living in cities. *Science*, 167(3924):1461, 1970.

[132] Jack L Nasar, Bonnie Fisher, and Margaret Grannis. Proximate physical cues to fear of crime. *Landscape and urban planning*, 26(1):161–178, 1993.

[133] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. Visual interestingness in image sequences. In *ACM Multimedia*, 2013.

[134] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *NIPS*, 2011.

[135] Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM TiiS*, 5(2):8, 2015.

[136] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *arXiv preprint arXiv:1511.07247*, 2015.

[137] Sean M Arietta, Alexei A Efros, Ravi Ramamoorthi, and Maneesh Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE TVCG*, 20(12):2624–2633, 2014.

[138] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes Paris look like Paris? *Communications of the ACM*, 58(12):103–110, 2015.

[139] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. Big data and big cities: The promises and limitations of improved measures of urban life. Working Paper 21778, National Bureau of Economic Research, 2015.

[140] Aditya Khosla, Byoungkwon An, Jasmine J Lim, and Antonio Torralba. Looking beyond the visible scene. In *CVPR*, 2014.

[141] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.

[142] Josef Sivic and Alexei A Efros. Urban-Scale Quantitative Visual Analysis. *Smart Cities*, page 43, 2014.

[143] Nikhil Naik, Ramesh Raskar, and César A Hidalgo. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *The American Economic Review*, 106(5):128–132, 2016.

[144] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A Bayesian skill rating system. In *NIPS*, 2006.

[145] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. *ECCV*, 2016.

[146] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[147] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[148] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[149] Vicente Ordonez and Tamara L Berg. Learning high-level judgments of urban perception. In *ECCV*, 2014.

[150] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[151] Michael Tiefelsdorf and Daniel A Griffith. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A*, 39(5):1193–1221, 2007.

[152] Gerda R Wekerle and Carolyn Whitzman. *Safe cities: Guidelines for planning, design, and management*. Van Nostrand Reinhold Company, 1995.

[153] Christina Pantazis. 'Fear of Crime', Vulnerability and Poverty. *British journal of criminology*, 40(3):414–436, 2000.

[154] Ralph B Taylor and Margaret Hale. Testing alternative models of fear of crime. *The Journal of Criminal Law and Criminology (1973-)*, 77(1):151–189, 1986.

[155] Marcus Felson and Ronald V Clarke. Opportunity makes the thief. *Police research series, paper*, 98, 1998.

[156] Wendy Hollway and Tony Jefferson. The risk society in an age of anxiety: situating fear of crime. *British journal of sociology*, pages 255–266, 1997.

[157] WARR Mark. Fear of victimization: Why are women and the elderly more afraid? *Social science quarterly*, 65(3):681, 1984.

[158] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[159] Stephen Hawking. *A Brief History of Time*. Bantam, 1998.

[160] S. McGreal, A. Adair, D. McBurney, and D. Patterson. Neural networks: The prediction of residential values. *Journal of Property Valuation and Investment*, 16(1):57–70, 1998.

[161] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. Urban perception of commercial activeness from satellite images and streetscapes. In *WWW*, pages 647–654, 2018.

[162] Desislava Hristova, Luca M. Aiello, and Daniele Quercia. The New Urban Success: How Culture Pays. *Frontiers in Physics*, 6, April 2018.

[163] Fei Tan, Chaoran Cheng, and Zhi Wei. Time-aware latent hierarchical model for predicting house prices. In *ICDM*, pages 1111–1116. IEEE, 2017.

[164] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *ACM SIGKDD*, pages 387–396. ACM, 2015.

[165] Xiaobai Liu, Qian Xu, Jingjie Yang, Jacob Thalman, Shuicheng Yan, and Jiebo Luo. Learning Multi-Instance Deep Ranking and Regression Network for Visual House Appraisal. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2018.

[166] Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. Real estate ranking via mixed land-use latent models. In *ACM SIGKDD*, pages 299–308. ACM, 2015.

[167] Paolo Buonanno, Daniel Montolio, and Josep Maria Raya-Vílchez. Housing prices and crime perception. *Empirical Economics*, 45(1):305–321, Aug 2013.

[168] Nicholas Boys Smith, Alessandro Venerandi, and Kieran Toms. Beyond location. Technical report, Create Streets, 2017.

[169] Robert D Dietz. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social science research*, 31(4):539–575, 2002.

[170] John R. Hipp and Adam Boessen. Egohoods as waves washing across the city: A new measure of "neighborhoods". *Criminology*, 51(2):287–327, 2013.

[171] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *ACM SIGKDD*, pages 785–794. ACM, 2016.

[172] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144. ACM, 2016.

[173] Michele Loberto, Andrea Luciani, and Marco Pangallo. The potential of big housing data: an application to the italian real-estate market. Technical report, Bank of Italy, 2018.

[174] Marc Augé. *Non-lieux. Introduction à une anthropologie de la: Introduction à une anthropologie de la surmodernité*. Le Seuil, 2015.

[175] C. Graif and Robert J. Sampson. Spatial Heterogeneity in the Effects of Immigration and Diversity on Neighborhood Homicide Rates. *Homicide Studies*, 13(3):242–260, 2009.

[176] Robert J. Sampson. Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science*, 277(5328):918–924, 1997.

[177] Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.

[178] Anne C Case and Lawrence F Katz. The company you keep: The effects of family and neighborhood on disadvantaged youths. Technical report, National Bureau of Economic Research, 1991.

[179] Clifford Robe Shaw and Henry Donald McKay. *Juvenile delinquency and urban areas.* University of Chicago Press, 1942.

[180] Travis Hirschi. *Causes of delinquency*. Transaction publishers, 2002.

[181] Robert J. Sampson. Neighborhood and crime: The structural determinants of personal victimization. *Journal of Research in Crime and Delinquency*, 22(1):7–40, 1985.

[182] Michael J Hindelang, Michael R Gottfredson, and James Garofalo. *Victims of personal crime: An empirical foundation for a theory of personal victimization*. Ballinger Cambridge, MA, 1978.

[183] Corina Graif, Alina Lungeanu, and Alyssa M Yetter. Neighborhood isolation in chicago: Violent crime effects on structural isolation and homophily in inter-neighborhood commuting networks. *Social Networks*, 2017.

[184] Robert J Sampson and Corina Graif. Neighborhood social capital as differential social organization: Resident and leadership dimensions. *American Behavioral Scientist*, 52(11):1579–1605, 2009.

[185] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime rate inference with big data. In *ACM SIGKDD*, KDD '16, pages 635–644, New York, NY, USA, 2016. ACM.

[186] David Weisburd, Elizabeth R Groff, and Sue-Ming Yang. *The criminology of place: Street segments and our understanding of the crime problem*. Oxford University Press, 2012.

[187] Martin B Short, P Jeffrey Brantingham, Andrea L Bertozzi, and George E Tita. Dissipation and displacement of hotspots in reaction-diffusion models of crime. *PNAS*, 2010.

[188] Martin A Andresen. Crime measures and the spatial analysis of criminal activity. *British Journal of criminology*, 46(2):258–285, 2006.

[189] Lucy W. Mburu and Marco Helbich. Crime Risk Estimation with a Commuter-Harmonized Ambient Population. *Annals of the American Association of Geographers*, 106(4):804–818, 2016.

[190] Cristina Kadar, Raquel Rosés Brüngger, and Irena Pletikosa. Measuring ambient population from location-based social networks to describe urban crime. In *International Conference on Social Informatics*, pages 521–535. Springer, 2017.

[191] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. The TimeGeo modeling framework for urban mobility without travel surveys. *PNAS*, 113(37):E5370–E5378, 2016.

[192] Martin A. Andresen. Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology*, 46(2):258–285, 2006.

[193] Brian G Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer, 2000.

[194] Nicholas G Polson, James G Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[195] Y C MacNab and C B Dean. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in medicine*, 19(17-18):2421–2435, 2000.

[196] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[197] Tomohiro Ando. Predictive bayesian model selection. *American Journal of Mathematical and Management Sciences*, 31(1-2):13–38, 2011.

[198] Sugie Lee, Chisun Yoo, Jaehyun Ha, and Jeemin Seo. Are perceived neighbourhood built environments associated with social capital? Evidence from the 2012 Seoul survey in South Korea. *International Journal of Urban Sciences*, pages 1–17, November 2017.

[199] Hyungun Sung and Sugie Lee. Residential built environment and walking activity: Empirical evidence of jane jacobs' urban vitality. *Transportation Research Part D: Transport and Environment*, 41:318–329, 2015.

[200] Carlos Caminha, Vasco Furtado, Tarcisio HC Pequeno, Caio Ponte, Hygor PM Melo, Erneson A Oliveira, and José S Andrade Jr. Human mobility in large cities as a proxy for crime. *PloS one*, 12(2):e0171609, 2017.

[201] Nick Malleson and Martin A Andresen. Exploring the impact of ambient population measures on london crime hotspots. *Journal of Criminal Justice*, 46:52–63, 2016.

[202] John F. Kain. A pioneer's perspective on the spatial mismatch literature. *Urban Studies*, 41(1):7–32, 2004.

[203] George E. Tita and Steven M. Radil. Making space for theory: The challenges of theorizing space and place for spatial analysis in criminology. *Journal of Quantitative Criminology*, 26(4):467–479, Dec 2010.

[204] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

[205] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[206] Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92, Dec 2016.

[207] Jameson L. Toole, Yu-ru Lin, Erich Muehlegger, Daniel Shoag, Marta C. González, and David Lazer. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.

[208] David Weisburd. Bringing social context back into the equation: The importance of social characteristics of places in the prevention of crime. *Criminology & Public Policy*, 11(2):317–326, 2012.

[209] Cristina Kadar and Irena Pletikosa. Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, 7(1):26, 2018.

[210] Ron S. Kenett, Danny Pfeffermann, and David M. Steinberg. Election polls—a survey, a critique, and proposals. *Annual Review of Statistics and Its Application*, 5(1), 2018.

[211] Isaiah Berlin. Historical inevitability. In Henry Hardy and Roger Hausheer, editors, *The proper study of mankind*. Vintage, 2013.

[212] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[213] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.

[214] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. 2012.

[215] Matt Aadland, Christopher Farah, and Kevin Magee. μ-shapes: Delineating urban neighborhoods using volunteered geographic information. *J. Spatial Information Science*, 12:29–43, 2016.

[216] Richard F Potthoff. Homogeneity, potthoff-whittinghill tests of. *Encyclopedia of Statistical Sciences*, 2006.