Analyzing Large-Scale Human Mobility Data to Address Societal Issues

by Hasan Alp Boz

Submitted to the Graduate School of Engineering and Natural Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy,

Sabanci University
December 2023

Analyzing Large-Scale Human Mobility Data to Address Societal Issues

APPROVED BY:

.....

.....

.....

.....

.....

DATE OF APPROVAL: 29/12/2023

Analyzing Large-Scale Human Mobility Data to Address Societal Issues

Hasan Alp Boz

Computer Science and Engineering, PhD Dissertation, 2023
Thesis Supervisor: Selim Balcısoy

Keywords: Computational Social Science, Human Mobility, Urban Analysis, Informed Policymaking

Abstract

Human mobility stands as an indispensable catalyst shaping the fabric of societies worldwide, serving as a critical component in understanding societal behaviors and influencing the formulation of effective policies. This dissertation explores the intricate interplay between human mobility patterns and the formation of impactful societal policies by employing mobility networks extracted from large-scale human mobility data with varying granularities, offering insights for informed policymaking through two distinct case studies. The first case study centers on advancing local economies, while the other scrutinizes the factors influencing community adaptability during the COVID-19 pandemic. Two primary investigations unfold in this dissertation. The prediction of business financial performance using customer co-location networks derived from credit card transactions, employing network modeling techniques rooted in human mobility data; and the analysis of neighborhood adaptability during the pandemic through smartphone-based mobility data, evaluating interventions' impacts on diverse sociodemographic groups and changes in mobility networks. It aims to offer innovative contributions by developing a novel framework for predicting business financial well-being utilizing privacy-enhanced network-based features extracted from customer co-location networks and providing insights into neighborhood adaptability during the pandemic, taking into account geographic and sociodemographic factors, in addition to amenity accessibility. Ultimately, this study aspires to equip policymakers with well-informed insights gleaned from human mobility data, fostering the formulation of adaptive, inclusive policies tailored to address the evolving societal landscape. The research presented in this dissertation holds the potential to significantly influence informed policymaking concerning communities' adaptability in the wake of exogenous shocks and the vitality of local economies.

Sosyal Sorunları Ele Almak İçin Büyük Ölçekli İnsan Hareketliliği Verilerinin Analiz Edilmesi

Hasan Alp Boz

Bilgisayar Bilimi ve Mühendisliği, Doktora Tezi, 2023
Tez Danışmanı: Selim Balcısoy

Anahtar Kelimeler: Hesaplamalı Sosyal Bilimler, İnsan Hareketliliği, Kentsel Analiz, Bilgiye Dayalı Politika Oluşturma

Özet

İnsan hareketliliği, toplumların dokusunu şekillendiren önemli bir katalizör olarak durmaktadır. Toplumsal davranışları anlamanın kritik bir unsuru olarak hizmet ederken etkili politikaların oluşturulmasını da şekillendirmektedir. Bu tez, büyük ölçekli insan hareketliliği verilerinden çıkarılan hareketlilik ağlarını kullanarak insan hareketliliği desenleri ile bilgiye dayalı toplumsal politikaların oluşumu arasındaki karmaşık etkileşimi keşfetmekte ve iki farklı vaka çalışması ile bilgi sağlamaktadır. İlk çalışma yerel ekonomilerin ilerlemesi üzerine odaklanırken, diğeri COVID-19 pandemisi sırasında toplumların uyum kabiliyetini etkileyen faktörleri inceler. Bu tezde iki temel araştırma ortaya konmaktadır. Kredi kartı işlemlerinden elde edilen müşteri bir araya gelme ağları kullanılarak işletmelerin finansal performansının tahmini; ve akıllı telefon tabanlı hareketlilik verileriyle pandemi sırasında mahallelerin uyum kabiliyetinin analizi, müdahalelerin çeşitli sosyodemografik gruplar üzerindeki etkilerini ve hareketlilik ağlarındaki değişiklikleri değerlendirir. Bu çalışma, müşteri bir araya gelme ağlarından çıkarılan gizlilik artırılmış ağ tabanlı öznitelikler kullanarak işletmelerin finansal iyilik halini öngörme için yeni bir çerçeve geliştirerek ve pandemi sırasında mahallelerin uyum kabiliyetine dair coğrafi ve sosyodemografik faktörleri, yanı sıra olanak erişilebilirliğini de dikkate alarak bilgi sunmayı amaçlamaktadır. Sonuç olarak, bu çalışma insan hareketliliği verilerinden elde edilen bilgiyle politika yapıcılarına donanımlı, evrilen toplumsal manzaraya uygun, adaptif ve kapsayıcı politikalar oluşturma imkanı sağlamayı hedeflemektedir. Bu tezde sunulan araştırma, dış kaynaklı şokların ardından toplumların uyum kabiliyeti ve yerel ekonomilerin canlılığıyla ilgili bilgi sahibi politika oluşturmayı önemli ölçüde etkileyebilecek potansiyele sahiptir.

To my beloved family...

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Selim Balcısoy for his constant support, guidance, and understanding throughout the past four and a half years, which also coincided with the distressful pandemic period. Without his encouragement and supervision, it would not have been possible to complete this dissertation.

In addition, I would like to thank Prof. Burcin Bozkaya, Prof. Nina Mazar, Dr. Ulku Ozturk, Dr. Mohsen Bahrami, and Dr. Yoshihiko Suhara for their invaluable comments and feedback during my PhD. Especially, I would like to express my gratitude to Dr. Mohsen Bahrami as he played a pivotal role in my academic journey with his constant guidance and leadership whenever I needed.

I would also like to thank Assoc. Prof. Dr. Öznur Taştan Okan, Asst. Prof. Dr. Onur Varol , Asst. Prof. Günce Orman, and Assoc. Prof. Dr. Vinicius Brei for accepting to take part in the dissertation jury and providing insightful comments and feedback on my work.

I would like to express my gratitude to BAVLAB members as well. I would like to thank Mert Gurkan, Fatih Oztank, Atra Bahceci, Atakan Saracyakupoglu, Berke Odaci, and Beyza Cokkececi for the good times in the lab. Also, I would like to thank Yasin Findik for his constant support over the last six years.

I am also thankful to Prof. Alex 'Sandy' Pentland for advising and hosting me in his research group at MIT Media Lab. It was a fruitful and enlightening six months that I will cherish with fond memories. From the lab, I would like to thank Prof. Esteban Moro, Dr. Takahiro Yabe, Isabella Loaiza Saa, Bernardo García Bulle Bueno, Yan Asadchy, Peter Edsberg Møllgaard, Levin Brinkmann, Massimiliano Luca, Agnese Sacchi, Tobin South, and Robert Mahari.

I am also thankful to the Scientific and Technological Research Council of Turkey (TUBITAK) for supporting part of my research under the BIDEB 2214-A fellowship program.

And finally, I would like to thank my family for their never-ending support and love throughout my entire life, without them, nothing would have been possible.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Society is a broad and interconnected web of individuals, groups, cultures, and norms that collectively form a functioning community. At its core, society represents the complex fabric within which people share common goals, behaviors, and organizational structures. In the intricate fabric of societies, a complex array of needs and goals is often met through the policies implemented by policymakers, i.e., public administrations, organizations, and governments [87]. These policies serve as the guiding threads embedded into the social framework, aiming to address fundamental challenges, promote stability, and foster advancement. From healthcare [124] and environmental sustainability [125] to economic growth [50] and exogenous shocks [105], societies are in need of appropriate policies to guide them in fulfilling these essential requirements.

Data-driven policymaking, a methodology utilizing cues and observations extracted from data to shape and enhance policies [39], has drawn considerable attention among policymakers due to its demonstrated effectiveness across diverse domains, from educational systems [48] to humanitarian aids [7]. Considering the well-being of societies, data-driven policymaking approaches require data sources that reflect the socioeconomic dynamics of the target area. Behavioral data sources constitute the backbone of data-driven policymaking methodologies. Amidst these behavioral data sources, one of the most compelling and intricately connected facets emerges, *human mobility*.

Mobility stands as an indispensable aspect of human civilization, embodying the essence of exploration and adaptation. The diverse forms of movement, ranging from daily commutes to international migrations, encompass a spectrum of motivations and consequences. In the intricate web of societal dynamics, human mobility emerges as a catalyst for change, a tool for addressing pressing challenges, and a

facilitator of progress.

At its core, the role of human mobility transcends mere transportation or relocation. It encapsulates the exchange of ideas, the fusion of cultures, and the dissemination of knowledge. Economic growth finds its momentum through the movement of workers and professionals across varying urban scales, fostering trade, investment, and technological developments. Human mobility serves as a driving force for evolution and advancement, yet often impaired by the disruptive impact of epidemics, the turmoil of wars, and the strain on social cohesion.

As policies endeavor to tackle societal needs, the flow of human movement, whether voluntary or forced, plays a pivotal role. The interplay between policies addressing pressing societal challenges and the impact of human mobility creates a landscape where considerations of displacement, adaptability, and movement patterns intersect with the fundamental pillars of societal well-being.

A substantial research effort has been devoted to uncovering how human mobility patterns can be collected, analyzed, and harnessed to understand societal behaviors and preferences, shedding light on the intricate nature of our movements. This exploration fuels a growing body of research [7, 104, 117] aimed at deciphering the complexities of human interactions, enabling a deeper comprehension of our communities' dynamics and informing strategies for policymaking.

By 2050, it's projected that 68% of the world's population will be living in cities [1]. This anticipated surge emphasizes the crucial role cities play as centers of opportunity but more importantly, underscores the pressing need for holistic policymaking that addresses a multitude of different social issues for the target residents. To this end, human mobility analysis stands out as an invaluable framework for assisting policymakers in designing responsive and inclusive societal policies.

Segregation, characterized by systematic disparities among diverse sociodemographic groups, is a threat to the cohesion of societies. To measure the segregation within an urban area, researchers traditionally employ low-granularity residential patterns obtained from census data. However, with fine-grained human mobility data, research [104] has shown that the daily encounters of individuals from different sociodemographic backgrounds enable an in-depth evaluation of segregation in an urban setting.

Urban crime analysis is another field that is enhanced by human mobility data. Caminha et al. [29] create mobility flows between urban areas, based on bus trips, with occurrences of property crime. The application of human mobility data extends to disaster management as well. Modeling the flow of masses during and after such a devastating incident plays a crucial role in the resilience and well-being of communities [177]. With the onset of the COVID-19 pandemic, a collective research effort focused on mobility analysis to comprehend and adapt to the drastic shifts in

human mobility patterns, which has been instrumental in elucidating the spread of the virus [136], gauging the effectiveness of containment measures [31], and devising strategies to navigate these unprecedented challenges [72].

In addressing societal challenges, human mobility analysis emerges as both a solution and a challenge in itself. Understanding and harnessing the intricacies of human movement presents an opportunity to develop targeted solutions, yet navigating the complexities of this data poses its own set of hurdles. The breadth and depth of insights gleaned from human mobility patterns can fluctuate based on the employed statistical frameworks, feature extraction methods, and the data source, leading to varying capabilities and outcomes.

This dissertation endeavors to spotlight the application of human mobility data based on mobility networks extracted from varying granularities in informed policymaking through two case studies. One focuses on informed policymaking aimed at enhancing local economies, while the other delves into analyzing the drivers behind communities' adaptability during the COVID-19 pandemic. The work presented in this dissertation is expected to have implications for informed policymaking on the adaptability of communities in the face of exogenous shocks and on the vitality of local economies.

1.2 Research Questions and Overview

This dissertation aims to answer the following question: How can human mobility data be effectively utilized to enhance and guide policymaking initiatives toward addressing societal issues? To this end, this dissertation first focuses on the well-being of local economies by employing a financial transaction data source. And lastly, for the analysis of neighborhood adaptability during the COVID-19 pandemic, smartphone-based large-scale mobility data is employed.

In Part I, I present a study on enhancing local economies that predicts the future financial performance of businesses based on customer co-location networks extracted from credit card transactions from an OECD country. From the constructed co-location networks, a set of network-based features are extracted for businesses to be fed into machine learning models. Here, the proposed set of network-based features addresses privacy concerns of financial information.

In Part II, I introduce an in-depth study on neighborhood adaptability indicators by using neighborhood-level mobility networks constructed from large-scale smartphone data that targets one of the economic hubs of the world, New York City. I analyze to what extent mobility networks changed from a topological perspective in different sociodemographic groups and conduct simulations on mobility patterns under hypothetical point-of-interest densities.

1.2.1 Part I: Enhancing Local Economies with Human Mobility Data

The study presented in Part I aims to answer the following questions:

(a) How could human mobility be used to predict business financial well-being instead of exclusively collected internal financial metrics?

(b) What data structures are needed for such an endeavor?

(c) What are the advantages of employing human mobility data for financial performance prediction?

The work presented in this part is motivated by social physics [16, 121] and computational social science [90], in which network modeling constitutes the backbone of the conducted study. Customer visitation patterns are treated as a proxy indicator for the economic activity of businesses. Business-level network-based features, i.e., centrality metrics and node2vec embeddings [64], are fed into machine learning models to predict the future financial performance of businesses.

1.2.2 Part II: Examining Neighborhood Adaptability for Informed Policymaking

The primary objectives of the study outlined in Part II are to address the following questions:

(a) How can we analyze the impact of non-pharmaceutical interventions (NPIs) on different sociodemographic groups in a way that captures the complexity of human mobility dynamics?

(b) What sociodemographic traits are significant in explaining the impact?

(c) Are there certain neighborhoods that act as a bridge for the virus to spread?

(d) How would this impact change under hypothetical scenarios involving varying point-of-interest densities?

This study primarily scrutinizes the evolving structure of mobility networks during the pandemic, employing robust network metrics to understand the dynamics within different sociodemographic groups. By using large-scale smartphone mobility data, I construct neighborhood-level mobility networks enriched with census attributes and quantify to what extent network structures changed for different sociodemographic groups. Moreover, to understand the spreading dynamics in an

urban area, I define bridge neighborhoods that act as intermediary areas for infection spreading and analyze the characteristic traits of the consequent neighborhoods. Lastly, I employed the Huff gravity model to simulate the mobility flows under hypothetical point-of-interest densities.

1.3 Contributions

The contributions of this dissertation are delineated as follows:

1. Enhancing Local Economies with Human Mobility Data (Chapter 4)

    - Based on social physics, a novel financial performance prediction framework for businesses is proposed based on customer co-location networks.
    - A novel approach is proposed to determine business performance labels while ensuring that the assigned labels show no biases concerning geographical locations, income levels of residents in the area, or the socioeconomic status of customers.
    - The proposed co-location networks are employed to produce network-based and node2vec features that provide a higher level of safeguarding against network construction attacks that aim to recover the private financial information of businesses.

2. Examining Neighborhood Adaptability for Informed Policymaking (Chapter 5)

    - Highlighting the role of geographic constraints, amenity accessibility, and sociodemographic characteristics in the adaptability of neighborhoods during the COVID-19 pandemic.

1.4 Structure of the Dissertation

Chapter 2 highlights the concepts and methodologies employed in studies concerning human mobility analysis. It introduces these methods in a manner accessible to casual readers while providing the essential mathematical foundations that underlie them. Chapter 3 gives an overview of data-driven policymaking approaches based on human mobility data and discusses the related work. Chapter 4 presents the novel use of human mobility data on businesses' financial performance prediction based on customer co-location networks. Chapter 5 presents the study on neighborhood-level mobility networks to analyze the impact of the COVID-19 pandemic on different sociodemographic groups. Finally, in Chapter 6, a comprehensive conclusion is drawn, highlighting the key findings and insights from the conducted analyses. Moreover, this chapter discusses the limitations of the conducted analyses while also illuminating potential future directions and areas for further exploration.

Chapter 2

Concepts & Methods

Mobility is an integral component of human lives. It is a complex and multi-faceted phenomenon as it's driven by a myriad of socioeconomic factors. Given its significant impact, policymakers around the world have dedicated resources to acquiring technologies and methodologies for collecting mobility patterns at different scales and granularities, which have been employed in designing and developing policies and systems such as transportation infrastructure [146], urban planning [68], and epidemic modeling [31].

To tackle such ambitious societal issues, the appropriate set of data collection resources and analytical frameworks are essential given the complexity and the scale of the potential impact on our societies. The case studies outlined in this dissertation extensively utilize these analytical frameworks alongside mobility datasets. Both case studies leverage spatial metrics to interpret and derive insights from the available mobility data. Moreover, to model the population-level mobility patterns in urban areas, mobility analysis frameworks from the literature are employed. In this chapter, I will first explain the existing data collection methods and provide an evaluation of the strengths and weaknesses inherent in each. In subsequent sections, well-established human mobility modeling approaches from the literature will be outlined.

2.1 Data Collection Methods

2.1.1 Survey-Based Methods

Historically, governments around the world have relied on on-site surveys to harvest public opinions and behavior patterns for distinct ends. For instance, the U.S. Census Bureau[1] has been collecting periodical census data by sending surveys to sampled households to gauge the sociodemographic and economic status of residents

[1] `https://www.census.gov/`

at varying administrative levels (e.g., census tracts). In this setting, to determine the mobility patterns of residents, they employed an array of questions related to the location of their workplace, their previous residency location, or their usage of public transportation, which can then be utilized to form commuting flows in urban areas or migration patterns at country scale.

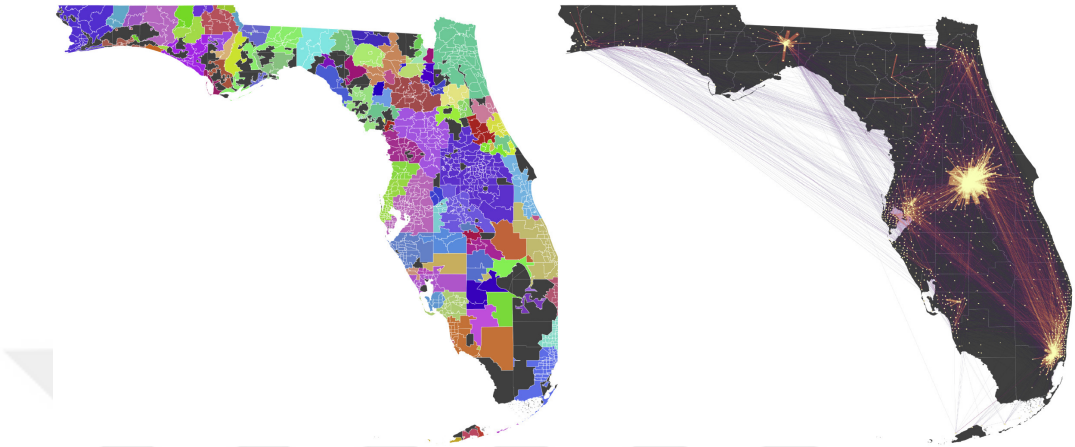

Figure 2.1: Mobility flows generated from residents' work locations. On the left, the counties in Florida, US are highlighted. On the right, the resulting flows between the counties are displayed.
Source: Image obtained from [18].

To construct mobility patterns in an urban setting, answers to such questions are utilized. For instance, to have a proxy of the mobility flow from urban areas A_i to A_j, the number of residents living in A_i and working A_j can be harnessed, which can then be generalized to every urban area to create flow networks in the selected region for the given time frame. The resulting network can then be enriched by the previously asked mode of transportation and travel time questions. On a larger scale, the same methodology can be deployed to assess the migration flow between cities as well as displayed in Figure 2.1.

However, such survey methods rely on the representativeness of the selected household sample, which could be quite biased in case several precautions [156] are not taken. Furthermore, the granularity of the analysis would be affected by the conducted survey. The deployed surveys aim to evaluate the overall behavioral patterns of the residents in an aggregated spatio-temporal domain. For instance, the U.S. Census Bureau releases aggregated 5-year and 1-year period estimates, which hinders the analyses that require a more nuanced spatio-temporal granularity.

Another drawback of the survey-based data collection methodologies comes from the heterogeneity problem. Surveys conducted in different locations in the world are inherently not uniform. Governments, depending on their political agenda and economic status, either do not incorporate mobility-related questions or do not publish

urban-level mobility patterns. Such discrepancies led researchers to combine survey-based data with more granular third-party datasets [7].

Survey-based approaches rely on labor-intensive, on-site questionnaires that aim to collect residents' aggregated socioeconomic behavioral patterns. They are based on the assumption that the respondents give faithful answers to the presented set of questions. However, surveys are inherently prone to biases, such as non-response [42] and social desirability biases [130]. Moreover, the respondents require a designated time slot to complete the given questionnaire. Researchers, driven by these concerns, began exploring alternative data sources on mobility patterns.

Despite their drawbacks, survey-based mobility patterns have been instrumental in developing data-driven policymaking. Researchers may rely on such survey data as confirmatory outlets to support their analyses in case the survey data lacks the required spatio-temporal granularity.

2.1.2 Call Records

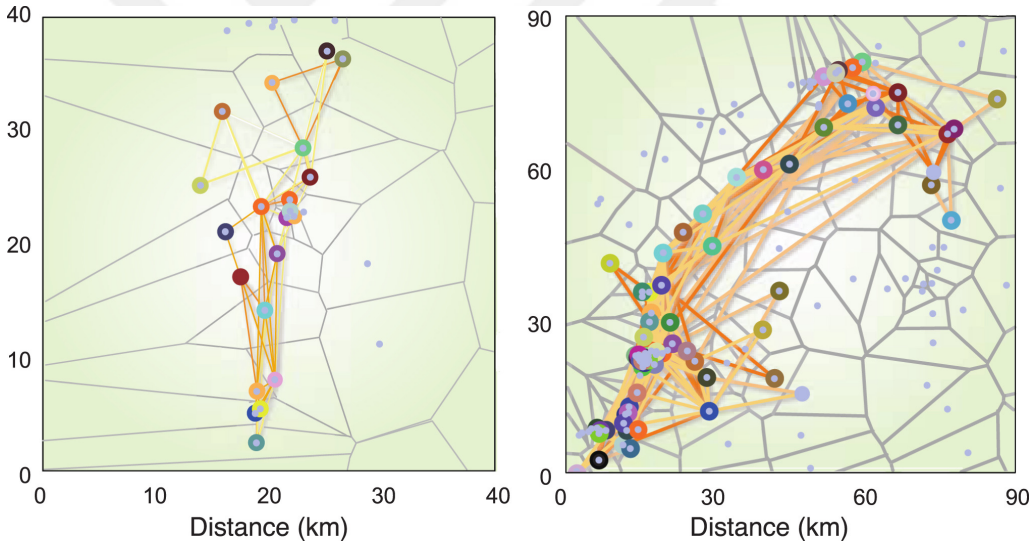

Figure 2.2: Trajectories of two mobile phone users (left and right) extracted from their corresponding CDR data. Each dot represents a cellular tower from which the call is recorded and edges between them connect consecutive calls to form trajectories over the course of the time frame.
Source: Image obtained from [145].

An alternative to survey-based approaches comes from utilizing the byproduct metadata of communication records. Every telecommunication (i.e., direct call or short message service) between two entities produces a special call log named Call Detail Record (CDR) that stores various metadata such as IDs of both parties, call starting time, call duration, and cellular tower information. Each cell tower is responsible for the data exchange in a designated geographical area. Pinpointing the

geo-location of cellular towers provides an estimate of mobile phone users' location throughout the day, which can then be used to form individual mobility trajectories as displayed in Figure 2.2.

CDR data is most often anonymized before being provided to the researchers. In addition, it also requires a pre-processing step to convert raw call metadata to trajectories [57]. The quality of the resulting trajectories relies on the frequency of recorded geo-locations. In addition, during a call, a signal may be transmitted over different nearby towers depending on the caller's position, which in turn may hinder the trustworthiness of the resulting trajectories. Furthermore, the number of cellular towers is correlated with the population density. In densely populated urban areas, there might be a multitude of different cellular towers to compensate for a high volume of communication. In contrast, in rural areas, the low number of cellular towers may curb the trajectory formation step.

Communication stands as an essential facet of human interaction. Across the globe, irrespective of prevailing economic conditions, mobile phones serve as the universal means through which people connect with one another. As a result, CDR data, and the ensuing mobility trajectories, stand out as a frequently employed data source due to its prevalence and well-established literature.

2.1.3 WiFi Logs

In bustling urban settings, such as university campuses, WiFi routers are strategically positioned throughout the area to cater to the high demand for internet access among the crowds. Users connect to nearby routers to access the Internet. Routers periodically produce device connection logs containing connection/disconnection timestamps, device MAC address, and other connection metadata. Based on device MAC addresses, a comprehensive pathway of connections can be assembled to map out the trajectory of a specific user, which can subsequently be translated into discernible mobility patterns, aided by the knowledge of router locations.

Researchers leverage WiFi connection logs to construct such mobility trajectories and analyze movement patterns in various urban densities. Gunce et al. [116] collect connection logs from a university campus as displayed in Figure 2.3, and tackle the dynamic network extraction problem based on the resulting mobility networks. Although cities have a lower router density compared to a university campus, a body of literature [78, 132, 159] focuses on city-wide WiFi connection logs to create and analyze mobility patterns.

In WiFi connection logs-based studies, the first concern is the anonymity of the users. To ensure privacy protection, device MAC addresses undergo anonymization before they are employed in an analysis. Furthermore, a significant pre-processing

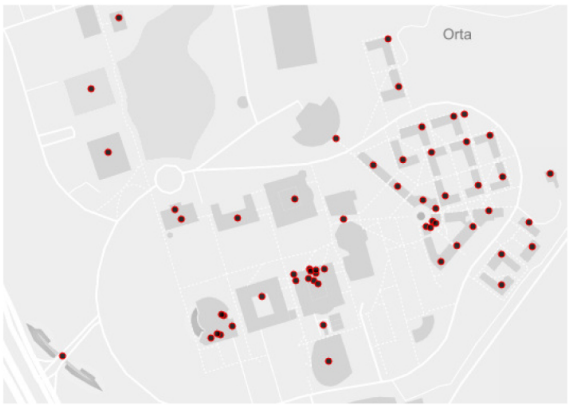

Figure 2.3: WiFi router distribution in a university campus. Every router is strategically positioned to ensure comprehensive coverage across the entire area. A priori known router positions are employed to create mobility trajectories from connection logs.
Source: Image obtained from [116].

step is required as well. Connection logs are produced periodically, which in turn may generate millions of records given the span of the time frame and number of devices. Based on proximity and the surrounding environment, a device might establish connections with multiple routers within a session, potentially leading to its identification in the vicinity of various routers. In such cases, to accurately estimate a device's location, pre-processing plays a crucial role.

2.1.4 GPS Data

Global Positioning System (GPS) is used to determine the precise geo-location of a device with the help of satellites around the globe. Periodically, the device iteratively transmits and receives a radio signal to the reachable satellites (at least 4), which in turn calculates the corresponding geo-location of the device on Earth. Due to its precision, the mobility trajectories derived from GPS data exhibit a high degree of accuracy.

In addition to the studies that involve custom-made GPS devices [128], researchers also rely on transportation systems as most of the vehicles are equipped with built-in GPS components. For instance, the New York City taxi trips dataset has been employed by researchers to understand the mobility patterns in a densely populated urban area [129]. Figure 2.4 displays the density of taxi trip locations in New York City, highlighting the gravity of Manhattan in the area.

With the emergence of smart mobile phones, GPS became an integral component of user experience as it constitutes the backbone of navigation applications. In addition, many mobile applications also harvest the geo-location of their users and share it with third-party entities. For instance, SafeGraph [131] is a data consortium

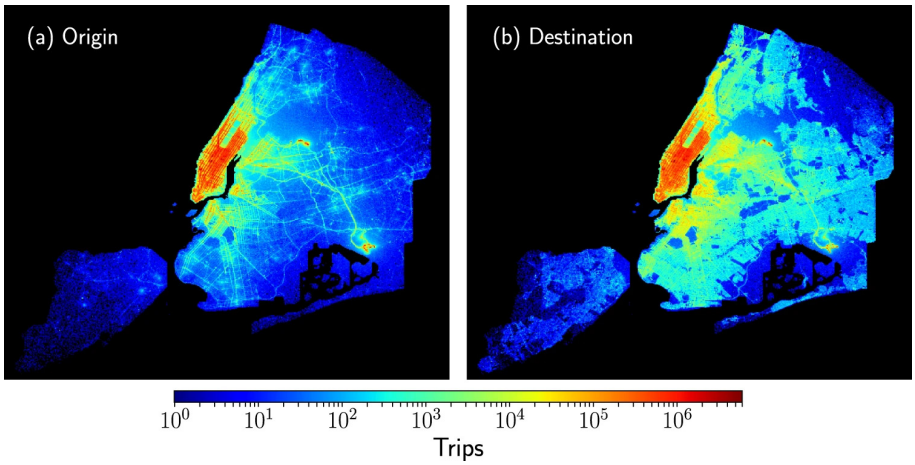

Figure 2.4: The density map of origin and destination of taxi trips that took place in New York City, harvested from GPS data.
Source: Image obtained from [129].

that collects periodic user geo-location data through a set of mobile phone applications. The resulting location data is then matched with point-of-interest locations in urban areas to determine their visit patterns throughout the target time frame.

Custom-made GPS devices present a challenge for researchers aiming to broaden their scope due to their specialized nature, often tailored for specific settings. Their limitations in terms of accessibility and standardization make it harder to collect comprehensive mobility patterns across diverse user groups or scenarios. Conversely, mobile phone-based GPS data offers increased accessibility and manageability. With the ubiquity of smartphones, this data source provides a more expansive and diverse pool of information, reflecting various demographics and behaviors. Moreover, the consistent integration of GPS capabilities into mobile devices streamlines data collection, enabling researchers to tap into extensive user-generated datasets seamlessly.

2.1.5 Financial Transactions

Economics plays an essential role in our lives. It's the mechanism through which consumers utilize their savings and wealth to access goods and services spread across time and space. Transactions between entities constitute the backbone of most economic systems. Considering its significance and impact, creating historical transaction records plays a vital role for governments, financial institutions, and even individuals. Such historical records, accompanied by appropriate metadata, may reveal invaluable insights regarding movement patterns.

Commuters rely on various forms of public transportation as their primary mode of commuting to reach their workplaces. Based on a report published by the New York City Planning Committee [114], each day, close to 1 million individuals commute into New York City for employment, accounting for roughly 20% of the city's

total workforce as of 2019. With the help of personalized transportation cards, commuters gain access to transportation services. Every time a commuter uses a transportation card, the timestamp of the transaction, the ID of the commuter, and the station ID are recorded, which in turn can be employed to construct timely movement trajectories for commuters. Xia et al. [172] employ transaction records produced from subway smart cards and analyze mobility patterns of thousands of individuals on weekdays and weekends. However, data obtained from public transportation systems are prone to be incomplete and biased toward commuters.

Despite the growing popularity of online financial transactions in today's economic landscape, traditional on-site payments, such as cash and credit cards, continue to hold considerable significance. Each transaction over a credit card, which is processed by a point-of-sale (POS) device, generates a set of metadata regarding the purchase, such as customer ID, payment amount, and store ID. Based on customer and store IDs, researchers can construct mobility data based on the historical purchase patterns of individuals. In such studies [14, 141], anonymized customer credit card transactions are employed to construct mobility patterns based on store IDs.

2.1.6 Online Behavior Data

Over the past three decades, the Internet has evolved into an indispensable part of human lives. Initially serving as a means of communication, its scope has expanded significantly; from financial transactions to health monitoring, a vast array of applications are all managed over the Internet. Moreover, social media platforms have been instrumental in shaping our everyday lives. These platforms have transcended mere communication tools; they have become hubs for social interaction, news dissemination, entertainment, and business networking. To enhance target marketing and better content recommendation, social media platforms, such as Facebook and Twitter, began periodically capturing geo-tagged data from user interactions. Because of their expansive user base, comprising millions of individuals globally, social media platforms generate invaluable geo-tagged data. This data provides intricate insights into the mobility patterns of diverse populations, spanning various backgrounds and locations. By extracting and analyzing this rich stream of geo-tagged information from social media outlets, researchers gain access to comprehensive snapshots of how different demographics move and interact within their respective environments.

Researchers began to compile proxy mobility patterns that emerged as a byproduct of user interaction or platform-specific functionalities. For instance, Facebook[2]

[2] `https://www.facebook.com/business/ads`

enables business owners to manage their advertisement campaigns targeting audiences from a certain sociodemographic cohort. An advertisement owner may impose a set of demographic criteria for their campaign. One such criterion selects the users who *used to live in country X* and now *live in country Y*. Based on this criterion, Spyratos et al. [148] formed migration flows between countries given users' current and previous residency locations.

Most of the social media platforms allow their users to geo-tag their posts. For instance, Twitter[3] allows their users to geo-tag (i.e., coordinates or place names) their tweets. Depending on the volume of posted geo-tagged tweets by a user, the geo-tags may be employed to create daily trajectories or aggregated visits to a particular area. Jurdak et al. [82] harvest geo-tagged tweets, construct individual mobility trajectories, and analyze the statistical properties of human mobility based on Twitter.

With the advance of Location-Based Social Networks (LBSN), user interactions are now facilitated based on geographical locations. FourSquare[4], a well-known LBSN, provides a platform for their users to log their visits (i.e., check-ins) to assorted point-of-interests and leave a comment as well. Such visit logs can then be collected over time to construct mobility patterns for individuals or aggregated patterns [40, 98, 108].

Behavior data extracted from social media outlets have not only been used for mobility analysis. Their usage covers a wide range of studies, from disaster response systems [106] to emotion analysis [120]. However, working with social media data requires additional precautions from a mobility analysis perspective. An extensive data cleaning step is necessary, such as removing highly suspicious user activities, incomplete records, and irregular movement patterns. In addition, to account for the sampling bias, the representation of the target urban area in the data must be checked accordingly to make sure that it is representative.

Table 2.1 offers a comparison of the described data sources essential for understanding mobility patterns. Survey-based methods provide aggregated socioeconomic behavior patterns and historical census data, yet they suffer from limited spatio-temporal granularity. Call records generate individual mobility trajectories and stand out as a prevalent, well-established data source, although challenges arise from anonymization and reliance on signal transmission and tower density for data quality. WiFi logs assemble comprehensive connection pathways, aiding mobility understanding, but require extensive anonymization and pre-processing while generating large volumes of data. GPS data yields high precision but faces limitations in accessibility and standardization for custom devices. Financial transactions offer

[3]https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo
[4]`https://location.foursquare.com/developer/`

insights into economic and commuting patterns, leveraging diverse sources like transportation cards and credit card transactions, but encounter issues of incompleteness, biases, and privacy concerns. Online behavior data captures diverse population behaviors from social media and location-based services, yet entails challenges of data cleaning, sampling biases, and representativeness issues. Overall, each data source presents unique advantages and drawbacks, catering to different analytical needs while demanding careful consideration of limitations and biases inherent in their collection and utilization for mobility analysis.

Data Source	**Advantages**	**Drawbacks**
Survey-Based Methods	Provides aggregated socioeconomic behavioral patterns; historical and periodic census data availability.	Limited spatio-temporal granularity.
Call Records	Produces individual mobility trajectories; prevalent and well-established data source.	Anonymization challenges; data quality relies on signal transmission and tower density.
WiFi Logs	Assembles comprehensive pathways of connections; aids in understanding urban mobility patterns.	Anonymization and preprocessing required; a large volume of data generation; accuracy reliant on device connections.
GPS Data	High precision in geo-location; applicable in both custom devices and mobile phones.	Limitations in accessibility and standardization for custom devices; potential privacy concerns.
Financial Transactions	Insightful in economic analysis and commuting patterns; diverse data sources (e.g., transportation cards, credit card transactions).	Incompleteness and biases in public transportation data; privacy concerns in credit card transactions.
Online Behavior Data	Captures diverse population behaviors; provides insights from social media and location-based services.	Data cleaning challenges; sampling biases; representativeness issues.

Table 2.1: Comparison of different mobility data sources.

2.2 Spatial Analysis

In analyzing the collected mobility data, commonly used spatial metrics from the literature are employed. In this section, metrics and methodologies used to characterize the ensuing mobility patterns will be explained in detail.

2.2.1 Displacement

To understand and characterize the nature of the obtained mobility patterns, the researchers made use of spatio-temporal metrics and frameworks. Among the pivotal traits regarding mobility patterns is the distance individuals traverse within a particular timeframe. To this end, understanding the distance distribution significantly contributes to characterizing mobility patterns. This involves quantifying the distance between consecutive visited locations sourced from mobility records, thereby depending on the data origin. With precise geo-locations, such as CDR and GPS data, it is possible to quantify the displacement between consecutive instances, offering a granular understanding of the movement patterns. To compute the displacement an individual covered within a specific timeframe, it's essential to identify the *stops*—instances when the individual paused movement and spent a significant amount of time. Such stop location can then be used to quantify the displacement, Δr, over a certain timeframe. In data sources where each record inherently denotes a stop—such as online behavioral data or financial transactions—the *jump length* metric is used as a quantifier for displacement. Analyzing the subsequent displacement distribution may reveal the overall mobility patterns to understand the likelihood of an individual traveling a distance at a certain time step.

A large body of literature focuses on the characterization of the displacement distribution. Based on empirical observations, researchers have observed mobility patterns following different distributions depending on the urban scale. A majority of the studies note that it follows a power law distribution, $P(\Delta r) \sim \Delta r^{-\beta}$. However, the findings from Noulas et al. [112] suggest that power law may not be able to capture intra-urban displacements. Other works have marked that short-distance displacement patterns can be fitted as an exponential curve, $P(\Delta r) \sim e^{-\lambda \Delta r}$. Alessandretti et al. [8] marks that displacement distribution may also be represented as a log-normal distribution, $P(\Delta r) \sim \frac{1}{\Delta r} \star e^{(-log \Delta r - \mu)^2 / 2\sigma^2}$.

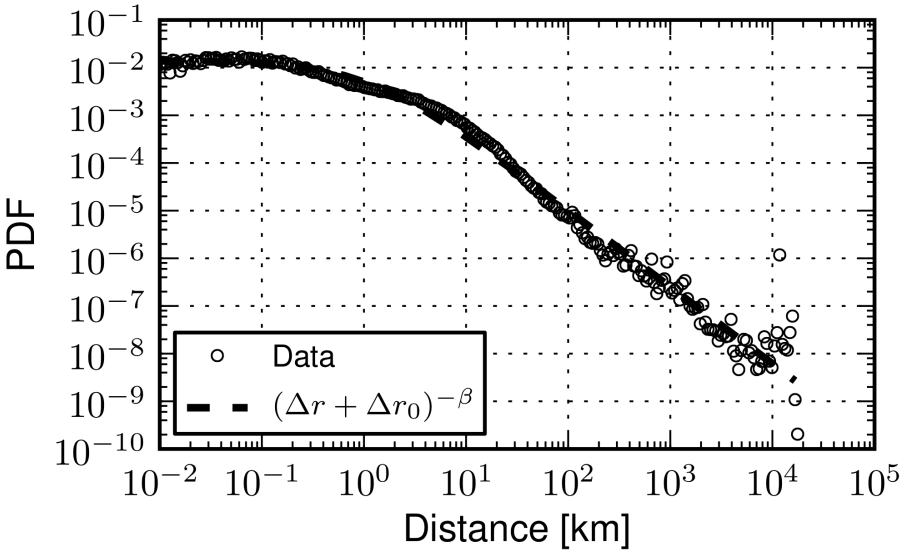

Figure 2.5: Displacement patterns following a power law distribution.
Source: Image obtained from [112].

The radius of gyration is another metric employed in the literature to evaluate the displacement characteristics of mobility data. The radius of gyration is calculated by measuring the root mean square distance between a central point, r_0, such as an individual's home location or the center of mass of their movement trajectory, and all the different stops or locations, r_i, visited by that individual as displayed in Equation 2.1. Papplalorda et al. [119] employ a modified version of the radius of gyration to understand the characteristics of returners and explorers, a class of individuals that differentiates based on their recurrent mobility patterns.

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_0)^2} \tag{2.1}$$

2.2.2 Visitation Patterns

In human mobility analysis, it's crucial to differentiate between locations based on their significance in terms of visitation patterns, i.e., how frequently they are visited in a certain time frame. To quantify the significance of a location, visitation rankings can be employed. In that case, the most visited locations, such as home locations and workplaces, will have higher ranks, whereas locations visited infrequently would receive lower rankings. Gonzalez et al. [62] analyze CDR data and rank locations based on total visits. Their results indicate that the rank of a location can be approximated by Zipf law, in which the probability of a user being situated at a location with rank R is $P(R) \sim 1/R$, meaning that a location's visitation rank is inversely proportional to its visitation frequency.

2.2.3 Origin-Destination Matrix

An origin-destination (OD) matrix is a fundamental tool in human mobility analysis that captures the flow of movements (aggregated mobility) between locations in differing urban scales, e.g., census tracts, counties, cities, countries, or custom tesselations. Given an urban scale, an OD matrix is constructed for n origin (start point) and m destination (endpoint) locations that result in a n x m matrix in which cell M_{ij} stores the mobility flow from origin location i to destination location j as depicted in Figure 2.6.

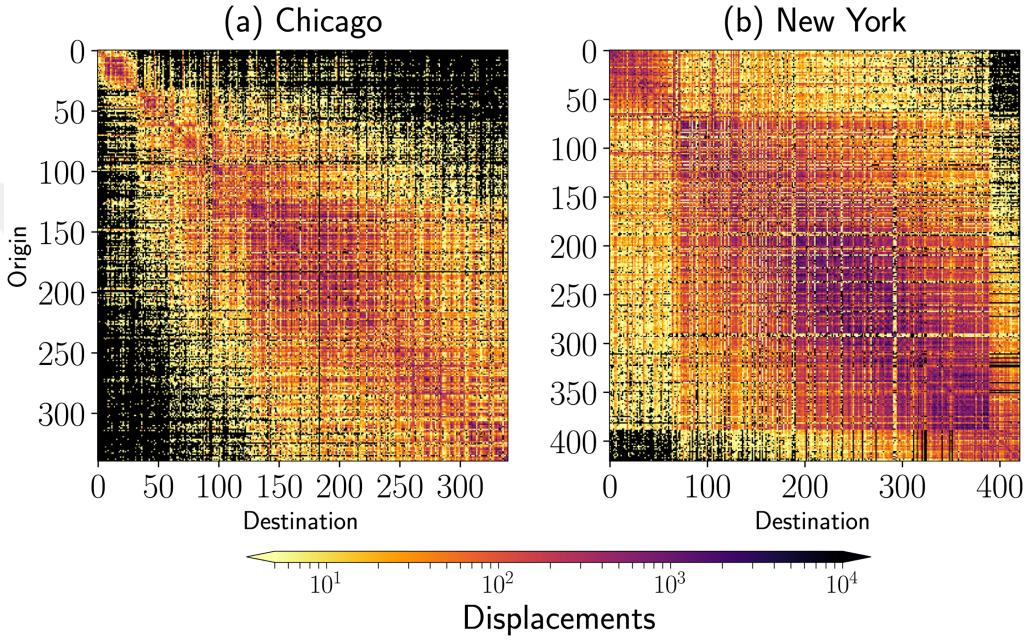

Figure 2.6: Origin-Destination matrix visualization depicting the density of visits in Chicago and New York.
Source: Image obtained from [97].

OD matrices are essential tools for analyzing aggregated mobility patterns in varying urban scales. In most settings, origin and destination locations are derived from administrative regions, such as census tracts or counties. The OD matrix can take the form of a square matrix, especially when the analytical context involves identical origin and destination locations within the studied area. Instead of relying on administrative regions, an urban area can also be partitioned into spatial tessellations, which can then be used to aggregate mobility flows from corresponding locations [138].

2.2.4 Home Location Estimation

In human mobility analysis, the sociodemographic characteristics of sampled individuals serve as crucial elements in understanding their mobility patterns. With

survey-based data, such traits are collected in tandem with their mobility records. However, with alternative data sources, researchers often have to infer or estimate the sociodemographic traits of sampled individuals primarily from their home locations, which can then be used to harvest matching census data. To infer the sampled individual's home location, researchers focus on nighttime stop locations, e.g., 8 p.m. to 6 a.m., and apply a frequency analysis. With CDR and GPS data, it is relatively easy to compute nighttime stay patterns and infer their home location. However, with other data sources, the researchers need to rely on custom algorithms and models to estimate the home locations of individuals. For instance, Mahmud et al. [100] employ an ensemble classifier model to infer Twitter users' home location based on their tweet patterns.

2.3 Mobility Models

A large body of literature focuses on constructing accurate human mobility prediction models, addressing both individual-level and population-level mobility predictions. Such models incorporate the spatio-temporal characteristics of mobility patterns in varying urban scales and timeframes. In this section, both individual-level and population-level mobility models will be discussed in detail.

2.3.1 Models for Individual Mobility Patterns

Individual mobility models are focused on predicting an individual's next stop location given their historical movement trajectories. Research [62, 119] shows that individual mobility patterns are far from being random; individual mobility patterns demonstrate regularity and predictability over a certain spatio-temporal domain.

2.3.1.1 Levy Flights

Random walk is a probabilistic framework for pinpointing an object's next location based on a spatial random displacement variable, ΔX_i, extracted from a displacement distribution $f(\Delta x)$. The location of the object i', after N discrete stops, is determined by $\sum_i^N \Delta X_i$, in which successive identically distributed random displacement variables are aggregated.

A subclass of random walks, named *Levy flights*, is found to be well-suited for modeling animal mobility patterns by [166], in which the displacement distribution is defined as $f(\Delta x) = \Delta x^{-(1+\beta)}$. Levy flights consist of small displacements with sporadic large displacements in between. Brockmann et al. [28] harvest individual mobility trajectories from the circulation of banknotes and mark that the resulting trajectories are approximated by Levy flights with β equal to 0.6.

2.3.1.2 Exploration and Preferential Return

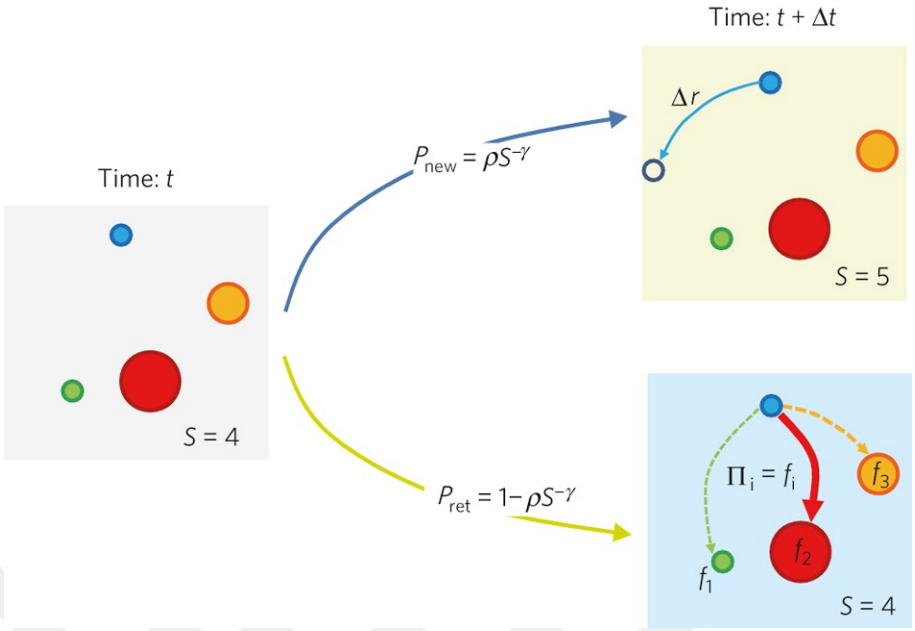

Figure 2.7: Each circle denotes a particular location, in which circle size represents the previous visit frequencies at time t. In the next time step, an individual may choose to visit a new location, i.e., exploration, or may go back to a previously visited location, i.e., preferential return.
Source: Image obtained from [144].

Random walk-based models do not respect an individual's desire to return to previously visited locations or explore new places. Song et al. [144] develop a model that incorporates individuals' exploration and preferential patterns based on their previous visits to different locations. The probability of an individual visiting a new location, i.e., exploration, is defined as

$$P_{new} = \rho S^{-\gamma}, \tag{2.2}$$

in which ρ and γ are model parameters and S is the number of unique locations visited by a random individual. On the other hand, in preferential return, the probability of visiting a previously seen location is denoted as the complementary of exploration.

$$P_{ret} = 1 - \rho S^{-\gamma} \tag{2.3}$$

In this setting, model parameters are bounded as $0 < \rho < 1$ and $\gamma \geq 0$. These two parameters control an individual's predisposition to explore a new location or

visit a previously seen location, given their mobility trajectories. In summary, the probability of visiting location i is computed based on an individual's prior visits to that particular location.

2.3.1.3 Recency Model

The preferential return model gives a greater weight to the locations that have been visited frequently. However, recently visited locations may also appear as a trend considering human mobility tendencies. Based on the preferential return phenomenon, Barbosa et al. [19] develop a model that also considers the recency of a location. They introduce two ranking variables, K_f and K_s, which respectively capture the frequency and recency of a particular location within an individual's trajectory patterns. In this setting, the preferential return is calculated in two folds based on empirical probability value α as

$$\begin{aligned} P^s_{ret} &= (1-\alpha)P_{ret} \\ P^k_{ret} &= \alpha P_{ret}, \end{aligned} \tag{2.4}$$

Depending on the α, the preferential return will either be giving more weight on recently visited locations or act as pure preferential return ($\alpha = 1$). As a result, the probability of visiting a previously seen location i, π_i, becomes $k_s(i)^{-v}$ and $k_f(i)^{-(1+\gamma)}$, when $\alpha < 1$.

2.3.1.4 Evaluation Metrics

Several approaches have been employed by the literature to evaluate the performance of models for individual mobility. An individual's ground truth mobility trajectory, T, consists of stops, l_i, at different time steps $T = \{l_1, l_, \dots l_N\}$ over a certain timeframe. A mobility generator model outputs a trajectory in a similar structure. To understand the performance of a model, one evaluation method is to analyze the statistical properties of the ground truth and generated mobility trajectories, such as displacement distribution, $P(\Delta R)$, radius of gyration, r_g, and mobility entropy which measures the diversity of visited locations by an individual with Shannon entropy. Figure 2.8 displays the comparison of ground truth with generated trajectories based on their statistical properties.

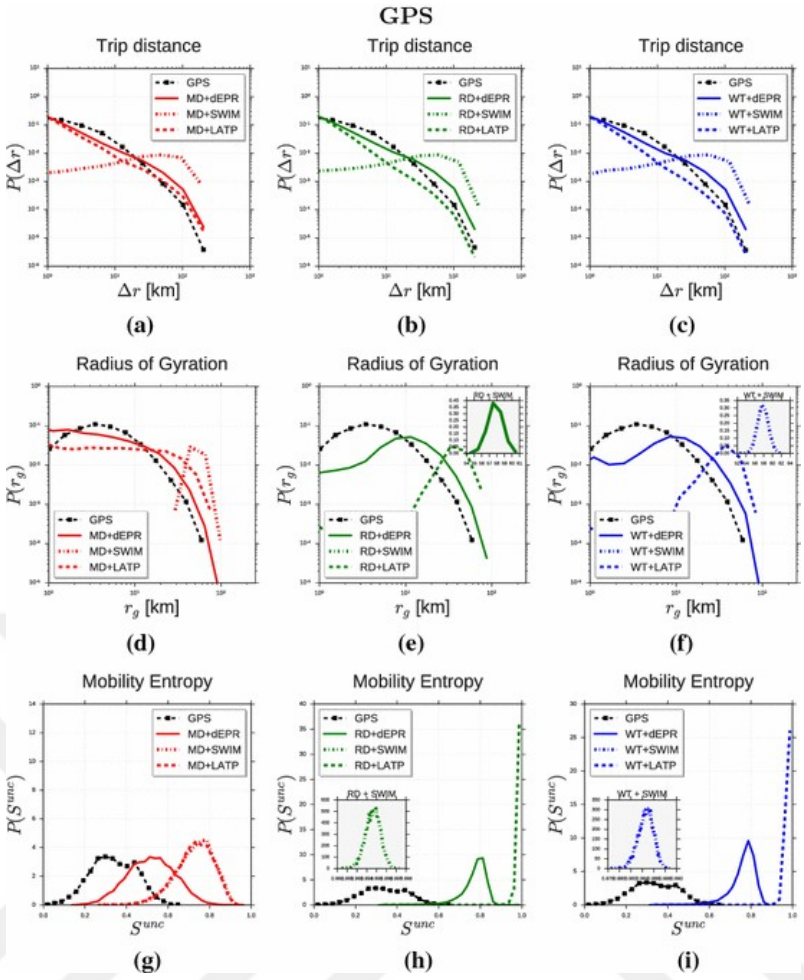

Figure 2.8: Model evaluation based on the distribution of statistical properties of ground truth (black squares) and the generated mobility trajectories by different models.
Source: Image obtained from [118].

To obtain point-estimate evaluation metrics, researchers applied error-based approaches, such as Root Mean Square Error (RMSE), and similarity-based approaches, such as Kullback-Leibler (KL) divergence, for the distribution of statistical properties. Given a ground truth distribution, y, and generated distribution, $\hat{y}$, RMSE accumulates the squared errors between sampled points, y_i and $\hat{y}_i$, as displayed in Equation 2.5.

$$RMSE(y, \hat{y}) = \sqrt{1/N \sum_{1}^{N} (y_i - \hat{y}_i)^2} \tag{2.5}$$

KL divergence, on the other hand, measures the non-symmetric difference between two distributions, which aims to quantify the information loss when using the generated distribution to approximate the ground truth. Equation 2.6 displays the

KL-divergence computation for y and $\hat{y}$.

$$D_{KL}(y, \hat{y}) = \sum_i y(i) \log(\frac{\hat{y}(i)}{y(i)}) \tag{2.6}$$

In order to evaluate the actual trajectories, researchers employ methods that treat a trajectory as a temporal sequence of objects. Dynamic time warping (DTW) is a metric that quantifies the similarity between two sequences considering their alignment cost. Given a ground truth trajectory, T, and a generated trajectory, $\hat{T}$, DTW considers an alignment P as a series of paired random locations from T and $\hat{T}$, i.e., $P = \{(T_{i_1}, \hat{T}_{j_1}), (T_{i_2}, \hat{T}_{j_2}), \ldots, (T_{i_K} \hat{T}_{j_K}), \}$, where K is $min(N, M)$ in which N is the length of the ground truth trajectory and M is the length of the generated trajectory. The cost of constructing such an alignment is defined as

$$cost(P) = \sum_{k=1}^{K} d(T_{i_k}, \hat{T}_{j_k}), \tag{2.7}$$

where d is a distance function, e.g., Euclidean distance. DTW evaluates all possible alignments between the ground truth and the generated trajectories, and finds the alignment with minimum cost as the similarity measure, $\min_P cost(P)$.

2.3.2 Models for Population Mobility Patterns

Aggregated mobility flows between urban areas offer crucial insights for decision-makers to shape data-informed policies across diverse domains like urban planning, public transportation, and access to amenities. Most often, flows are represented as an OD matrix. The aim of a population-level mobility model is to create a synthetic OD matrix that closely mirrors the original data.

2.3.2.1 Gravity Model

Newton's gravity law states that the gravitational attraction between two objects is proportional to their mass and inversely proportional to their distance. Based on Newton's gravity law, the Gravity model computes the mobility flow, F_{ij}, between the populations of two urban areas, and the distance, r_{ij}, between them as displayed in Equation 2.8.

$$F_{ij} = \frac{P_i P_j}{r_{ij}^{\gamma}} \tag{2.8}$$

The parameter γ controls the weight of the distance between urban areas and is empirically estimated from data. For each area in the OD matrix, Equation 2.8 produces the number of trips from corresponding locations.

2.3.2.2 Huff Model

Another model based on the gravitational attraction approach is the Huff gravity model [75], in which the mobility flow is estimated between urban areas and points of interests (POI). Huff model assigns each POI an attraction score based on their utilities, such as store area. In this setting, the distance between urban areas and POIs again negatively affects the estimated mobility flows. Huff model estimates the probability of residents at urban area i visiting POI j as the following

$$P_{ij} = \frac{A_j^{\alpha} / D_{ij}^{\beta}}{\sum_{k=1}^{n} \frac{A_k^{\alpha}}{D_{ik}^{\beta}}}, \tag{2.9}$$

A_j is the attractiveness POI j and D_{ij} is the distance between urban area i and POI j. Model parameters α and β control the weights of attractiveness and distance are empirically estimated based on the data.

2.3.2.3 Radiation Model

The gravity model solely takes the distance between two locations into consideration to account for the mobility flow. The seminal work by Stouffer [150] claims that an individual's choice of mobility from urban area i to j does not solely depend on the distance; rather the number of *intervening opportunities* along the path. In this setting, the definition of an opportunity depends on the social context, such as job selection or nearby amenities. Based on this idea, Simini et al. [139] proposed the Radiation model which considers the opportunities for an individual along the path from urban area i to j. Equation 2.10 displays the computational framework of the radiation model

$$T_{ij} \;=\; T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{2.10}$$

where T_{ij} is the average mobility flow from urban area i to j, m_i and n_j are the populations of each urban area. The distance between i and j, r_{ij}, is used as the radius of the circular region centered on urban area i. Thereby, s_{ij} is the population that resides inside that circular region. And lastly, T_i is the total number of individuals that originate their mobility from urban area i. In contrast to the

gravity-based models, the radiation model is a parameter-free model.

2.3.2.4 Evaluation Metrics

Population-level mobility models generate an OD matrix as their output. The evaluation of these models involves comparing the computed OD matrix with the ground truth OD matrix. The Pearson correlation measures the linear relationship between two data objects. The resulting correlation coefficient, r, describes the relationship between data instances, where a coefficient value of 1 depicts the perfectly positive relationship, while -1 depicts the perfectly negative relationship. A correlation coefficient value of 0 means that the two data objects are not related at all. In population-level model evaluations, Pearson correlation is applied to the rows of the ground truth and the generated OD matrices, and then their average is taken into consideration. Equation 2.11 displays the computation of Pearson correlation of ground truth flow array, F, and the generated flow array, $\hat{F}$.

$$r = \frac{cov(F, \hat{F})}{\sigma(F)\sigma(\hat{F})} \tag{2.11}$$

Another well-established evaluation measurement is the Common Part of Commuters (CPC), which compares the ground truth and the generated OD matrices. Equation 2.12 displays the computational framework of CPC, in which OD_{ij} represents the ground truth flow from urban area i to j, while $\hat{OD}_{ij}$ stands for the generated flow between the same urban areas. CPC returns a score as a non-negative number between 0 and 1, where a score value of 0 indicates bad performance, while a score value of 1 marks perfectly generated flow instances.

$$CPC(OD, \hat{OD}) = \frac{2\sum_{ij} \min(OD_{ij}, \hat{OD}_{ij})}{\sum_{ij} OD_{ij} + \sum_{ij} \hat{OD}_{ij}} \tag{2.12}$$

2.4 Network Approaches

Networks are powerful tools for analyzing the relationship between entities. A network, $G = (N, E)$, consists of nodes, N, representing the entities in a system and the edges, $E = \{\{x, y\} \mid x, y \in N\}$, define the relationship between nodes. A network G can also represented as an OD-matrix-like data structure named adjacency matrix, A, in which rows and columns represent the nodes of the network and the corresponding cells, A_{ij}, stores the connectivity information between node i and j, i.e., $A_{ij} = 1$ if they are connected, else $A_{ij} = 0$.

In human mobility analysis, networks are constructed to capture the mobility patterns between locations in varying urban scales, such as neighborhoods, cities, and countries. In addition, mobility networks can be constructed for both individual-level [71] and population-level [31] analyses.

However, the majority of studies utilizing mobility networks are focused on population-level tasks to reveal the role of locations in urban mobility patterns. In a population-level setting, an OD matrix that records the mobility flows between urban areas constitutes the backbone of the network construction procedure. An OD matrix is employed to create a weighted directed network, in which each edge is assigned with a weight w_{ij} representing the flow from urban area i to j.

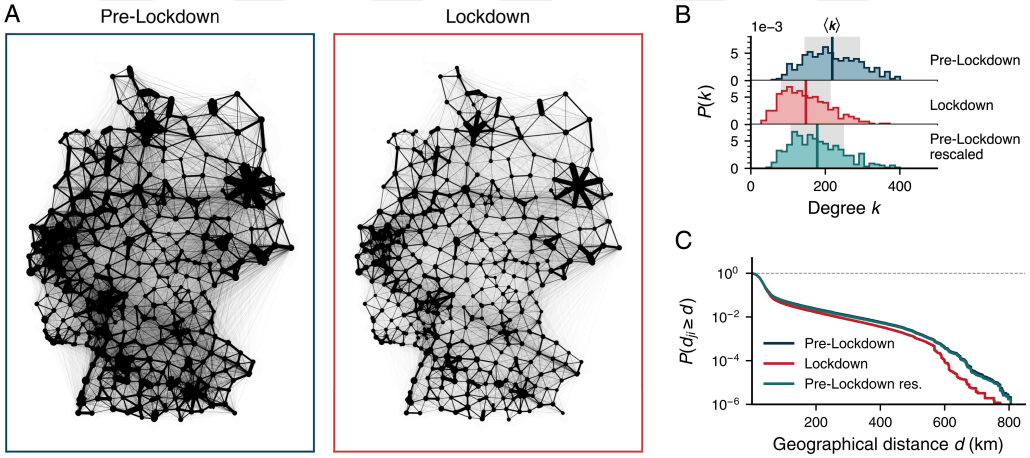

Figure 2.9: The effect of COVID-19 pandemic lockdowns on mobility networks constructed for the counties in Germany. With the onset of the pandemic, the ties between counties are weakened as it is observed through the mobility networks.
Source: Image obtained from [133].

Analyzing the topological network structure with well-established network metrics may reveal invaluable insights regarding mobility patterns. *Degree* of a node v quantifies the number of edges directly connected to v. In a directed network, node degree can be broken down into out-degree, i.e., the number of edges that originate from node v, and in-degree, i.e., the number of edges that end that node v. A high degree of a node indicates its gravity in the resulting mobility network.

The *node strength* is a metric that takes the edge weights into consideration. A node v's strength is defined as the sum of all edge weights, w_{vj}, that are incident to v. Analyzing the node strength may reveal the overall mobility ties. However, node strength alone does not constitute an evaluation metric for assessing a node's centrality in a network since it does not take which nodes are incident to v.

Centrality metrics evaluate a node's topological position and connectivity within a network, assigning a score that reflects its influence, tailored to the specific computational framework used for centrality analysis. *Betweenness* measures a node's centrality based on the frequency of its presence along the shortest paths connecting pairs of nodes within the network. Formally, betweenness computes node v's centrality, B_v, as the following

$$B_v = \sum_{st} \frac{n_{st}^v}{g_{gt}}, \tag{2.13}$$

where n_{st}^v represents the number of shortest paths from node s to t that also traverse target node v, while g_{st} is the total number of shortest paths from s to t. Betweenness can also be applied to directed networks as well, in which shortest paths will be adjusted to account for the direction. Betweenness centrality computation relies on the shortest path counts, which results in a range of values scaled by the number of node pairs. To put the betweenness score into a pre-determined range, betweenness scores are normalized by the total number of node pairs (excluding node v). For undirected graphs, normalization is done by $(N-1)(N-1)$, and for directed graphs, it is $(N-1)(N-1)/2$ so that betweenness scores will be scaled down in range [0, 1]. The betweenness centrality of a node represents its influence in terms of acting as a broker between other nodes in the network.

Closeness is another measure that evaluates a node's centrality based on its average distance to other nodes in the network. Closeness considers a node's centrality in terms of ease of reachability to other nodes in the network. Formally, it is defined as the inverse of the sum of distances so that as the average distance gets smaller a higher closeness score will be obtained. The distance between nodes, $d(s,t)$, is measured based on the path length of the shortest paths between nodes. In this setting, to obtain normalized closeness values, the resulting reciprocal sum is multiplied by $N-1$, where N is the number of nodes in the network.

$$C_v = \frac{N-1}{\sum_s d(s,v)}, \tag{2.14}$$

The node degree treats each connected node equally. However, in most network

settings, nodes possess varying scales of importance, which might not be fully captured by a uniform consideration of connections. *Eigenvector* centrality considers a node's centrality based on the importance of its connected nodes. Eigenvector centrality of node v, x_v, is calculated as

$$x_i = \frac{1}{\lambda} \sum_s A_{s,v}\, x_s, \tag{2.15}$$

where λ is a non-zero constant, $A_{s,v}$ denotes connectivity between target node v and other nodes s, x_s is the eigenvector centrality of v's connected nodes. Equation 2.15 can be rewritten in the matrix form based on the adjacency matrix A.

$$AX = \lambda X \tag{2.16}$$

In this formulation, X is the centrality vector. Since the adjacency matrix consists of non-negative values, there will always be a unique largest eigenvalue λ, which in turn yields the centrality measurements for all the nodes in the network.

Transitivity in networks quantifies the presence of loops, i.e., if node x is connected to y and y is also connected to node z and transitivity implies that node x is connected to z as well. Analyzing the transitivity of mobility networks may reveal the spatial interplay between locations in an urban area. The transitivity of a network is calculated based on *triads* which consist of two edges with a shared node. For instance, in the above example, the nodes x, y, and z constitute a triad centered on y. Moreover, the same set of nodes also constitutes a *triangle*, also named closed triads, since they are all connected. In this setting, the transitivity is the ratio between the number of triangles in the network and the number of triads, i.e., paths of length 2 as displayed in Equation 2.17. In Equation 2.17, the number of triangles is multiplied by 3 to account for the direction of path traversals in the network.

$$C = \frac{\text{(number of triangles) x 3}}{\text{number of triads}} \tag{2.17}$$

The same idea can also be applied to understand the connectedness around a certain node in the network. *Local clustering coefficient*, C_i, measures the likelihood of two neighbors of a node i being connected to each other. To this end, the local clustering coefficient for node i is calculated as the ratio between the number of triangles through node i and the maximum number of triangles that could be constructed through node i. For undirected networks, the local clustering coefficient is

defined as

$$C_i = \frac{2T_i}{deg_i(deg_i - 1)}, \tag{2.18}$$

where T_i is the total number of triangles through node i and deg_i is the degree of the same node. The local clustering coefficient returns a value between 0 and 1, in which 1 corresponds to the case where all the neighbors of node i are connected.

To understand the dynamics within the neighborhood of a node, *ego-networks* stand out as useful analysis tools. An ego-network is extracted from a network with respect to an *ego node*, also named as the focal node, in which the neighborhood consists of the nodes, i.e., *alters*, that are incident to the ego node. Ego-networks can be employed to analyze the spatial context of an urban area. Moreover, Ego-networks can also be used to extract features in modeling analyses.

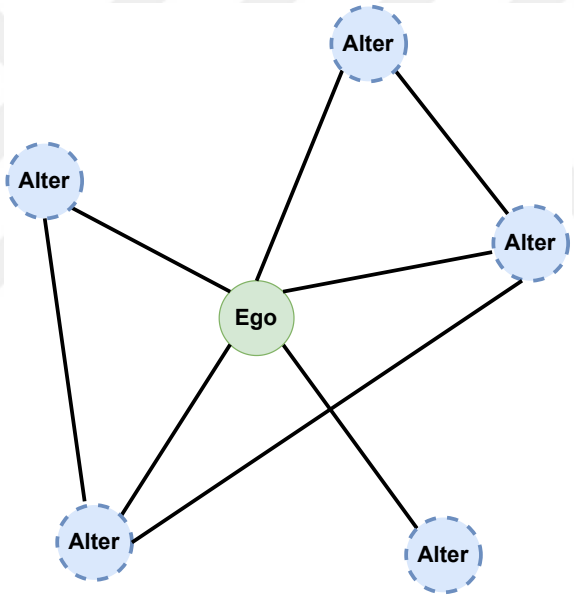

Figure 2.10: An exemplary ego-network, in which ego-node and its alter nodes are highlighted..

Constructing representative feature vectors is a crucial step in predictive analyses. For graph-related predictive settings, such as node-level prediction, Grover and Leskovec [64] proposed the node2vec framework that generates node embeddings in a low-dimensional space that aims to preserve the neighborhood structure of the network in the resulting node embeddings. The proposed framework is based on random walks and has two parameters, the return parameter p, which controls the probability of revisiting a node, and the in-out parameter q, which controls the range of visits, e.g., in case $q < 1$, resulting random walks would have a tendency to visit further away nodes. Node2vec is a well-established framework that is used in the

literature in graph-related predictive tasks.

2.5 Machine Learning Methods

Machine learning methods stand out as essential tools to capture the underlying dynamics, regularities, and patterns of the data to be employed in predicting the unseen data. Machine learning methods are mainly divided into two categories; supervised, unsupervised, and semi-supervised learning [10]. In this dissertation, supervised machine learning methods, in particular classification models, are used to predict the financial performance of businesses.

2.5.1 Supervised Learning

In a supervised learning setting, the data instances are presented with a representative feature vector and its corresponding label. Given training data with N observations, $\{(x_1, y_1), (x_1, y_1), \ldots, (x_N, y_N)\}$, each data point i is represented by its feature vector, x_i, and its label, y_i. The main objective of a supervised machine learning model is to construct a mapping from the input feature space X to the output space Y. To this end, each machine learning algorithm employs a specific approach to learn this mapping based on the provided labeled data, i.e., ground truth.

Depending on the nature of the target output values, supervised models are divided into two groups. In classification tasks, the target label comprises categorical outputs, where each label represents a distinct category or class (binary or multiclass) that characterizes specific groups within the dataset. In case the target outputs are numerical, the task shifts to regression. Regression models are employed to predict continuous values or quantities, aiming to estimate an output based on input features.

Logistic regression is a well-established statistical model that is mainly used for binary classification but can also be extended to multiclass classification settings as well. In logistic regression, a decision boundary is constructed to decide which class a data point belongs to based on the logistic function $p(x) = 1/(1 + e^{-(\beta_0+\beta_1 x)})$. In a multiclass task, the decision boundary can be created in a one-vs-rest setting, in which each class is compared against all other classes individually.

Support Vector Machines (SVM) [41], which is another approach based on decision boundaries, construct a hyperplane (or a set of hyper-planes) to separate the instances belonging to different classes. SVM models aim to find a hyperplane that maximizes the margin, i.e., the distance between the hyperplane and the closest training data points across all classes.

Tree-based models are frequently employed in the literature due to their interpretability and capability to handle nonlinear relationships and interactions within the data. In this dissertation, the random forest model [69], an ensemble of randomly generated trees, is employed. The random forest method creates a series of trees, in which each tree is built with a random subset of the feature set on a random sample of training instances. The final prediction is made based on a combination method, such as voting, that considers the predictions from constructed trees.

Naive bayes is a probabilistic framework that is based on Bayes' theorem. For a given data point, i, and its corresponding feature vector, x_i, naive bayes predicts the class label, k, among the set of labels, C, that yields the highest posterior probability, $P(C_k|x_i)$, that is computed based on the Bayes' theorem as $p(x_i|C_k)p(C_k)/p(x_i)$. Given the target labels, the naive bayes model assumes that the provided input features are conditionally independent, which may not be the case in many real-world problem settings, hence the name naive.

2.5.2 Evaluation Methods

Evaluation is a critical step in any predictive analysis setting to gauge the effectiveness of the utilized models. To have a robust assessment of the developed models, cross-validation is applied, in which the training data is split into k folds. In an iterative manner, the model is trained on $k-1$ folds and evaluated on the remaining fold. This process is carried out until each fold is used for training. The results of each fold are aggregated, such as averaging, and reported.

Confusion matrices, as depicted in Table 2.2 for a binary outcome, are useful tools for understanding the class-level performance of the employed model. In Table 2.2, row identifiers are used to highlight the ground truth and column identifiers are depicting the predicted labels. Each cell of the table gives the number of data points in that predictive setting, e.g., true positives give the number of predictions that are accurately classified as positives while false positives are wrongly classified as positive data points.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2.2: An exemplary confusion matrix representation.

Based on a confusion matrix, several evaluation metrics can be defined.

Accuracy: Fraction of accurately classified data points.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

Precision: Fraction of data points that are correctly classified as positives.

$$\text{precision} = \frac{TP}{TP + FP}$$

Recall: Fraction of actual positives that are classified as positives by the model.

$$\text{recall} = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall.

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

AUC: Area under the received Receiver Operating Characteristic (ROC) curve. ROC displays the relationship between the true positive rate, $TP/(TP + FN)$, and the false positive rate, $FP/(FP+TN)$, at various threshold values, usually depicted as a curve. To have a scalar metric that evaluates a ROC, the area under the curve is calculated. A random model would yield equal true positive and false positive rates for each threshold value, which returns a 0.5 AUC. Models with higher AUC scores perform better.

The evaluation metrics listed above are mainly defined for binary outcomes. However, the same set of evaluation metrics can also extended to multiclass prediction tasks as well. To this end, micro and macro averaging can be employed for

precision and recall. With micro averaging, each equal weight to observations in each class, e.g., total true positives are calculated. On the other hand, with macro averaging, class-level scores are simply averaged. To calculate the AUC in a multiclass setting, classes are binarized to calculate the true positive and false positive rates either in a one-vs-rest or a one-vs-one scheme.

Chapter 3

Related Work

Considering the role of human mobility in an array of different domains, e.g., urban design [68], disaster [177], and resource management [7], the number of human mobility studies have been growing steadily over the past decade. In this chapter, a large body of the existing literature will be discussed from distinct perspectives. First, existing human mobility models will be examined from geographical approaches to deep learning-based frameworks to inform the reader about recent mobility modeling achievements. Next, studies aiming to improve data-driven policymaking based on human mobility patterns will be elucidated, in which the critical role played by empirical data in informing and shaping effective policy frameworks across various domains will be emphasized. And lastly, studies that aim to unveil how human mobility data can be employed to gain economic and business insights will be investigated.

3.1 Modeling Human Mobility

Exploring human behavior has long captivated researchers, with a particular focus on analyzing human mobility patterns. Transportation and urban planning were the main drivers behind human mobility analysis in the 1950s [30], in which survey-based data and estimations based on pre-calculated probabilities constituted the backbone of the conducted analyses. Given the profound significance and far-reaching implications of targeted analyses, the accurate modeling of human mobility stands as a paramount goal for policymakers.

The emergence of enabling technologies, such as GPS and CDR, paved the way for researchers to study human mobility patterns quantitatively. Earlier approaches conceptualized human mobility models in a spatial setting, in which spatial dependence forms the basis of the conducted studies. Gravity model [184], exploration and preferential return [119] and recency model [19] are some of the examples from such spatial models. Spatial statistical properties, such as displacement distribu-

tion ($P(\Delta)$), and visitation patterns constitute the backbone of spatial methods, in which the resulting model aims to capture the actual distribution. Simini et al. [139] developed the radiation model for population-level human mobility analysis the concept of opportunity, a non-spatial specification, is introduced. In Figure 3.1, the opportunity for an individual is considered as job offers from neighboring counties. In contrast to the gravity model, which solely considers the spatial drivers for mobility modeling, the radiation model outputs a greater performance by incorporating the available job opportunities in proximity.

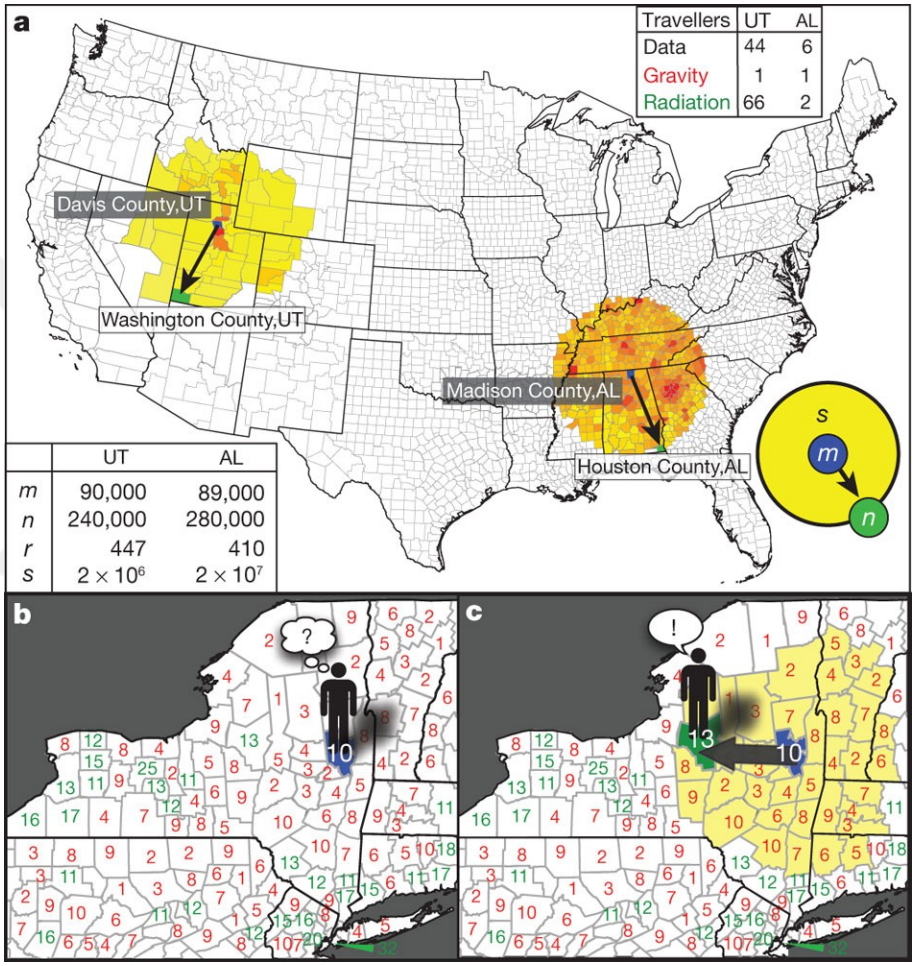

Figure 3.1: **a)** Two pairs of counties from the U.S. with similar inbound and outbound mobility flows. Moreover, the distance between paired counties is also comparable. Since the distance is the same, the gravity model outputs uniform flows between county pairs, contradicting the U.S. Census data. **b-c)** An individual considers the available job opportunities in nearby counties in proportion to the county populations. The numbers of counties represent the job attractiveness. The individual then chooses the closest county with the highest job opportunity.
Source: Image obtained from [139].

Incorporating non-spatial features, such as human behavior and urban features, into mobility models has been a well-established method in the literature [9, 15, 23, 32, 138]. Alis et al. [9] employ a radiation model-based approach to human migration analysis, in which population is not the only indicator for the flow of people between countries. They propose a new indicator, named *the urbanization index*, that captures the urban dynamics of the origin and destination locations as a weighted sum of locality features, $\sum_k = w_k f_k$, where f_k represent an urban feature, such as the number of amenities and population density. The inclusion of non-spatial features yields a higher performance compared to the original radiation model.

Bahrami et al. [15] extend the Huff gravity model by integrating an array of additional features for location attractiveness, such as parking area availability, visitor loyalty, and the diversity of the businesses in the vicinity. Furthermore, they take into account the sociodemographic similarity of the visitors to the broader sociodemographic makeup of the visited location. In their study, each neighborhood is treated as an individual mobility center, for which separate Huff exponents are obtained. Figure 3.2 displays the density plot of the resulting exponents belonging to different parameters.

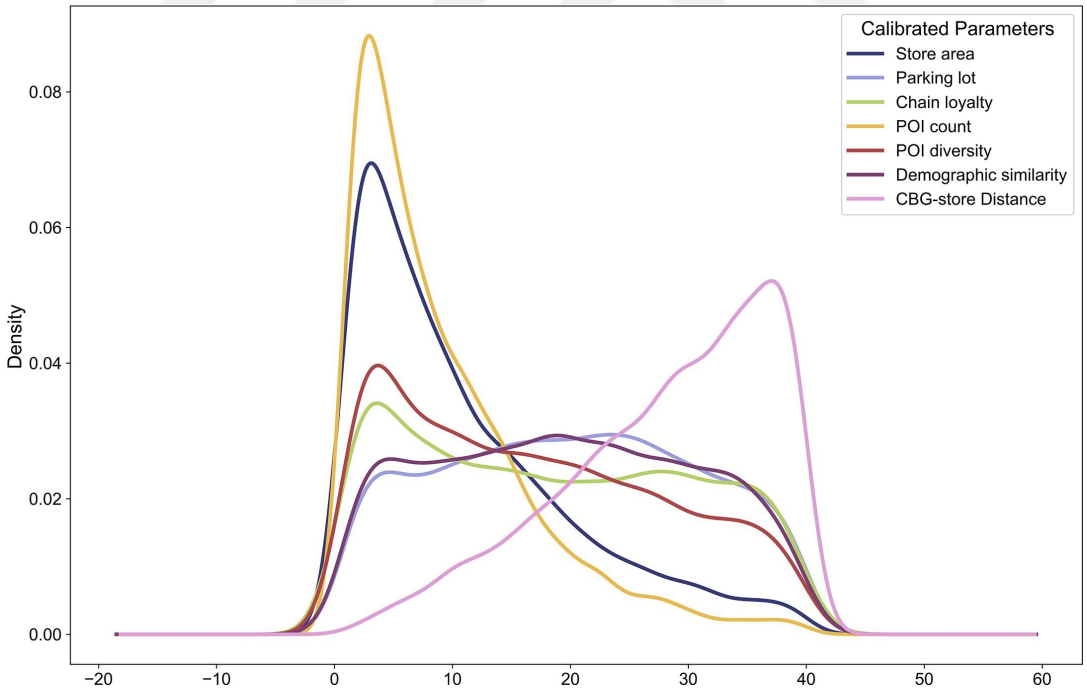

Figure 3.2: Density plot of the calibrated Huff model parameters by Bahrami et al. [15]. In addition to the conventional distance parameter, the attractiveness of a location is modeled as a multitude of non-spatial parameters.
Source: Image obtained from [15].

As large-scale mobility data becomes increasingly accessible and AI experiences notable advancements, researchers are notably drawn to employing deep learning methods for modeling human mobility patterns. Deep learning-based approaches have been utilized for varying mobility tasks, from individual-level trajectory prediction [5, 17, 33, 154, 179] to population-level flow estimation [43, 80, 94, 153, 167].

Simini et al. [138] propose the Deep Gravity model, in which a multitude of spatial and non-spatial features, such as distance, road network, transportation, and amenities in the vicinity, are incorporated as inputs to a feed-forward network. The proposed model expects the merged feature vectors of origin and destination locations, in addition to their distance. In a sense, the model considers every single cell of an OD matrix based on their merged feature vectors and outputs a probability value that accounts for the mobility flow from origin to destination. The proposed model outperforms the traditional gravity model by 66% percent in the common part of commuters (CPC) evaluation metric.

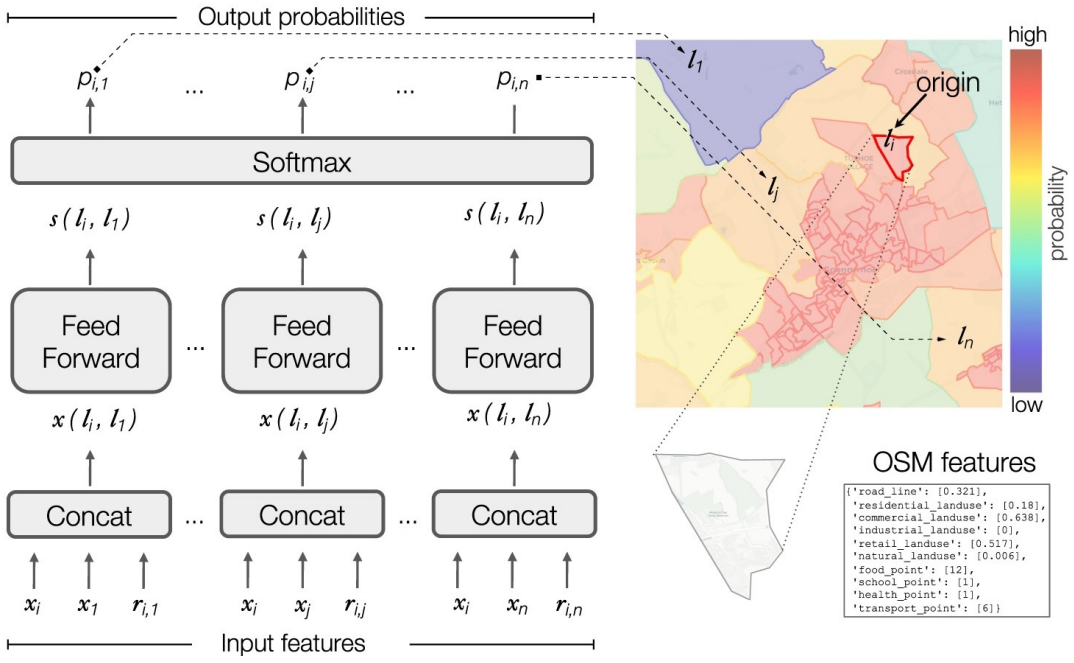

Figure 3.3: Model architecture of the Deep Gravity model [138] for population-level flow estimation task. The feature vectors of origin and destination locations are merged, in addition to the distance between them. The Softmax layer's output serves as the probability assessment for residents originating from location i to visit the destination locations.
Source: Image obtained from [138].

3.2 Economic & Business Insights

Considering the complexity of socioeconomic phenomena, researchers have relied on a multitude of different data sources to gain insight into these complex systems. Human mobility patterns stand out as one of the key economic indicators [79] as it is a product of assorted sociodemographic interactions, economic systems, and available transportation infrastructure. In this section, the role of human mobility analysis in understanding economic systems at varying scales, from individual economic outcomes to global economic development assessment, will be examined.

3.2.1 Individual Welfare & Local Economies

The welfare of societies depends on the interplay of a diverse set of phenomena, such as social cohesion [22], environmental sustainability [125], access to quality education and healthcare [124], and economic stability [50]. Given the nature of human mobility data, wherein movement patterns unveil spatial interactions, researchers have vigorously invested efforts to uncover how this data can be harnessed to understand and explain the intricacies of individual welfare and local economies.

Traditionally, the welfare of individuals is evaluated by socioeconomic indicators, such as demographic traits, savings rate, debt-to-income ratio, and credit score. However, such explicit indicators are recorded by private financial institutions and are most often not allowed to be shared with third-party entities for further development. In order to evaluate the financial performance of individuals based on behavioral patterns, researchers investigated the role of human mobility patterns. Singh et al. [140] employ anonymized credit card transaction data, obtained from a financial institution in an OECD country, to predict an individual's financial performance based on their movement patterns. The utilized data provides credit card transactions with the location of the point-of-sale (POS) device, which is then employed to create spatial mobility metrics, namely *exploration*, *exploitation*, and *plasticity*. The proposed mobility-based feature vectors are then fed into a classification model to predict individuals' categorical financial performance in the next time step. Their results indicate that the proposed mobility-based feature vectors outperform conventional demographics-based feature sets by 30%.

In a similar fashion, Agarwal et al. [6] combine credit card transactions and call records to predict an individual's financial well-being. An array of features devised from communication patterns are enhanced with spatial mobility patterns from credit card transactions, e.g., regularity and diversity. The proposed set of features is then given to a classification model with a binary outcome, an individual having financial distress or not.

In order to delineate the main sociodemographic drivers behind individual-level economic behaviors, Di Clemente et al. [46] analyze credit card transaction data in tandem with CDR data to shed light on customer purchasing clusters in urban areas. To this end, they construct purchase sequences extracted from credit card transaction records and enrich them with spatial mobility metrics obtained from mobility networks that ensue from CDR data. As a result of the conducted analysis, six customer behavior clusters representing different purchasing patterns are obtained, which can then be utilized to understand the individual-level local economic dynamics in an urban area.

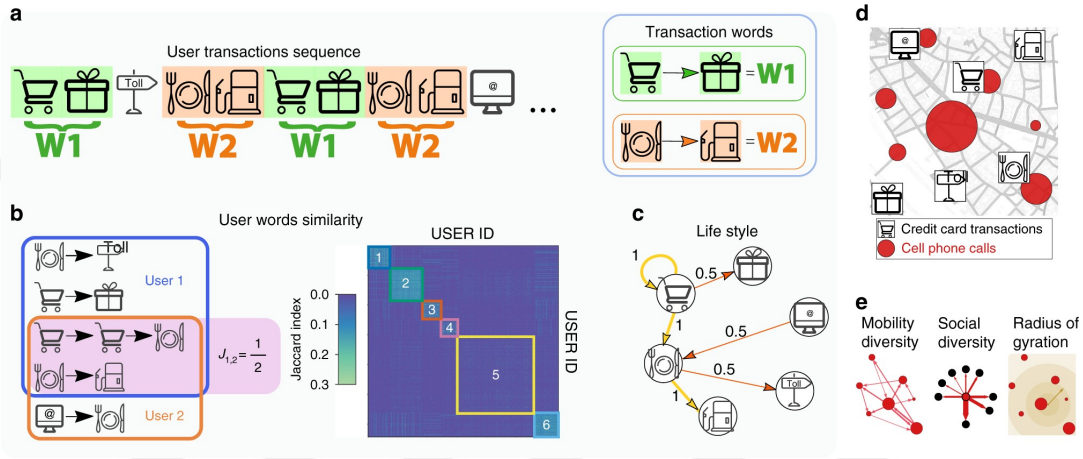

Figure 3.4: Combining credit card transactions with CDR data to shed light on purchasing patterns in urban areas by Di Clemente et al. [46].
Source: Image obtained from [46].

Local economies rely on the well-being of businesses scattered in an urban area. Such businesses frequently rely on loans from financial institutions to ensure the safety of their financial performance [126]. The role of local businesses plays a pivotal role in the well-being of societies. As per the U.S. Small Business Administration (SBA), local businesses employed 61.7 million individuals, accounting for 46.4% of the private sector's total employed workforce in 2022 [161]. Considering their importance for economies, financial institutions require predictive models to assess a business's future financial performance from a holistic view.

In the literature, machine learning-based methodologies have emerged as essential tools for predicting business performance [53, 73, 85, 107]. To this end, the majority of the studies rely on quantitative risk models that consider internal financial metrics such as earning per asset, equity per asset, and debt ratio [20, 35, 37]. Gallucci et al. [60] employed financial metrics, details on bank-firm interactions, and corporate governance variables within a Bayesian model to enhance the accuracy of predicting loan defaults. Kim et al. [85] employed a tree-based majority voting ensemble method to forecast business failure. In addition to financial fea-

tures, they incorporated the recession indicator computed by the National Bureau of Economic Research (NBER) as a macroeconomic feature, providing a holistic view of the overall economic status.

In addition to business-level indicators, researchers have looked for individual-level financial clues as well, for instance, Tang et al. [155] consider the business owner's credit information in quantitative risk modeling. Such approaches aim to overcome the lack of data issue, however, these methodologies still require exclusively collected internal data. The scarcity of available data has prompted researchers to explore alternative datasets and proxies as potential substitutes. For instance, Te [157] employs web mining to harvest proxy business performance features from online outlets, e.g., TripAdvisor and OpenStreetMap, to predict business growth.

To address the impact of local economies on SMEs, Fernandes and Artes [54] introduced a novel variable utilizing ordinary kriging. This variable aims to enhance the assessment of credit default risk among businesses. Yoon and Kwon [180] demonstrated the significant value of credit card transaction data in providing insightful information about the financial status of businesses, in which an SVM model is employed for bankruptcy prediction, including variables like sales fluctuation and patterns derived from credit card transactions as essential predictors.

Recently, human mobility patterns have been employed by researchers for business performance prediction as well. Bahrami et al. [15] study business closure problem, in which given a set of stores belonging to a business (e.g., commercial chains), the task of the decision-maker is to identify the store to be closed. In this setting, they employ an enhanced gravity model to perform iterative simulations based on ground truth mobility patterns in tandem with purchasing records. The proposed method, in the end, identifies the store that would cost the minimum revenue loss.

Maintaining customer retention stands as a critical factor for all types of businesses. Conventional methods rely on customer demographic features, macroeconomic indicators, and other financial information to predict the customer churn problem. Kaya et al. [84] construct spatial features from credit card transaction data that describe a customer's purchasing patterns over the observed time span. In this setting, a customer's purchasing patterns are represented by diversity, loyalty, and regularity indicators that are obtained from credit card transaction data. The proposed feature set outperforms the demographic-based features.

Zhu et al. [183] study the customer churn problem in a similar setting, in which the given individual-level mobility trajectories are utilized to construct spatial mobility metrics, such as diversity and loyalty, in addition to the semantic correspondence of the employed trajectories, e.g., diversity of the visited location. The resulting features are fed into a sequential deep-learning model to predict customer churn in the

next time step.

Exogenous shocks, e.g., natural disasters and pandemics, heavily impact the financial performance of local businesses as well. Estimating the cost of such shocks plays a crucial role in determining overall economic damage. To this end, Yabe et al. [178] employ a counterfactual analytical setting to analyze the visitor distribution of businesses affected by Hurricane Maria in Puerto Rico. Based on large-scale mobility data, the researchers analyze business resilience under exogenous shocks by revealing the impact of the hurricane on local businesses in terms of visit distributions.

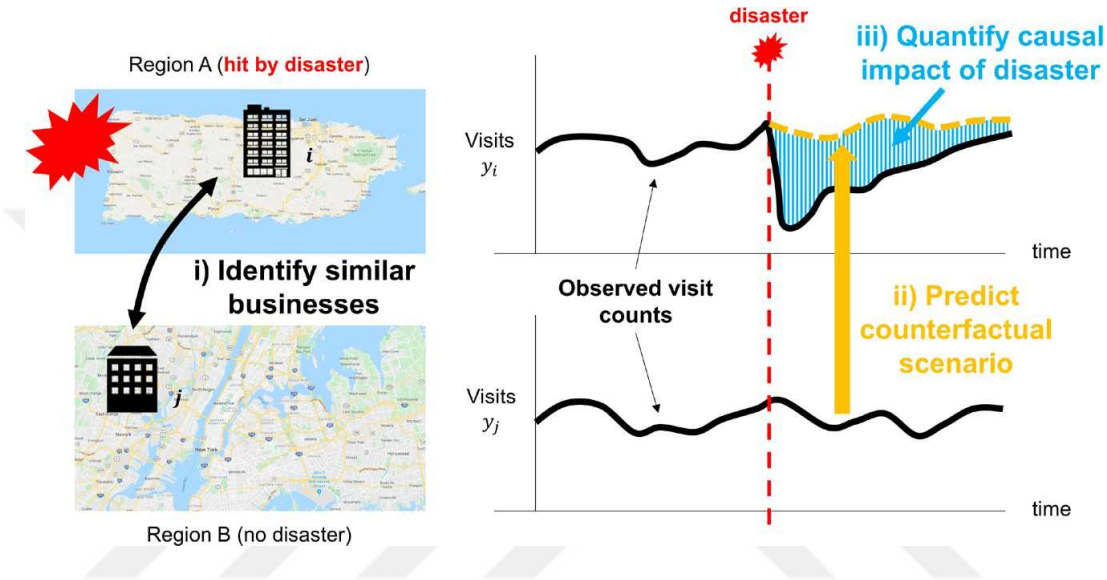

Figure 3.5: Quantifying the impact of a disaster on local businesses by assessing the counterfactual customer visits [178].
Source: Image obtained from [178].

Accurately estimating the customer inflow is a crucial task for new business openings, considering their future financial performance. Conventional approaches rely on on-site surveys to estimate the potential customer foot traffic for new business openings. Large-scale mobility data enables researchers to develop data-driven solutions to tackle this problem. Nie et al. [111] focus on the gas site selection problem and employ GPS mobility data collected by taxi trips. The employed mobility data is used to extract spatial features, in addition to the features obtained from the temporal refueling dynamics.

Liu et al. [96] presents a human mobility-based analytical framework for point-of-interest demand modeling in urban areas, i.e., identifying the type and quantity of point-of-interests that are needed in a certain urban setting. The researchers employ the taxi trips dataset and estimate the trip activity with the help of the Foursquare check-in dataset. And lastly, a latent factor analysis-based model is utilized to infer the point-of-interest demand in the target urban area.

3.2.2 Socioeconomic Development

Understanding socioeconomic development stands as a pivotal task for policy-makers, enabling the design of more effective policies tailored to address the intricate needs and ever-evolving dynamics of urban areas. Existing methods rely on on-site surveys and census polls to unveil the socioeconomic development in urban settings at varying scales. As accessibility to big data sources increases, researchers [58, 93, 117, 174] are increasingly pivoting towards integrating these sources into methodologies for socioeconomic development assessment.

Frias-Martinez et al. [58] employ CDR data and extract a set of spatial metrics to investigate their correlation with socioeconomic development. In this context, the socioeconomic development of urban areas is represented in five ordinal categories. The researchers then match the spatial metrics extracted from CDR data with socioeconomic categories and measure their linear correlation. The results of the conducted study indicate that areas with high socioeconomic value have a positive linear relationship between the radius of gyration, the number of cell towers in the urban area, and the covered area by call records.

In a similar setting, Pappalardo et al. [117] study the socioeconomic development of cities by just looking at the residents' mobility volume and diversity. To this end, socioeconomic development is measured over per capita income, education level, unemployment rate, and a custom indicator named deprivation index, which is a combination of low-quality socioeconomic indicators. Moreover, the mobility patterns are extracted from CDR data and turned into the radius of gyration and diversity of visited locations. The results of the study show that socioeconomic indicators diversity has a higher correlation with socioeconomic indicators in contrast to the radius of gyration as demonstrated in Figure 3.6.

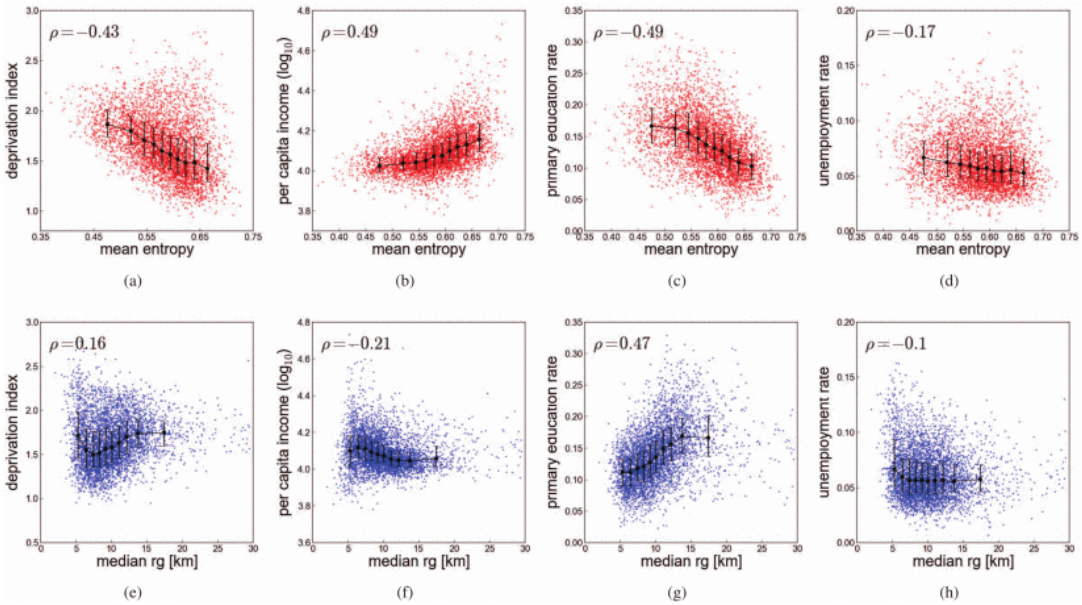

Figure 3.6: The relationship between socioeconomic indicators and spatial mobility metrics, i.e., the radius of gyration and diversity of the visited location measured with Shannon entropy [117].
Source: Image obtained from [117].

3.3 Informed Policymaking

Government and public administration policies wield substantial influence, shaping our lives in profound ways. Policies serve as the means to tackle societal issues by creating and enforcing laws, rules, and recommendations [165]. To enhance the effectiveness and fair treatment of affected subpopulation groups, a data-driven policymaking approach has been influential with the availability of big data sources. In this context, human mobility analysis emerges as a crucial avenue for policymakers, encapsulating a diverse array of intricate spatio-temporal and sociodemographic phenomena. Researchers have explored novel methodologies and indicators aimed at facilitating informed policymaking through the utilization of human mobility patterns.

3.3.1 Disaster Management

During exogenous shocks, such as earthquakes and tsunamis, having a holistic view of human mobility patterns enables a deeper understanding of population movements and aids in orchestrating effective disaster response and recovery strategies. Existing studies in the literature construct models targeting mobility patterns under such exogenous shocks [25, 51, 66, 70, 147] to design better policies regarding the management of such events.

Song et al. [147] focus on the earthquakes in Japan and analyze how mobility

flows took place during and after such a catastrophic incident. To this end, the researchers employ GPS records of approximately 1.5 million individuals over 3 years and construct individual-level mobility trajectories. In the proposed mobility model, a hidden Markov model constitutes the backbone of the overall prediction pipeline, in which the trajectory samples from past disaster periods are used to train the model. The model is evaluated using trajectories from an unseen disaster time frame. The resulting model's performance showcases its robustness in predicting and adapting to unforeseen disaster-induced mobility patterns, reinforcing its reliability for real-time applications during crises.

In addition to the disaster response mobility models, the researchers focused on disaster-related behavioral patterns as well. Han et al. [66] utilize mobility patterns obtained from geo-tagged tweets collected before, during, and after Hurricane Matthew. In the study, the mobility patterns are analyzed considering the flow to the evacuation zones. Their findings suggest that during such disasters, human mobility patterns follow log-normal distributions during the evacuation phase. More significantly, it is noted that humans tend to embrace large cities farther away instead of resorting to smaller but nearby cities.

Mobility patterns are not only used for developing evacuation models but also for evaluating the preparation and protective mechanisms. Li et al. [92] consider residents' preparedness for hurricanes by analyzing the spatio-temporal dynamics of visits to essential point-of-interests. Prior to a hurricane, it is recommended that residents acquire essential medical and emergency supplies. In the conducted study, the researchers analyze residents' visits to pharmacies, grocery stores, gas stations, and home improvement stores. The neighborhoods in the analyzed urban area are then clustered based on the relative change of visits to such essential point-of-interests during the preparation period compared to the baseline time frame. The findings of the study offer promising insights for policymakers, enabling them to prioritize assistance for neighborhoods in dire need.

3.3.2 Urban Segregation

Social cohesion and diversity are essential elements for urban development, especially considering the role played by the diversity of networks in economic growth [49]. However, urban areas are experiencing social segregation with increased urbanization, migration, and industrialization [56]. To this end, researchers have spent significant efforts [52, 103, 104, 110, 175] to unveil the urban segregation dynamics through the lens of mobility patterns.

Moro et al. [104] exploit large-scale mobility data belonging to 4.5 million individuals from 11 cities in the United States and analyze the income segregation in

these urban settings. Income segregation is quantified in two settings, place-level and individual-level, in which observed income quartile distributions are measured. In the place-level analysis, the evenness of the income quartile distribution of visitors is measured. In the individual-level analysis, previously obtained place-level segregation indicators are employed.

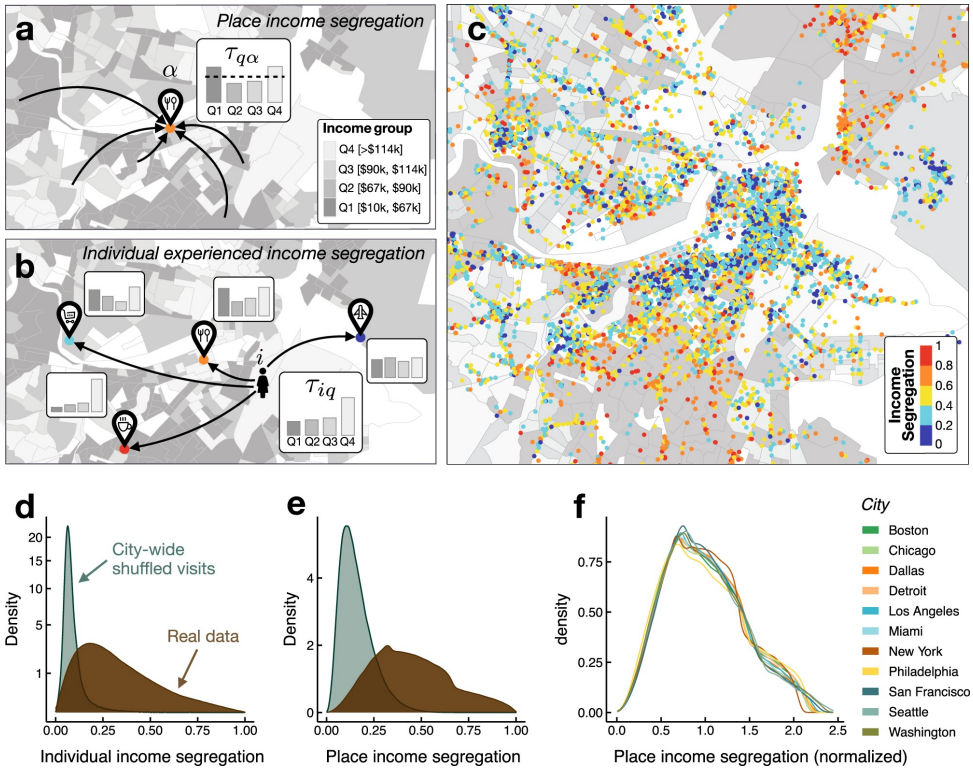

Figure 3.7: Place-level and individual-level income segregation in Boston, MA, measured by leveraging large-scale human mobility patterns. Segregation is gauged by the evenness of visitors' income quartile distribution [104].
Source: Image obtained from [104].

Nilforoshan et al. [110] construct temporal encounter networks between individuals obtained from smartphone mobility data of approximately 9 million users from multiple metropolitan areas with varying urban densities. In this context, economic segregation is measured over the correlation between an individual's socioeconomic status and the mean socioeconomic status of their encounters in a specified time frame. The findings of the conducted study highlight that large cosmopolitan urban settings have more experienced segregation due to the availability of assorted activity outlets, and spaces suited for specific income groups.

3.3.3 COVID-19 Pandemic

The COVID-19 pandemic has triggered profound and diverse changes in multiple aspects of our lives. A significant body of literature exploits human mobility data to enhance epidemiological models [44, 168, 169, 158, 160], in which spatial metrics extracted from mobility patterns are used in modeling the spread of virus. Given the close link between movement and the virus's transmission [173], many enforced Non-Pharmaceutical Interventions (NPIs) aimed at restricting people's mobility. These reduced mobility patterns resulted in socioeconomic consequences that reverberated across varying levels. Researchers have not only analyzed the effect of COVID-19 on populations [31, 59, 86] but also considered the behavioral changes induced by the pandemic [76, 175, 181].

The work by Chang et al. [31] analyzes the overall impact of the pandemic in major cities in the US by constructing dynamic mobility networks between census block groups to points-of-interest coupled with an epidemiological model. The main contribution of the work states that instead of applying uniform business closures, restricting the maximum occupancy for certain points-of-interest, such as restaurants, would curb the spread of infections without affecting local businesses as much as full-scale closures.

Galeazzi et al. [59] construct country-level mobility networks and analyze the national travel patterns during the pandemic in three different European countries. Their findings indicate that a greater mobility reduction took place in long-distance travels while sort-distance connections are observed to be preserved, which is also confirmed by the work of Schlosser et al. [133], in which structural changes of county-level mobility networks are analyzed.

The pandemic has significantly reshaped our behavioral patterns. Yabe et al. [176] conduct an insightful analysis using mobility data from four prominent US cities, spanning a period of three years, and focus on the change of encounters between distinct income diversity groups. The study highlights that due to lowered urban-level exploration trends, the mixed encounters of diverse income groups have decreased significantly.

Hunter et al. [76] employ large-scale smartphone mobility data belonging to approximately 1.5 million users from 10 major cities in the US and analyze the walking behavior of residents in pre-pandemic and pandemic periods concerning two walking behavior classes, recreational and utilitarian. Their results highlight that utilitarian walking patterns have decreased significantly during the pandemic while recreational walking patterns were observed to diminish as well but managed to recover and bounce back as the pandemic unfolded.

Chapter 4

Local Business Performance Prediction with Customer Co-Location Networks

Businesses rely on loans from financial institutions, e.g., banks, to sustain their economic vitality. To this end, these financial institutions require predictive analytical frameworks to evaluate the credit risk of a business. Existing approaches are constructed based on exclusively collected internal financial indicators to assess the financial performance of a business, which are highly private and might not adhere to a certain standard to develop a universal framework. In this study, customer co-location networks constructed over credit card transactions are employed to assess a business's financial performance.

To evaluate a business's financial performance level, a novel evaluation framework based on the relative changes in sales, attractiveness to passers-by, and customer relationships is introduced. This approach can be seamlessly incorporated into the models by financial institutions to determine the risk of financial decisions. Next, based on a novel analytical setting, we propose a social network among businesses in an urban area and construct a network based on a customer co-location network that reflects the shared customer bases. The resulting co-location network is employed to extract a set of network-based features to be fed into predictive models. The proposed set of network features is evaluated considering the conventional revenue and customer information-based business performance features.

The results of the conducted analysis indicate that the performance of the proposed network-based features is comparable to that of well-established revenue-based and customer-based features. This suggests that the proposed co-location network effectively captures businesses' performance levels and yields promising results for future research endeavors. In addition, we argue that the proposed network-based

features possess a crucial characteristic considering the privacy concerns by providing a higher level of safeguarding against attacks to recover the financial attributes of the target businesses, which in turn fosters data sharing between financial institutions and third-party entities for further development and research.

The contributions of the conducted study are as follows.

- A novel approach is proposed to define business financial performance labels in a multi-criteria setting, which is also verified that does not exhibit any biases concerning the home location of businesses, their customers' socioeconomic distribution, and the income distribution of the residents in the home location.
- In a novel setting, a social network among the businesses in an urban area is constructed based on the co-location patterns of the shared customers. The resulting customer co-location network is then employed to extract centrality features, (i.e., node degree, betweenness, closeness, and eigenvector centralities) and diversity indicators within the topological neighborhood of a business, which at the end constitutes the backbone of the proposed network-based features. In addition, we also employ node embeddings to predict business financial performance.
- Based on large-scale credit card transaction data, we show that the proposed network-based and node2vec features perform on par with the conventional financial features. Moreover, the proposed network-based features ensure a higher level of safeguarding concerning the sensitive information of businesses.

4.1 Background

As per data from the U.S. Bureau of Labor Statistics, from 1994 to 2019, approximately 33% of newly established small and medium-sized enterprises experienced failure within their initial two years, with only around 50% surviving beyond their first five years [162]. Given the pivotal role businesses play in maintaining economic vitality, it becomes imperative for governments, public administration offices, and financial institutions to actively support and monitor the short-term performance of businesses.

Additionally, the majority of businesses depend heavily on loans from financial institutions, such as banks, to sustain their economic vitality [126]. Given the significant failure rates among businesses, financial institutions must meticulously evaluate businesses when making decisions regarding loans. These institutions are in search of methodologies that are able to explain and analyze the factors influencing the economic troubles of businesses. As a result, financial well-being and business

performance prediction are crucial for financial institutions as they gauge the risk level concerning a particular loan decision.

In the literature, a significant majority of the business financial performance prediction studies employ exclusively collected private internal metrics, e.g., debt-to-income ratio, which can not be shared with third-party entities. However, businesses often engage in operations constrained by resources, facing challenges due to the scarcity of certified financial statements and publicly available information concerning debt, equity, or liquidity [21, 38, 180]. Furthermore, local businesses are not obliged to publish financial reports in a periodical manner, which in turn makes it difficult for financial institutions to oversee and follow the financial dynamics of the target area.

Considering the competitive landscapes where businesses are run, privacy concerns regarding financial records are highly prevalent. On the other hand, the scarcity or unavailability of businesses' internal information and financial records presents a significant obstacle to developing and enhancing predictive models. Sharing data with third-party entities has long been a delicate matter for financial institutions, in which legally binding yet time-consuming Non-Disclosure Agreements (NDAs) between entities provide some degree of security for the data.

This study aims to explore the evaluation and prediction of business performance while considering the concerns about data privacy and the safety of data sharing. Our goal is to fulfill these objectives without directly accessing the internal financial metrics of the businesses.

Based on social physics [16, 121] and computational social science [90], which employ data-driven methodologies to comprehend human behavior, we adopt a similar approach to analyze business financial performance. We employ large-scale credit card transaction data and machine learning models to predict businesses' future performance without directly using their private internal financial indicators.

Dong et al. [47] show the importance of social interactions in understanding purchasing patterns across communities with diverse backgrounds. We utilize customers between businesses as bridges to be considered as indicators for financial performance prediction.

Research in social physics shows that interactions and knowledge exchange through networks can enhance productivity [171, 63, 122, 127, 36, 11]. Networks constitute the backbone of social and economic functions, shaping the fabric of societies [49], which provides a framework to analyze the dynamics of social and economic interactions in an urban area, offering a means of analysis of business financial performance.

In our interconnected social fabric, businesses thrive within a network. We propose the establishment of a business network based on credit card transaction data, operating on the basis that customers can serve as connectors between businesses. To

this end, we employ well-established centrality metrics as indicators of businesses' performance within the network. We argue that the proposed network features metrics offer comprehensive signals for improved understanding and prediction of business financial performance.

4.2 Methods

4.2.1 Data and Preprocessing

The credit card transaction data, spanning a single year from July 2014 to July 2015, from a well-established bank in an OECD country [4] constitutes the backbone of the conducted study. The data provides two sets of information on transactions, customer demographics, and purchases.

In the customer data, each customer is identified by their anonymized ID. In addition, the data provides their demographic traits, such as age, income, education level, marital status, and home location identified with their district ID. In the target urban area, districts have an average area of 150 square kilometers with an average population of 380K residents.

The table contains anonymous business ID, business ISO 18245 category code (MCC) [3], and business district ID. Businesses may belong to various categories such as grocery stores and supermarkets, restaurants, gas stations, clothing stores, and bookstores, which are described by their MCCs.

The transaction data firstly provide business-related information, anonymous business ID, business ISO 18245 category code (MMC) [3], and their home location. In addition, the purchase-related features provide the purchase timestamp and the amount, which captures 4,507 businesses, 62,194 customers, and 2,511,527 transactions. Table 4.1 presents a detailed view of the employed credit card data.

Attribute	**Value**
Timeframe	12 months
Number of customers	62,194
Number of businesses	4,507
Number of transactions	2,511,527
Average transactions per customer	40.38
Average transactions per business	557.25

Table 4.1: The utilized credit card data, comprising 2,511,527 credit card transactions, are presented. Each transaction is recorded as a row and includes the customer ID, transaction amount, transaction date, business ID, and the business category.

As our objective is to determine the future performance levels for businesses, we eliminate non-discretionary MCCs, including government-owned establishments,

parking lots, lodging facilities, and similar categories. Subsequently, to enhance the robustness of our analyses, we exclude businesses with fewer than 10 transactions per month during each period. In addition, only the MCCs with more than 10% of the filtered businesses are considered and the districts with less than 10 businesses are removed, which results in 1,977 businesses in three different business categories as demonstrated in Table 4.2.

Description	**MCC**	**Business Count**
Grocery Stores, Supermarkets	5411	1,118
Men's and Women's Clothing Stores	5691	464
Service Stations (with or without ancillary services)	5541	395

Table 4.2: The resulting business categories and their business counts after the preprocessing.

Table A.1 presents a comprehensive list of business categories along with the corresponding transaction and business counts.

4.2.2 Approach

4.2.2.1 Customer Co-Location Networks

Networks representing a specific phenomenon demonstrate inherent capabilities for inferring the future state of entities, e.g., nodes. Previous research efforts have demonstrated that the structural properties of nodes in a network can encode the potential for improvement in future time steps [26, 27, 123]. Drawing inspiration from these findings, we define and construct a co-location network using the transaction data.

Let $v_i \in V$ represent a vertex for business i (shown by m_i) and $e_{ij} \in E$ is an edge connecting businesses v_i and v_j. A co-location network is an undirected graph $G = (V, E, W)$, where the weight $w_{ij} \in W$ is defined as:

$$w_{ij} = |\{c_k | \exists(c_k, m_i) \in \mathcal{D} \wedge \exists(c_k, m_j) \in \mathcal{D}, c_k \in \mathcal{C}\}| \tag{4.1}$$

where $\mathcal{D}$ denotes the transaction data and $\mathcal{C}$ is the set of customers. The edge weight w_{ij} captures the number of unique customers who visited and made transations at both businesses i and j. A_i denotes the ego-network of business i that captures the topological neighborhood of the business in graph G. Figure 4.1 illustrates the process of the creation of a co-location network from credit card transactions.

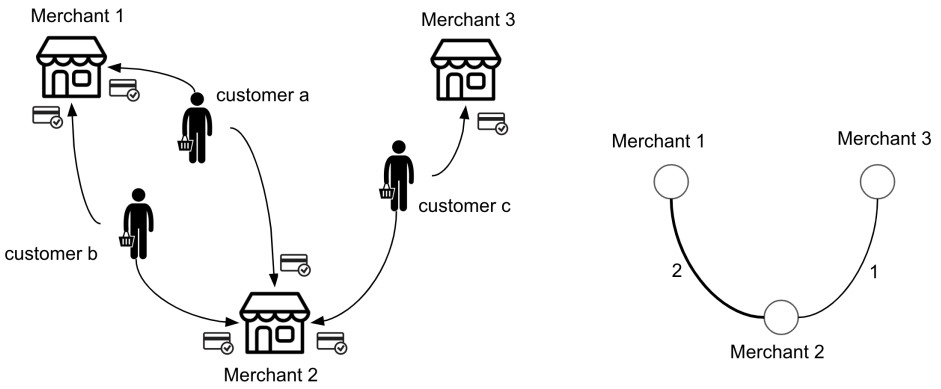

Figure 4.1: The proposed customer co-location network, which is constructed between businesses based on their shared customer bases.

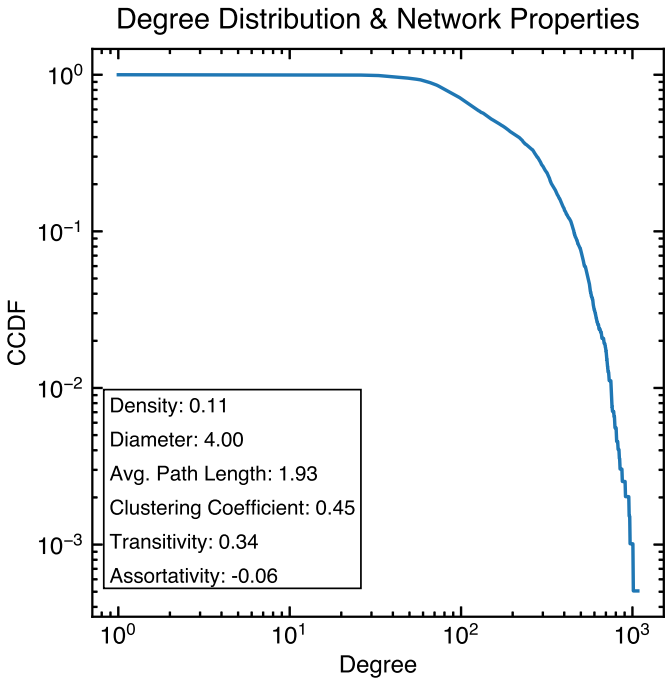

Figure 4.2: Complementary cumulative distribution function (CCDF) of resulting degree distribution and the descriptive statistics of the resulting co-location network.

4.2.2.2 Network-Based Features

In the proposed framework, in order to predict the future financial performance of a business, we extract network-based features from the constructed customer co-location network. To this end, four well-established centrality measures, i.e., degree and strength, betweenness, closeness, and eigenvector centrality, and two diversity metrics constitute the proposed features. Network centrality metrics are able to reflect the significance of a node based on its topological position in a network [163]. In what follows, the employed centrality metrics are introduced, in addition to the

two diversity metrics aiming to explain a business's ability to attract customers.

Degree and Strength considers the connections of a node and quantifies its centrality based on these connections. The degree of a node is obtained by calculating the total number of edges a node has. In parallel, node strength focuses on the edges as well while accounting for the edge weights. The node strength is computed as the sum of edge weights incident to the target node. In the context of co-location networks, node strength is identical to the number of unique customers shared between two businesses. A business's degree and strength values are able to act as proxy measures considering its revenue.

Betweenness centrality measures the importance of a node within a network based on the number of shortest paths that pass through it. In other words, the betweenness of a node reflects the interconnectivity of the node's neighboring nodes, assigning greater significance to nodes that act as bridges. In the context of co-location networks, betweenness reflects the number of businesses indirectly linked to a shop through their direct connections.

Eigenvector centrality the significance of a node while factoring in the significance of its neighboring nodes [61]. In the proposed customer co-location networks, the eigenvector centrality approximately represents the likelihood that a random customer will patronize a specific merchant.

Closeness centrality measures how closely connected a node is to all other nodes within a network. In the context of customer co-location networks, closeness centrality assesses a business's ability to access information from other businesses in its network through its customers.

In order to reflect the local dynamics of a business's interactions with distinct geographies and business categories, we introduce two diversity metrics based on a business's ego-network.

Geographical diversity quantifies the geographic diversity within the ego-network of the target business i, which is denoted as D_g^i. In this context, the district ID is employed to calculate the geographic distribution of a business's topological neighbors. Shannon entropy [134] is utilized in Equation 4.2, which is a widely employed diversity metric.

$$D_g^i = \sum_{h \in I_H^i} -p_h^i \log p_h^i \tag{4.2}$$

p_h^i represents the proportion of the edge weights of business i linked to other businesses from district h, relative to all districts that share edges with business i, denoted by I_H^i.

Business-category diversity quantifies the diversity of business categories among

the topological neighbors of business i. The business-category diversity, D_c^i, of a business is obtained with respect to its ego-network alters with Equation 4.3.

$$D_c^i = \sum_{b \in I_B^i} -p_b^i \log p_b^i \tag{4.3}$$

p_b^i represents the proportion of the edge weights of business i linked to other businesses from business type b, relative to all business types that have edges with business i, denoted by I_B^i.

4.2.2.3 Financial Performance Label Definition

The primary goal is to devise a financial performance metric that can accurately assess a business's performance within a competitive context. Existing approaches in the literature have utilized a variety of financial and subjective indicators to propose diverse performance definitions for businesses [91]. Nonetheless, employing insights considering a business's internal objectives might not always be achievable. Even if achievable, there may be no trivial method to render such cues into a metric that reflects the business's market position and competitiveness. Hence, we utilize three concrete metrics, including business sales, attractiveness to potential customers, and customer relationship data, to formulate a novel financial performance metric.

Given the spatio-temporal dynamics of the employed credit card transaction data, as well as the findings from preliminary observations, we opt for a duration of 6 months, dividing the data into two equal periods. Next, we compare a business's revenue, number of unique customers, and number of transactions over the first 6 months with those in the consequent 6 months. Then, the rate of change for each metric is computed using Equations 4.4 through 4.6.

$$\Delta R_{t+1,t}^i = (R_{t+1}^i - R_t^i)/R_t^i \tag{4.4}$$

$$\Delta C_{t+1,t}^i = (C_{t+1}^i - C_t^i)/C_t^i \tag{4.5}$$

$$\Delta N_{t+1,t}^i = (N_{t+1}^i - N_t^i)/N_t^i \tag{4.6}$$

R_t^i, C_t^i, and N_t^i represent the revenue, number of unique customers, and number of transactions of business i in period t (the first 6 months), respectively. Similarly, R_{t+1}^i, C_{t+1}^i, and N_{t+1}^i denote the revenue, number of unique customers, and number

of transactions of business i in period $t+1$ (the remaining 6 months). It's crucial to note that the revenue of a business is obtained by aggregating the transaction amounts that took place at the business during the observed period. Subsequently, the rates of change derived from Equations 4.4 through 4.6 for business i are compared with the median rate of change of the same indicators across all businesses based on their business category (MCC). By comparing the rate of change in these indicators among businesses with respect to their MCC, the seasonality issue can be mitigated, which would have posed a significant threat in case the utilized transaction data fails to cover the full extent of such effects.

Next, each business is assigned a binary labeled based on whether their rates of change are above or below the median using Equations 4.7 through 4.9. The binary labeling approach is a standard practice in the literature [101].

$$I_R^i(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta R_{t+1,t}^i \ \geq \ median(\Delta R_{t+1,t}^{b_i}) \\ 0 & otherwise \end{cases} \tag{4.7}$$

$$I_C^i(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta C_{t+1,t}^i \ \geq \ median(\Delta C_{t+1,t}^{b_i}) \\ 0 & otherwise \end{cases} \tag{4.8}$$

$$I_N^i(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta N_{t+1,t}^i \ \geq \ median(\Delta N_{t+1,t}^{b_i}) \\ 0 & otherwise \end{cases} \tag{4.9}$$

The resulting binary labels are then aggregated as displayed in Equation 4.10 for each business. The aggregated labels are employed to assign the final financial performance label based on their values. If the aggregated value is equal to 3, meaning that the business performs better across all three indicators compared to half of the other businesses, the business is considered as *well-performing*. Such businesses can be considered low-risk investments since they have outperformed at least half of their counterparts. Conversely, if the aggregated value is equal to 0, it signifies that the business is below the median rate of change in all three indicators in the same business category. Such businesses are labeled as *poorly-performing*. The remaining businesses are classified as *medium-peforming* with moderate risk levels. In total, three performance labels are proposed as demonstrated in Equation 4.11.

$$I_S^i(t+1) = I_R^i(t+1) + I_C^i(t+1) + I_N^i(t+1) \tag{4.10}$$

(4.11)

$$L^i(t+1) \rightarrow \begin{cases} \text{'well-performing'} & if \quad I_S^i(t+1) = 3 \\ \text{'medium-performing'} & if \quad I_S^i(t+1) = 2 \text{ or } I_S^i(t+1) = 1 \\ \text{'poorly-performing'} & if \quad I_S^i(t+1) = 0 \end{cases}$$

4.2.2.4 Analytical Setting

In this research, we employ machine learning methodologies to assess the efficiency of the proposed network-based features obtained from the customer co-location network. Considering the supervised learning approach, it is necessary to generate input characteristics and establish target labels.

In our study, we employ two well-established baseline feature sets from the literature, namely revenue-based and customer-based features. Prior work by Yoon and Kwon [180] has showcased the utility of revenue-based features as a surrogate for internal financial data in predicting business failure. Additionally, research such as that by Anderson et al. [12] and Simester et al. [137] has employed customer-based features to predict the success of new products.

On top of the employed credit card transaction data, the socioeconomic characteristics of the customers and the home location of the businesses are utilized as well. Based on the home location of a business, the sociodemographic characteristics, e.g., population, and average household income, are extracted. Prior research [36, 83, 109] has illustrated how the quantity and variety of points-of-interest (POIs) and amenities in a business's vicinity can enhance its appeal to transient customers and augment its market potential. Based on such findings, a POI dataset from *here.com*[1], a company specializing in digital map production, is utilized. The obtained POIs are classified into twelve categories, such as transportation hubs, hospitals, and educational institutes.

The geographic neighborhood of the businesses from the credit card data and the retrieved POI database are combined to analyze the diversity of POIs located within the 200-meter radius of a target business. Next, the quantity and business category diversity of the POIs inside the extracted area are computed. The resulting features constitute the business-based features, which are depicted in Table 4.3 and are used as fixed inputs in all the employed predictive models.

[1] `https://www.here.com/`

The features based on customers are extracted from information pertaining to individuals engaging in transactions with the businesses provided by the credit card transaction data. While certain businesses focus on a particular customer demographic group, others appeal to a diverse range of demographics. The incorporation of these features is intended to consider the allure of each business across various customer segments. These features enable us to assess whether the businesses of interest effectively draw their intended customers from various demographic cohorts. Table 4.3 presents the list of customer-based features.

Yoon et al. [180] employed revenue-based features on bankruptcy prediction, in which the sale, revenue, and customer profitability indicators are utilized. In this context, we rely on revenue-based features outlined in Table 4.3

The primary contribution of this study lies in the Network features. Based on the first six months, the customer co-location network is employed to extract network-based features. Moreover, we derive these features outlined in Table 4.3, encompassing four centrality metrics and two diversity metrics based on a node's placement within the network framework.

Finally, as a privacy-enhanced alternative, we focus on low-dimensional node representations using the proposed network structure and topology. This is achieved by employing the node2vec algorithm [64], which produces a feature vector consisting of 128 dimensions, with the return parameter p and in-out parameter q set to 1.3 and 1.2, respectively. While it is feasible to adjust the hyperparameters of the node2vec algorithm [65], instead of delving into and implementing optimization approaches, we adhere to a common method of comparing different dimension choices, e.g., 50, 100, 128, 200, and 300, in the downstream task of predicting businesses' financial performance.

All feature sets are solely derived from credit card transaction data occurring within the first six months, i.e., the observation period, which ensures that no data from the prediction period influences the predictive models, hence preserving integrity.

Feature sets			
Business	Customer	Revenue	Network
- Business MCC - District population - District's average per month household income - Surrounding POI count - Category diversity of surrounding POIs	- Number of unique customers - Age mean - Income mean and median - Education diversity - Gender diversity - Marital status diversity - Employment type diversity - Home district diversity - Work district diversity - Distinct home district count - Distinct work district count	- Period total revenue - Number of transactions - Number of unique customers	- Degree - Closeness centrality - Eigenvector centrality - Betweenness centrality - Ego network geo diversity - Ego network MCC diversity

Table 4.3: The list of feature sets employed in the predictive models.

Considering the predictive analyses, we employ four machine learning models [67] based on varying assumptions and approaches, namely: Multi-class Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB), which are common options for both linear and non-linear models, and they are prominent in a wide array of domains. Their widespread adoption is evident in their extensive usage across the literature, including business failure prediction studies.

The ten-fold cross-validation method is employed to assess classification performance, reporting the area under the receiver operating characteristic curve (AU-ROC) [67] as our primary evaluation metric. AU-ROC measures the ability of classification models to differentiate distinct classes. In the conducted study, given the absence of significant label imbalance, AU-ROC serves as an appropriate evaluation criterion. To accommodate multi-class labels, we utilize the one-vs-rest (OVR) evaluation approach, wherein AU-ROC is computed for each class against the rest, and the resultant scores are averaged. Throughout the cross-validation, AU-ROC is computed on the test fold, yielding mean and standard deviation values across the iterations.

4.3 Results

4.3.1 Label Analysis

Out of the 1,977 examined businesses, 590 (29.84%) are categorized as well-performing, 818 (41.37%) as medium-performing, and 569 (28.78%) as poorly-performing. To validate the effectiveness of these labels in predicting business financial performance and to explore the impact of business location and customers' sociodemographic attributes, such as income, on the proposed labels, further analyses are undertaken.

4.3.1.1 Label Indication

To examine the effectiveness of the proposed labels in distinguishing between well-performing and poorly-performing businesses and to gain insights into businesses' longer-term performance, we analyze businesses with similar levels of revenue, customer counts, and transaction counts in the first six-month period but differing labels, i.e., poorly-performing vs. well-performing, considering their performance in the subsequent period.

To accomplish this, we initially categorize the revenue, transaction count, and number of unique customers of businesses during the initial six months into quartiles. A random subset of the structured data is represented in Table 4.5. Subsequently, we select pairs of businesses with identical MCCs and identical quartiles of revenue, transaction count, and number of unique customers.

ID^m	$L^m(t+1)$	$Q_R^m(t)$	$Q_C^m(t)$	$Q_N^m(t)$
119811014	well-performing	Q4	Q3	Q3
119811011	poorly-performing	Q4	Q3	Q3
119811067	medium-performing	Q2	Q1	Q1
119811051	medium-performing	Q2	Q1	Q1
119811017	well-performing	Q1	Q3	Q3
119811002	poorly-performing	Q1	Q3	Q3
119811051	well-performing	Q1	Q2	Q3
119811018	poorly-performing	Q1	Q2	Q3

Table 4.4: Table illustrating an exemplary business's quartiles for businesses' revenue, transaction count, and unique customer count.

Table 4.5:

In this setting, only the pairs with opposite labels, e.g., poorly-performing vs well-performing, are retained based on the performance indicators in the following six months. The retained pair of businesses must belong to the same quartiles of business revenue, transaction count, and unique customer counts, and while one business is labeled as well-performing, the other business must be assigned a poorly-performing label. In the employed credit card transaction data, there exist 11,813 business pairs that adhere to the mentioned criteria. An exemplary instance from the resulting data table is depicted in Table 4.6, which utilizes the information provided in Table 4.5.

Subsequently, by employing the OLS method, we compute three fitted line slopes for each business, considering their revenue, transaction count, and number of distinct customers as dependent variables and months as the independent variable. Analyzing the resulting slopes offers a comprehensive indication of a business's financial performance regarding each variable over a span of a single year. Equation

ID^{m1}	ID^{m2}	$L^{m1}(t+1)$	$L^{m2}(t+1)$	$Q_R^{m1,m2}(t)$	$Q_C^{m1,m2}(t)$	$Q_N^{m1,m2}(t)$
119811014	119811011	well-performing	poorly-performing	Q4	Q3	Q3
119811017	119811002	well-performing	poorly-performing	Q1	Q3	Q3
119811051	119811018	well-performing	poorly-performing	Q1	Q2	Q3

Table 4.6: Business pairs exhibiting identical quartiles for revenue, transaction count, and unique customer count, yet featuring opposing performance labels.

4.12 presents the closed-form notation of the employed model, in which β_1 represents the value utilized as the line slope.

$$
\begin{aligned}
&Y = \beta_0 + \beta_1 X + \epsilon \\
&Y = (Y_1, ..., Y_{12})^\intercal,\ X = (1, ..., 12)^\intercal,\ \text{and}\ \epsilon = (\epsilon_1, ..., \epsilon_{12})^\intercal
\end{aligned}
\tag{4.12}
$$

Figure 4.3 displays time series plots for three scenarios outlined in Table 4.6. Sub-figures, i.e., 4.3a, 4.3b, and 4.3c, portrays time series of the business pairs' revenue trends, i.e., $4.3a_R$, $4.3b_R$, and $4.3c_R$, transaction counts, i.e., $4.3a_N$, $4.3b_N$, and $4.3c_N$, and the number of unique customers, i.e., $4.3a_C$, $4.3b_C$, and $4.3c_C$, over a span of a single year, which is divided into two six months long time frames by the blue vertical lines. Business pairs are drawn from identical quartiles of performance indicators in the initial period, potentially displaying similar trends during that period. Conversely, the analyzed pairs possess opposite labels in the following time period. Figure 4.3 demonstrates that during the first six months, business pairs' performances are on par independent of their financial performance label. However, in the second six months, a pronounced shift takes places and their performances begin to set apart.

In the subsequent step, we compare the β_1 values for each pair of merchants. For each slope pair, if the merchant labeled as well-performing has a higher slope, we assign 1 to the slope indicator and otherwise assign 0, and then sum the three resulting indicators. If the well-performing merchant has higher β_1 values in all three indicators, the sum will be equal to 3. Conversely, if the well-performing merchant has lower values in all three indicators, the sum will be equal to 0. The closed-form expressions of these calculations are presented in Equations 4.13 to 4.16.

Next, the resulting β_1 values of each business pair are compared. Considering each pair, the slope indicator is assigned a value of 1 if the well-performing business has a higher slope; otherwise, a value of 0 is assigned. We then aggregate the three slope indicators. The resulting aggregation will be a value of 3 if the well-performing business exhibits higher β_1 values across all three indicators. In the opposite case, a

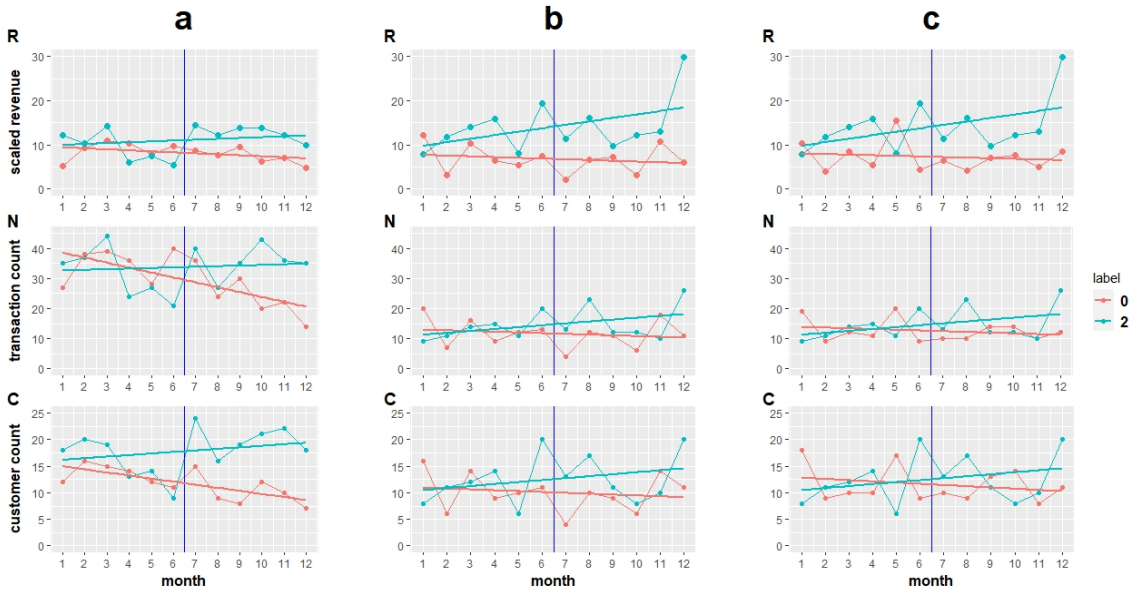

Figure 4.3: Three time series representing distinct business pairs that share identical quartiles of revenue, transaction count, and unique customer count in the initial period but possess opposite labels. Within each letter-tagged subplot, the time series plots depict revenue, top row (R), monthly transaction count, middle row (N), and monthly distinct customer count, bottom row (C), accompanied by the fitted line for each business. In the figure, the color red designates poorly-performing businesses, while the color blue signifies well-performing businesses.

value of 0 will be obtained. The closed-form expressions are highlighted in Equations 4.13 through 4.16.

$$I^R_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{R_{well-performing}} \geq \beta_1^{R_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{4.13}$$

$$I^C_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{C_{well-performing}} \geq \beta_1^{C_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{4.14}$$

$$I^N_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{N_{well-performing}} \geq \beta_1^{N_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{4.15}$$

$$I^S_{\beta_1} = I^R_{\beta_1} + I^C_{\beta_1} + I^N_{\beta_1} \tag{4.16}$$

Table 4.7 presents the findings of this investigation. It is clear that in over

90% of the business pairs, the well-performing businesses exhibit better long-term performances over all three metrics. As a result, the outcomes of the conducted analysis validate the reliability and veracity of the labeling method in discriminating businesses based on their performances as assessed by three objective measures.

Sum Indicator ($I^S_{\beta_1}$)	Business pair count	Business pair percentage
0	160	1.35%
1	239	2.02%
2	673	5.69%
3	10,741	90.92%

Table 4.7: Number and percentage of business pairs categorized by their aggregated indicators.

4.3.1.2 District-Level Analyses

In order to analyze the relationship between the proposed performance labels and their geographic locations, three statistical analyses are conducted.

The proportion of each performance label in each district is computed, which represents the likelihood of a business being classified into each performance label within a specific district. Additionally, the relative ratios of performance label pairs for each district are calculated. Next, a correlation analysis to investigate the potential relationship between the proportions and relative ratios of performance labels with the census population and average household income of their respective districts. The findings indicate no significant correlations between the performance labels of businesses and the census population or residents' income level in their home districts as displayed in Figure A.1.

A Chi-squared test is conducted between district IDs and performance label proportions converted into categorical variables. This test is appropriate given the small sample size (33 districts), and the contingency table analysis is deemed valid. The results of the conducted chi-squared test, ($\chi^2(2560, n = 99) = 2629, p = 0.167$), reveal no significant dependencies are observed between the distributions of performance label proportions and relative ratios with the district IDs, which states that the proposed financial performance labels display no bias regarding the home district of the businesses.

In order to quantify the performance label distribution inequality at the district level, we employ the statistical inequality analysis [104] formulated in Equation 4.17, which considers the inequality, $Inequality^i_L$, of the distribution of performance labels in district i in terms of the proportions of the three labels within that district, which is denoted as $p^i_{L_k}$.

$$Inequality_L^i = \frac{3}{4} \times \sum_{k=1}^{3} |p_{L_k}^i - \frac{1}{3}| \tag{4.17}$$
$$(Inequality_L^i \in [0, 1])$$

The inequality framework in Equation 4.17 yields a score of 0, in case a uniform distribution is considered. In contrast, in case only a single performance label is dominant in a district, the employed inequality framework yields a score of 1, the highest score. Table 4.8 presents the summary statistics of the label inequality scores at the district level.

Range	Mean	Median	Standard deviation	Inter-quartile range
[0.042 , 0.388]	0.188	0.164	0.094	0.11

Table 4.8: Summary statistics of the inequality measures for performance labels at the district level.

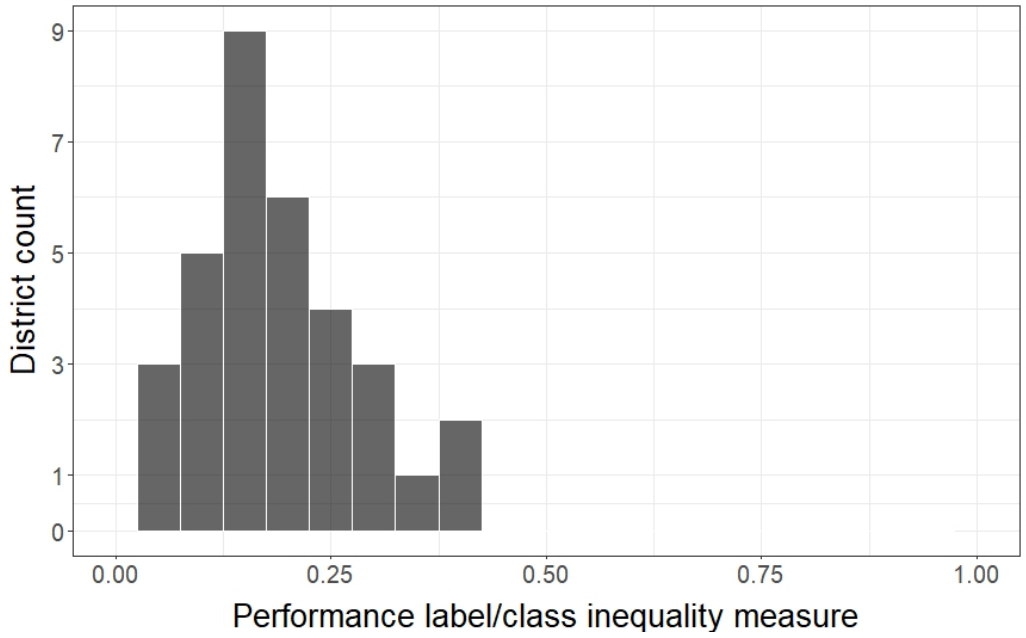

Figure 4.4: Histogram illustrating label inequality at the district level.

Theoretically, the employed inequality framework yields a score between 0 and 1, the perfectly equal and unequal cases, respectively. The histogram in Figure 4.4 and Table 4.8 show that the inequality scores predominantly cluster towards lower values. In particular, the highest obtained inequality score, 0.388, in this setting is less than 40% of the maximum value. Additionally, the statistical metrics, i.e., median, mean, and interquartile range, further reinforce the trend of the distribution displaying low inequality scores.

To conclude, the results derived from the conducted analyses confirm that there is no noticeable relationship between the proposed financial performance labels for businesses and their corresponding home districts.

4.3.1.3 Customer-Level Analysis

In order to investigate the relationship between the proposed financial performance labels and the customers' income levels, we conduct an additional statistical analysis. First, each business's customer income distribution is categorized based on quartiles, in which mean and median values are obtained. Next, a chi-squared test is conducted to analyze the relationship between the proposed financial performance labels and the quartiles of their corresponding customer's mean and median income. The results of the conducted statistical test on the mean income, $\chi^2(6, n = 1977) = 5.2951, p = 0.506$, and the median income, $\chi^2(6, n = 1977) = 2.8613, p = 0.826$, reveal that the income level of customers has no relationship with the proposed performance labels. An additional statistical analysis of customers' role in edge generation is presented in Table A.4 and Table A.5.

The district-level and customer-level analyses reveal that the proposed financial performance labels display no dependency or bias concerning their home locations or the socioeconomic characteristics of their customers, which confirms the resilience and validity of the labeling methodology introduced and applied in the study.

4.3.2 Predictive Setting

In order to predict the financial performance label of a business in the next time step, we consider the list of features from different settings, in which 5 of them are derived from the businesses, 3 features pertain to revenue information, 11 are acquired from customer characteristics, and 6 features are extracted from the proposed customer co-location network. The resulting customer co-location network comprises 2,011 nodes and 217,422 edges, forming a single strongly connected component. Alongside the 6 network-based features, we construct a feature vector of 128 dimensions for each business with the node2vec embedding algorithm.

Four different supervised machine learning models, namely logistic regression, random forest, support vector machines, and naive bayes, are employed. The AUROC scores of each feature set are displayed in Table 4.3.2 broken down by the employed classifiers. The results show that random forest-based approaches outperform other classifiers in most settings, which is also supported by the existing studies in the literature that emphasize the effectiveness of tree-based ensemble models on multi-class prediction tasks [55, 81, 143].

Moreover, Table 4.3.2 show that the proposed network-based features perform on par with the conventional revenue and customer-based features which are frequently employed by both the literature and the industry. In some cases, the models deployed with the proposed network-based and node2vec features outperform the customer-based features. However, the node2vec and network-based features do not demonstrate a striking advantage over the conventionally employed feature sets. However, the combination of the proposed network-based and traditional revenue and customer-based features increases the prediction performance, which may indicate that the network-based features are able to introduce a business's topological position information and enhance the model performance.

Feature set	NB	SVM	LR	RF
(A) Revenue	0.565	0.557	0.587	0.586
(B) Customer	0.558	0.561	0.579	0.571
(C) Network	0.561	0.556	0.575	0.579
(D) node2vec	0.537	0.556	0.567	0.572
(A) + (B)	0.566	0.566	0.588	0.591
(A) + (B) + (C)	0.569	0.567	0.598	0.609

Table 4.9: The results of the four machine learning models, namely naive bayes (NB), support vector machines (SVM), logistic regression (LR), and random forest (RF), on conventional features set, i.e., revenue and customer-based features, and the proposed network-based and node2vec features, evaluated with AU-ROC metric.

Figure 4.5 displays the impact of each feature on prediction accuracy, quantified by their impact on the mean decrease in accuracy based on random forest, the highest-performing algorithm among all the employed machine learning models. The figure highlights that financial features, e.g., revenue and the number of distinct customers, take place in higher ranks while the proposed network-based features are also significant, in particular, the impact of the degree centrality is vital to the model performance. Moreover, out of the top ten features determined by the mean decrease in accuracy, four are attributed to network-based features.

Table 4.3.2 and Figure 4.5 highlight that the proposed network-based features are able to perform on the same level as the conventionally employed revenue and customer-based features, where their AU-ROC evaluation scores differ by small margins, which might be the result of the correlation between the features from distinct feature sets. Table A.2 depicts a correlation table between the features employed in the conducted study.

Finally, considering the effectiveness of the node2vec features, we dimensionality reduction approach is taken to analyze the distribution of financial performance labels in a low-dimensional feature space. To this end, we focus on two different dimensionality reduction methods on the node2vec features, Principal Component

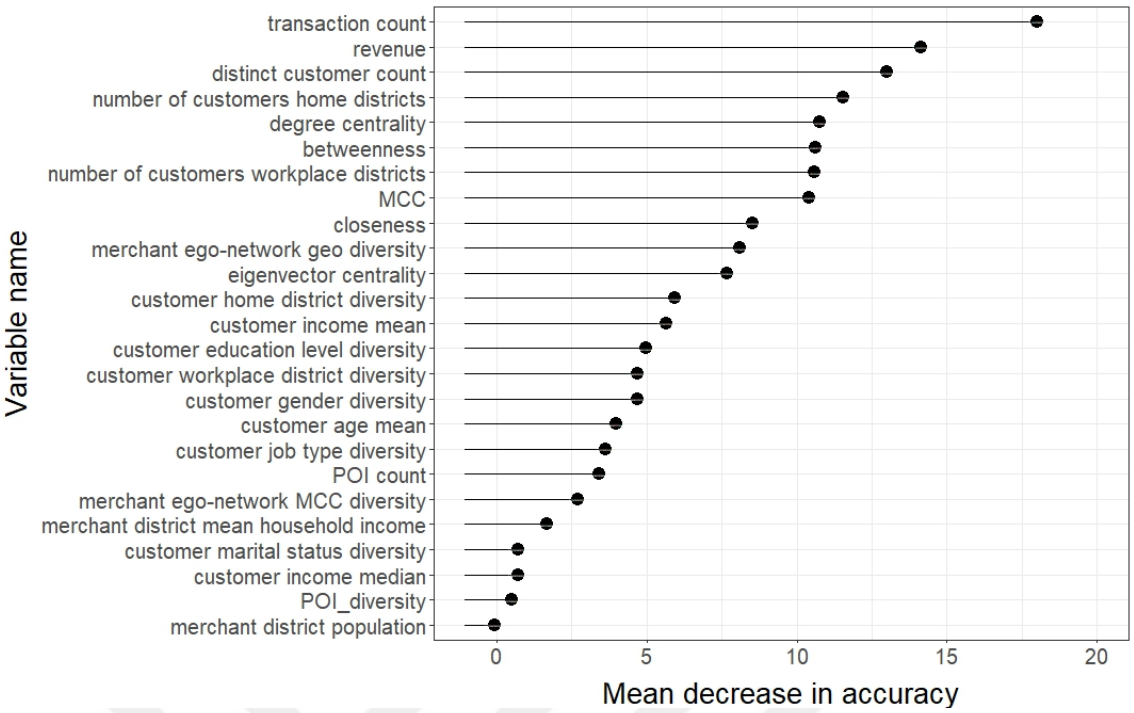

Figure 4.5: Ranking of features based on their importance considering the mean decrease in accuracy on the random forest classifier.

Analysis (PCA) [170] and T-distributed stochastic neighbor embedding (t-SNE) [164]. In this context, the two-dimensional feature space is obtained in which performance labels are color-coded as displayed in Figure A.3, which in turn does not reveal any clustering patterns considering the proposed financial performance labels.

4.3.3 Privacy Implications

In an intensely competitive landscape, the financial data of businesses holds a significant level of privacy concerns. Nevertheless, financial institutions face the necessity of assessing business risk to make well-informed decisions about loans. Consequently, accessing this confidential information becomes inescapable for such institutions. However, in the context of data sharing, ensuring the protection of businesses' revenue and customer-related data from unauthorized disclosure becomes crucial.

The conducted research in this dissertation, firstly, demonstrates that the proposed network-based features perform on par with the traditional features in predicting the future financial performance of businesses. Moreover, the proposed feature sets, provided in a tabular format, provide a higher level of privacy beyond the case where raw financial records are anonymized, which leads to the development of a more secure data-sharing environment with third-party entities. The resulting privacy-enhanced features hold a significant promise considering the facilitating

seamless collaboration and know-how sharing.

The proposed set of network-based features is relatively resilient against attacks aiming to infer the financial metrics of businesses given that the relevant purchase-related summary statistics are absent, which is almost always the case as financial institutions do not share such data. However, because of the evident correlation between financial indicators and network-based features, e.g., degree centrality, certain clues regarding the financial information may be employed in estimating the target attributes.

In order to deal with this problem, the proposed customer co-location network can be employed to produce node2vec node embeddings and be shared with other entities for downstream prediction tasks, which is a justified usage and illustrated in Table 4.3.2. The node2vec features are able to perform on par with conventional revenue and customer-based features while providing a higher level of safeguarding.

Adversarial attacks aiming to reconstruct a graph from node embeddings are presented in the literature. However, such methods are not able to recover the full extent of the original network. In addition, financial institutions are able to mitigate such issues by employing appropriate defense mechanisms considering downstream analysis tasks. One of the approaches focuses on perturbation [135, 182], in which the dimensions of the resulting node embeddings are systematically adjusted in a privacy vs accuracy tradeoff. To address this concern, as a simple iterative approach, the least significant feature vector from the node2vec matrix may be eliminated until no significant alterations are detected in the classification performance. This method is both simple and efficient, ensuring that our proposed approach does not compromise the classification performance.

4.4 Discussion

In local economies, businesses contribute substantially to employment and economic activities. Moreover, financial reports indicate that businesses live in a fragile economic environment, in which economic and financial downturns are significantly affecting the lifespan of businesses. In such an environment, businesses revert to loans and external investments from financial institutions. Considering the high failure rates, financial institutions are in need of predictive models to evaluate the risk and future financial performance of businesses. In this context, due to the inherent dynamics of the local businesses, non-standardized financial records and indicators introduce further issues for financial institutions. Moreover, privacy concerns regarding the available financial records are another issue for financial institutions to collaborate with third-party entities to develop data-driven methods, which emphasizes the need for methods for rapid and secure data-sharing methodologies regarding the downstream task of predicting the financial performance of businesses.

In the conducted research, credit card transaction data is employed, which are highly prevalent purchasing patterns in modern economies, to construct a social network for businesses in an urban area. To this end, a novel computational framework is presented, in which customer co-location networks constitute the backbone. The resulting co-location network is used to extract well-established network-based features, namely the centrality measures, which offer to encapsulate a business's topological importance and connectedness inside the network. The results of the conducted predictive analysis show that the proposed network-based features are able to predict the financial performance of businesses on the same level as the conventionally employed revenue and customer-based features.

In addition to the ability to predict the financial performance of businesses on par with the conventional feature sets, the proposed network-based features provide a higher level of safeguarding as opposed to utilizing raw financial indicators. The privacy-enhancing implications provide a secure data-sharing environment for financial institutions for further collaborations with third-party entities. However, the correlations between financial indicators and some of the network-based features, e.g., degree centrality, may introduce relatively minor privacy issues. To this end, node2vec node embeddings offer a solution to this problem. Although there exist adversarial attacks aiming to reconstruct original networks based on node embeddings, several perturbation-based mitigation strategies can be employed.

In the conducted study, one of the limitations arises from the employed credit card transaction data, which solely provides credit card-based transactions and ignores cash and online transactions, which are actively used in modern economies. In countries with less transparency and higher levels of informal businesses, the

problem becomes even more significant. In these cases, incorporating other mobility data sources, such as GPS and smartphone data [142], can be an option to construct customer co-location networks.

As another limitation, in the utilized credit card transaction data, there are no signals to differentiate business size, i.e., small and medium-sized enterprises (SME) and non-SMEs. Although financial indicators, such as transaction volume, could have been analyzed to categorize the businesses, no such categorization has been applied in this study. Developing separate models specifically designed for SMEs and non-SMEs would have increased the model performances, as their topological landscapes would take place distinctly considering their economic activities.

Despite its limitations, inspired by the literature on computational social science [89], a social network for businesses is proposed, in which shared customer bases constitute the backbone of the proposed network. To this end, based on credit card transaction data, customer co-location networks are constructed, from which a set of network-based features are harvested to predict a business's future financial performance in a novel evaluation setting. The predictive results demonstrate that the proposed network-based features perform on par with the conventionally employed revenue and customer-based features while providing a higher level of safeguarding and enhanced privacy levels, which facilitates a more secure data-sharing environment for financial entities.

Chapter 5

Neighborhood Adaptability Indicators During the COVID-19 Pandemic

The COVID-19 Pandemic has led to numerous non-pharmaceutical interventions (NPI) aiming to curb the spread of new infections in varying urban scales by limiting the mobility of masses, e.g., travel bans between countries, curfews or lockdowns in urban areas, and business closures. Considering the crucial role of mobility in a vibrant economy and the welfare of societies, NPIs transformed the spatio-temporal mobility patterns. However, the imposed NPIs do not affect the target population uniformly. The impact of NPIs varies significantly based on the sociodemographic composition of different populations in an urban area. Urban areas with different social fabrics exhibited a distinct pattern of adaptability in response to the imposed NPIs. In this work, our goal is to highlight the main drivers behind the adaptability of urban areas considering their sociodemographic traits. To this end, we present an application of human mobility patterns in a network setting to analyze the adaptability of urban areas. We analyze large-scale mobility data, encompassing the period from January 2019 to December 2020, from one of the economic hubs of the world, New York City (NYC), and create weekly mobility networks between neighborhoods based on accumulated point-of-interest (POI) visits.

Unlike previous research, our goal is to examine the evolving structure of mobility networks throughout the pandemic using well-established network metrics. This methodology allows us to swiftly identify and scrutinize the shifting dynamics and connections within these networks, offering crucial insights into how urban areas adapt amidst a public health crisis. Leveraging the insights obtained from the mobility networks, we focus on COVID-19 hotspots to explore the neighborhoods functioning as bridges between these hotspots and others, unraveling the contribut-

ing factors. Subsequently, employing the Huff gravity model, we investigate how enhanced accessibility to essential businesses like grocery stores might curtail interactions between COVID-19 hotspots and surrounding neighborhoods, potentially lowering infection rates and preserving lives.

5.1 Datasets

5.1.1 Safegraph Mobility and Point-of-Interest Dataset

The Safegraph dataset [131] offers detailed longitudinal geo-location data sourced from a diverse array of smartphone applications. This data is collected from millions of users from the United States of America and Canada who have consented to share their real-time location information. Instead of releasing individual mobility patterns, Safegraph publishes aggregated visit counts to point-of-interests (POI) originating from a particular urban area in a weekly manner. Safegraph aggregates mobility patterns based on *census block groups* (CBG), which is an administrative region that hosts 600 to 3000 residents. Figure 5.1 displays the hierarchical relationship between administrative regions in the US.

Safegraph provides weekly visit counts from census block groups to the compiled list of POIs. We focus on the time span from January 2019 (before the pandemic) to December 2020 (during the pandemic). The mobility patterns are geographically filtered to obtain the visits made in the New York Metropolitan Area, which results in 6,463 CBGs and 333,241 POIs in total (129,517 POIs are located in NYC).

5.1.2 Google COVID-19 Community Mobility Reports

The dataset, curated by Google, is designed to offer valuable insights into mobility trends, aiding in the comprehension of how communities respond to interventions during the COVID-19 pandemic. It aggregates data from diverse smartphone applications, including Google Maps. This dataset depicts the shifting dynamics of mobility trends across regions and various POI categories compared against a pre-pandemic normal day, established as a baseline. This baseline is derived from the median values across a five-week period from January 3rd to February 6th, 2020. The POI categories encompass retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential areas.

5.1.3 COVID-19 Cases

The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [13] offers a comprehensive COVID-19 data catalog that provides global in-

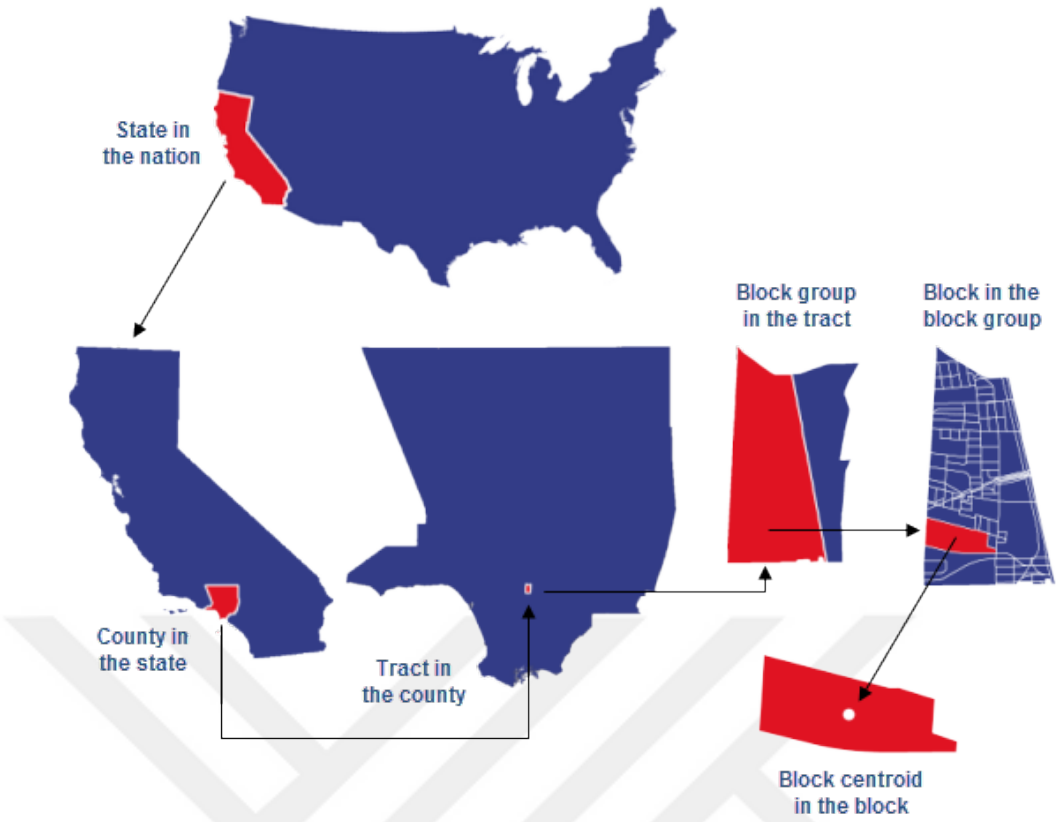

Figure 5.1: The hierarchical relationship between administrative regions in the United States of America. Counties in each state consist of census tracts and each census tract is comprised of census block groups.
Source: Image obtained from `https://learn.arcgis.com/en/related-concepts/united-states-census-geography.htm`.

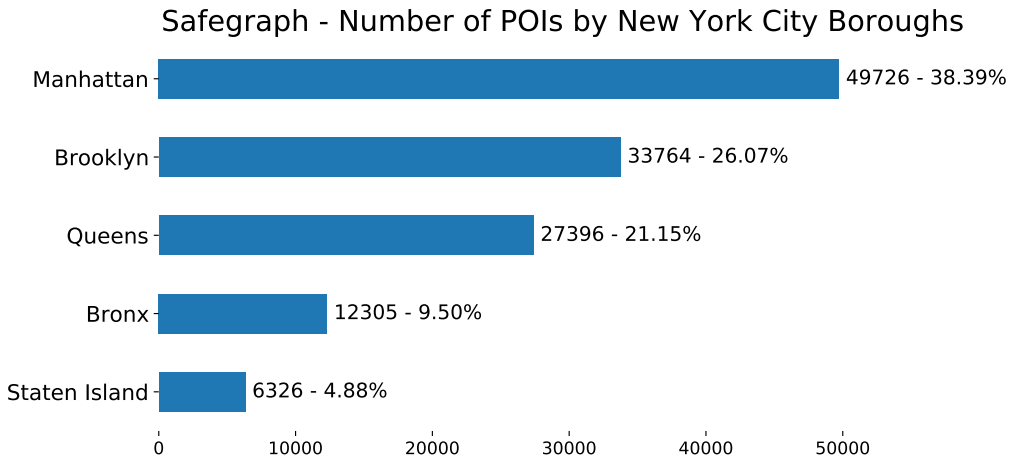

Figure 5.2: POI distribution provided by Safegraph in the five boroughs of New York City.

formation, catering to varying levels of granularity to suit different analytical needs. In New York City, data on metrics like new cases, test counts, and deaths are reported based on ZIP Code Tabulation Areas (ZCTAs). It's important to note that

a CBG might fall within the boundaries of multiple ZCTAs. To calculate the estimated weekly cases per CBG, we utilized a weighted average approach. This method involves considering the population ratio of each CBG within its respective ZCTA alongside the corresponding COVID-19 case rate. Assume S is the group of ZCTAs whose boundaries overlap with CBG_i. Then the number of COVID-19 cases for CBG_i at time t are estimated using the following formulations.

$$w_{ij} = \frac{population\ of\ \ CBG_i\ \ in\ \ ZCTA_j}{population\ \ of\ \ ZCTA_j} \tag{5.1}$$

$$CCBG_i^t = \sum_{j \in S} (w_{ij} \times case_j^t) \tag{5.2}$$

$CCBG_i^t$ denotes the estimated COVID-19 cases in CBG_i at time t, and $case_j^t$ is the number of COVID cases in $ZCTA_j$ at time t.

5.1.4 The United States Census Data

The U.S. Census Bureau's American Community Survey (ACS) releases demographic estimations at a Census Block Group (CBG) level. For our analysis, we utilized the most recent 5-year ACS data, gathered in 2019, to extract various demographic features. These include total population, median household income, education levels, commuting duration, and racial distribution for each CBG. Additionally, these features are portrayed with their respective percentile levels to provide a comprehensive representation.

5.1.5 New York Metropolitan Area

In 2019, the New York Metropolitan Area emerged as the most populous (hosting 19.22 million residents) and standing as a powerhouse in the US economy with a GDP of 1.522 billion dollars [149]. Encompassing counties from four states—New York, New Jersey, Connecticut, and Pennsylvania—the metropolitan area comprises of 7,809 census tracts and spans 23 counties. The POIs, as provided by Safegraph, are filtered based on their geographical proximity and relevance to the New York Metropolitan Area. New York City, home to approximately 9 million residents, stands as the largest city within the New York Metropolitan area, encompassing 6,493 CBGs. NYC is segmented into five distinct administrative divisions known as boroughs: Manhattan, Brooklyn, Bronx, Queens, and Staten Island.

5.2 Methods

5.2.1 Constructing Mobility Networks

We consider the mobility patterns between CBGs as weighted directed networks, $G^{(t)} = (V^{(t)}, E^{(t)})$, for each time step t, in which the nodes $V^{(t)}$ correspond to CBGs and edges $E^{(t)}$ correspond to the weekly number of visits originating from a CBG to another with weight $w_{ij}^{(t)}$ capturing the number of visits from CBG_i to CBG_j in time step t. The time steps are comprised of weeks spanning from January 2019 through December 2020.

In the Safegraph dataset, the cumulative count of visits from CBGs to a particular POI is recorded. To capture the overall mobility patterns between CBGs, we take the POIs in a specific CBG as mobility aggregators. In this setting, the CBG hosting the destination POI is designated as the target CBG. This implies that the set of POIs, denoted as $P_i = \{p_1, ..., p_n\}$ within CBG_i, serve as a representation for capturing the inbound mobility to CBG_i. The accumulated number of visits to P_i forms the basis for the weighted incoming edges associated with CBG_i. Moreover, the distances within the road network connecting CBGs are integrated as edge attributes.

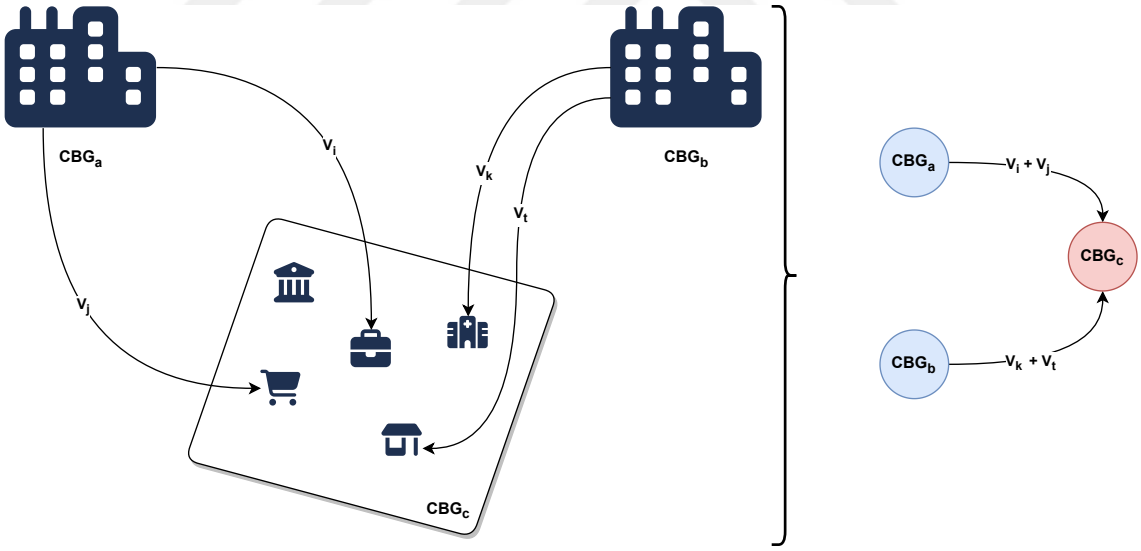

Figure 5.3: Constructing mobility networks among the CBGs through accumulating visits to POIs. The number of visits, V, from each CBG$_x$ to the POIs located in the target CBG are aggregated to be used as edge weights in the resulting mobility networks.

5.2.2 Temporal Dynamics of Mobility Networks

To illustrate the evolving network structure over time, our approach initiates by examining the dissimilarities between paired weekly networks in 2019 and 2020.

Subsequently, we classify nodes based on demographic groups to delve deeper into this analysis. Following this, we explore the correlation between centrality metrics and these demographic groupings to gain insights into their relationship within the network.

5.2.2.1 Dissimilarity Analysis

To create the feature vectors for the synchronized weekly networks of 2019 and 2020, we adopt the ego-network-based node features outlined by Berlingerio et al [24]. These features serve as key components in characterizing the nodes within the networks.

Node Features
⋆ Node In & Out Degree
⋆ Node Strength
⋆ Local Clustering Coefficient
⋆ Average Degree & Average Clustering Coefficient of Node's Neighborhood
⋆ Number of Edges & Alters in Node's Ego-Network
⋆ Number of Out-going Edges from Node's Ego-Network

Table 5.1: The list of node features to be used in the dissimilarity analysis.

To calculate the dissimilarity score between the paired weekly networks, we consolidate the resulting node feature vectors into a unified vector. This process involves deriving network features by employing statistical aggregation methods like standard deviation, median, kurtosis, and skewness on the individual node features. Next, in the node-level dissimilarity analysis, we directly employ paired node feature vectors without aggregation, generating a distinct score for each individual node. For both analyses, Canberra distance [88] is applied on the paired feature vectors, P and Q.

$$d(P,Q) = \sum \frac{|P - Q|}{|P| + |Q|} \tag{5.3}$$

5.2.2.2 Analyzing Centrality Metrics

Neighborhoods with different demographic distributions display varying responses to the imposed NPIs over time, due to their different needs and socioeconomic characteristics. CBGs hosting a substantial amount of POIs emerge as common destinations. By utilizing centrality metrics, we illustrate the temporal evolution of CBGs' topological significance within the network. In particular, we focus on *betweenness*, *in-degree*, *out-degree*, and *self-visit ratio*. To approximate the number of visits made by individuals to the POIs within their home CBG, we define a custom metric named self-visit ratio as

$$S_c^t = \frac{W_l^t}{W_l^t + W_o^t}, \tag{5.4}$$

where W_l^t is the sum of weights on self-loops and W_o^t is the sum of weights on the outgoing edges in time step t for CBG c. The self-visit ratio serves as an indicator of the frequency with which residents tend to visit locations within their immediate locality.

5.2.3 COVID-19 Hotspots and Bridge CBGs

The mobility within each CBG is influenced by numerous parameters intricately tied to social dynamics, resulting in a complex interplay of factors. Given the clear correlation between mobility and the spread of COVID-19 infections [31, 77, 173], CBGs exhibiting higher mobility rates might pose a potential challenge in containing the prevalence of the disease. We identify the CBGs that are located in the top weekly new COVID-19 cases as COVID-19 *hotspots* and the CBGs that interact frequently with such hotspots as COVID-19 *bridge* CBGs. To identify and analyze potential spreaders or bridges, we leverage the visit frequencies between CBGs. To this end, we consider a two-week time span as the incubation period for the emergence of new cases [31, 77]. Starting from March 2nd, 2020, the date of NYC's first reported case, we identify the CBGs in the top weekly new cases quartile in week t, which constitute the hotspots for that time step. Next, the CBGs that have an outgoing edge to the identified group in week $t-2$ are recorded. The resulting pool of candidate bridge CBGs is then applied a percentile-based frequency filtering to determine the COVID-19 bridge CBGs. To accomplish this, we explore the CBGs with occurrence frequencies in the 75^{th} percentile, considering them as the final group of bridge CBGs. This analysis aims to uncover any distinctive socioeconomic and demographic traits within the bridge CBGs, shedding light on the reasons behind their distinct mobility patterns.

5.2.4 Huff Gravity Model

To simulate the mobility patterns under hypothetical POI densities, we employ the Huff gravity model. In this setting, we analyze the hypothetical mobility flow between bridge and hotspot CBGs under varying POI distributions. To this end, we frame the attractiveness of a POI as its store area in square meters.

$$P_{ij} = \frac{\frac{A_j^\alpha}{D_{ij}^\beta}}{\sum_{k=1}^{n} \frac{A_k^\alpha}{D_{ik}^\beta}} \tag{5.5}$$

P_{ij} is to the probability of a resident at CBG i visiting POI j among the available POI set k. The exponents α and β control the weight of distance and attractiveness, i.e., store area. Due to a notable decrease in movement activities amid the COVID-19 pandemic, we aggregate CBG visits into census tract-level data to ensure a more robust amount of observations, enhancing the model's fitting accuracy. We treat each census tract as an independent mobility center and employ the Particle Swarm Optimization (PSO) technique to estimate a pair of parameters, α and β, for each census tract.

As we treat each census tract as an individual mobility center, we individually fit a model for all 109 census tracts, acquiring respective α and β exponents. Various methods in the literature are available for estimating these exponents, including ordinary least squares [74], geographically weighted regression [151], and PSO techniques. Recent studies indicate that the PSO technique tends to yield more accurate estimates for the α and β exponents [15, 95, 152]. Therefore, we employ PSO with an objective function that maximizes the Pearson Correlation between the ground truth and the estimated visit ratios. Equation 5.6 showcases the objective function, where the minimized coefficient is negated to align with the requirements of the employed software package. Figure 5.4 displays the resulting coefficient distribution between the ground truth and predicted visits, which highlights a respectable median coefficient value of 0.6.

$$\begin{aligned} a &= \text{actual visit ratios} \\ e_{\alpha,\beta} &= \text{estimated visit ratios with respect to } \alpha \ \& \ \beta \\ \arg\min_{\alpha,\beta} &= 1 - pearsonr(a, e_{\alpha,\beta}) \end{aligned} \tag{5.6}$$

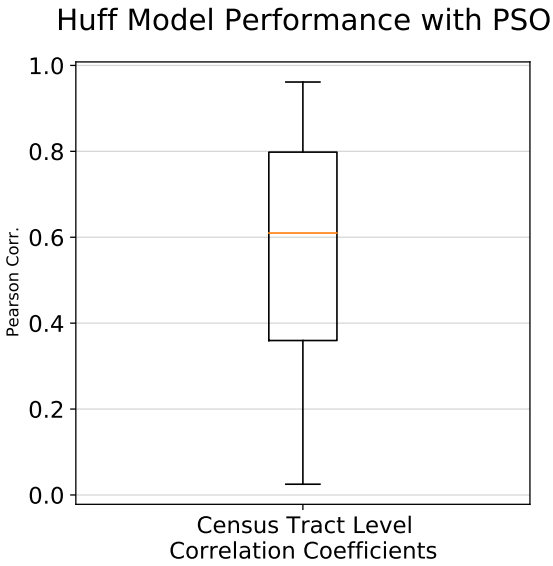

Figure 5.4: Huff gravity model evaluation under PSO estimation. For each census tract, separate α and β parameters are calculated. To evaluate the performance, we consider the correlation between ground truth visit records and the predicted visits. The resulting coefficient distribution, with a median of 0.6, is displayed.

5.3 Results

To elucidate the impact of the COVID-19 pandemic on various socioeconomic groups, our first undertaking involves scrutinizing the weekly fluctuations in network structures at the neighborhood level. In this context, we derive node-level feature vectors that encapsulate statistical properties, providing a concise summary of the individual ego-networks within these neighborhoods. [24]. The node feature vectors we extract are subsequently employed to calculate dissimilarities between paired weekly networks, contrasting those from 2019 with their counterparts in 2020. Subsequently, we illustrate the dynamic shifts in centrality metrics across various sociodemographic groups and showcase their inherent variability. Our analysis advances by examining potential COVID-19 bridges—neighborhoods engaging in frequent interactions with COVID-19 hotspots, particularly CBGs hosting higher numbers of infected residents. Our emphasis lies in studying the inbound and outbound edges between CBGs, observed over two-week periods to align with the typical incubation time of the virus. [31]. Finally, with the help of the Huff gravity model, we analyze the changes in visits to hotspot CBGs in hypothetical scenarios that deal with differing points-of-interest densities.

5.3.1 Demographic Disparities: Unfolding Dynamics in Mobility Networks Across Time

5.3.1.1 CBG-Level Analysis

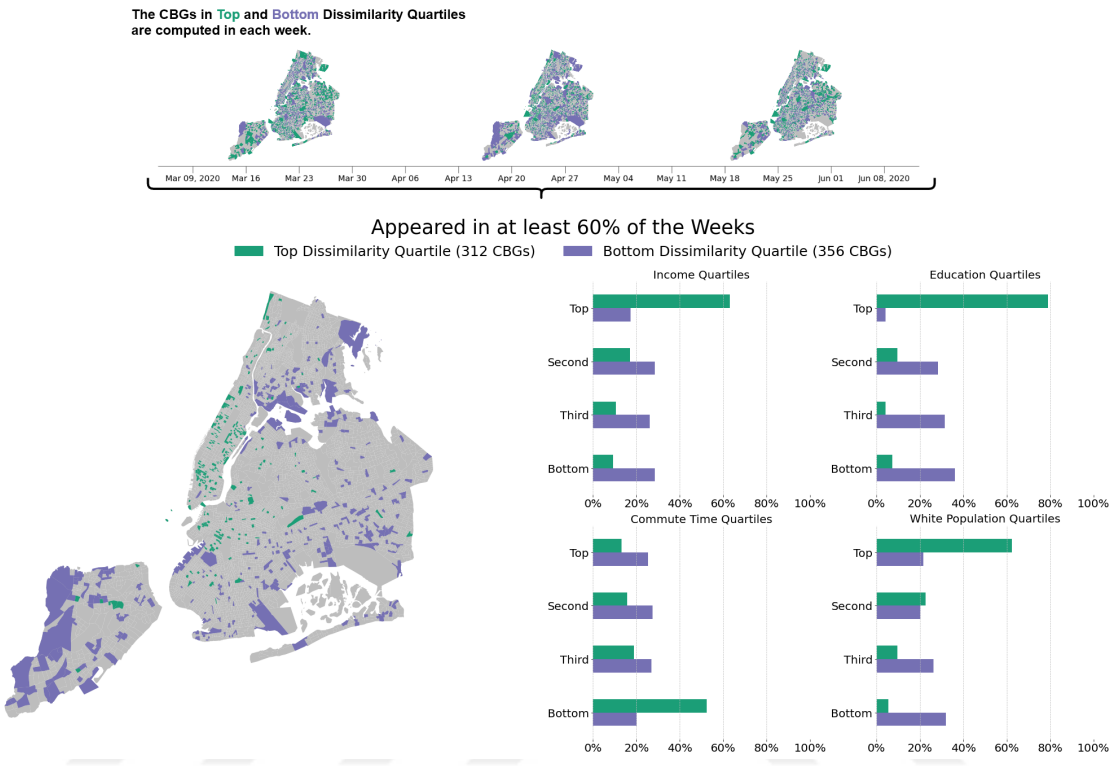

Figure 5.5: The socioeconomic distribution of CBGs undergoing the most pronounced shifts in mobility patterns, highlighted in green for a minimum of 60% of the observed time steps, contrasts with the least changed, marked in purple. The uniform color scheme serves not only to delineate the spatial distribution but also to reflect the demographic characteristics of each group. This approach effectively illustrates how these characteristics are distributed across quartiles, emphasizing socioeconomic traits. It's worth noting that while significant socioeconomic characteristics are discernible for the top-quartile CBGs, they are absent for the bottom-quartile.

We compute dissimilarity scores at the node level between paired weekly networks of 2019 (pre-pandemic) and 2020 (pandemic) using the acquired ego-network feature vectors. At each time step, CBGs are ranked according to their dissimilarity scores. To highlight variations among CBGs displaying divergent behaviors, we focus on those residing in the top and bottom dissimilarity quartiles. Two cohorts are formed, consisting of CBGs consistently present in these quartiles during the initial wave of the pandemic, from March to June 2020. Figure 5.5 highlights the spatial and socioeconomic distribution of the CBGs appearing in the top and bottom dissimilarity quartiles across at least 60% of the time steps. To illustrate the demographic profile of each cohort, we demonstrate their distributions across socioeconomic traits

through quartiles. This visualization employs a color scheme that aligns consistently with the spatial distribution representation. Setting a threshold at 60%, we achieve a balanced count of CBGs in each group, bolstering the significance of our comparison. Notably, within the CBGs exhibiting the most pronounced shifts in mobility patterns and falling into the top dissimilarity quartile, a conspicuous concentration is observed within Manhattan, the socioeconomic hub of New York City. Within this group of CBGs, a noteworthy demographic distribution becomes apparent: 63% are positioned in the top quartile for income, 79% in the highest education quartile, 62% in the top quartile for the white population percentage, and 52% in the bottom quartile for commute time. No clear socioeconomic profile emerges for the bottom dissimilarity quartile, representing the CBGs with minimal changes in their mobility patterns. While the quartile distributions provide some insights into the residents, there is a discernible decreasing trend from the bottom to the top quartiles concerning income, education, and the percentage of the white population.

5.3.1.2 Node Centrality Analysis

In constructed mobility networks, where each node corresponds to a CBG, centrality metrics play a pivotal role in offering valuable insights into a node's significance within the network [45, 102]. These metrics are crucial for identifying standout CBGs based on population flow dynamics.

Within this framework, examining temporal shifts in centrality metrics reveals interaction patterns among diverse socioeconomic communities, providing insights into intricate mobility behaviors from a network perspective. To this end, we conduct our analysis based on two well-established centrality metrics, *betweenness* and *degree*, in addition to a custom metric named *self-visit ratio*.

Betweenness: This centrality score considers how often a CBG appears along the shortest paths in a network and is the sole node centrality metric showing a noteworthy difference among the chosen demographic groups. As highlighted in Figure 5.6-A, CBGs in the top income quartile maintained a higher betweenness value until the onset of March 2020 (the beginning of the pandemic), signifying their crucial role in bridging the flow of masses. Nevertheless, there was a sudden decline in the betweenness values among the top-income CBGs after the onset of the pandemic, whereas less affluent CBGs experienced an increase in their betweenness scores. In other words, less affluent CBGs progressively served as connectors among the nodes in the mobility network, but this trend persisted only until September 2020, coinciding with the revival of economic activity. A similar correlation is evident when examining education levels as depicted in Figure 5.7. CBGs with lower educational attainment exhibited a higher betweenness score within the same time frame.

Degree: Node degree analysis reveals that income and education significantly

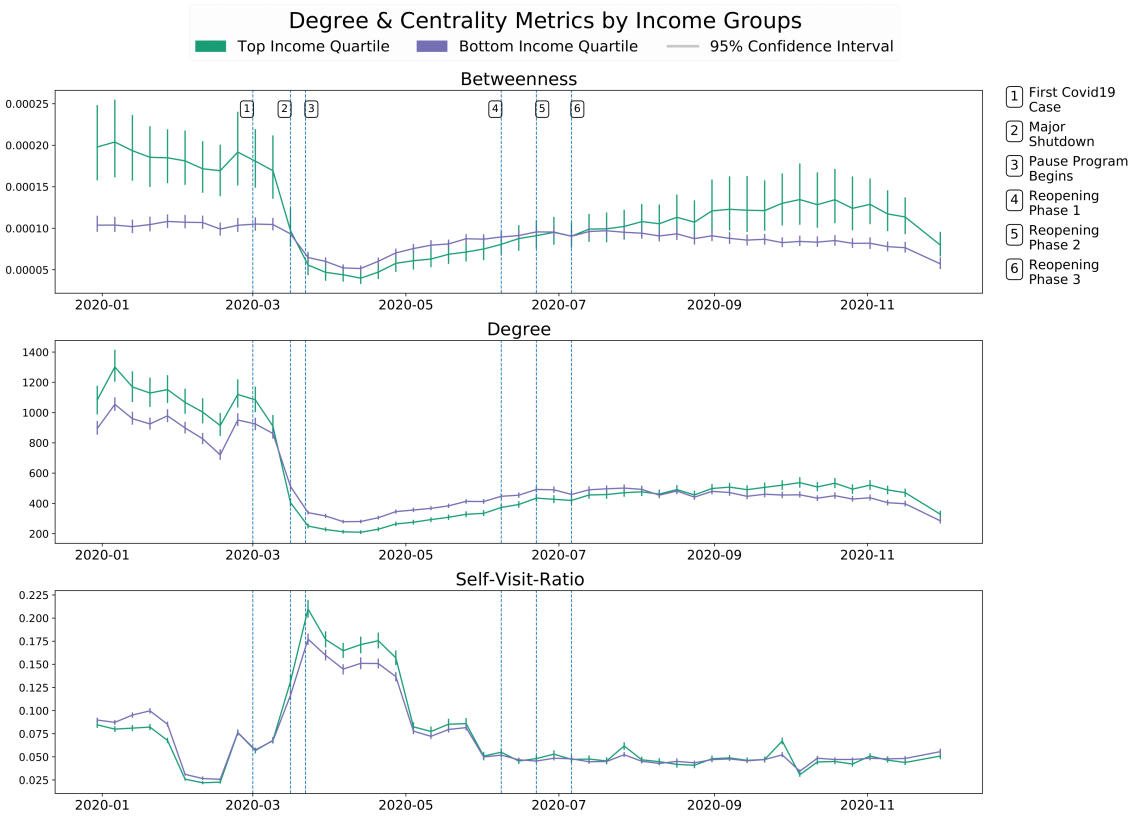

Figure 5.6: Analyzing the temporal shifts of three key centrality metrics—(A) betweenness, (B) total-degree, and (C) self-visit ratio—in both the top and bottom income quartiles. The vertical line segments represented in the graphs depict a 95% confidence interval.

influence the distribution of degree centrality values. As displayed in Figure 5.6-B, wealthier CBGs were more successful in reducing their mobility compared to less affluent neighborhoods.

Self-Visit Ratio: Figure 5.6-C depicts the trend of the self-visit ratio concerning the top and bottom income quartiles. The self-visit ratio explains the fraction of visits made in the home CBG. From March to June 2020, encompassing the first wave and the most pronounced decline in mobility, CBGs in the top-income quartile exhibited a higher rate of visits to the POIs within their home CBGs. Conversely, residents of lower-income CBGs displayed a lower self-visit ratio. The gap between income groups began to diminish with gradual re-openings that took place in June 2020.

5.3.2 COVID-19 Hotspots, Bridge CBGs & the Case of Staten Island

In order to analyze the neighborhoods and their sociodemographic characteristics that act as a mediator for the virus to spread, we define bridge CBGs. To this end, for

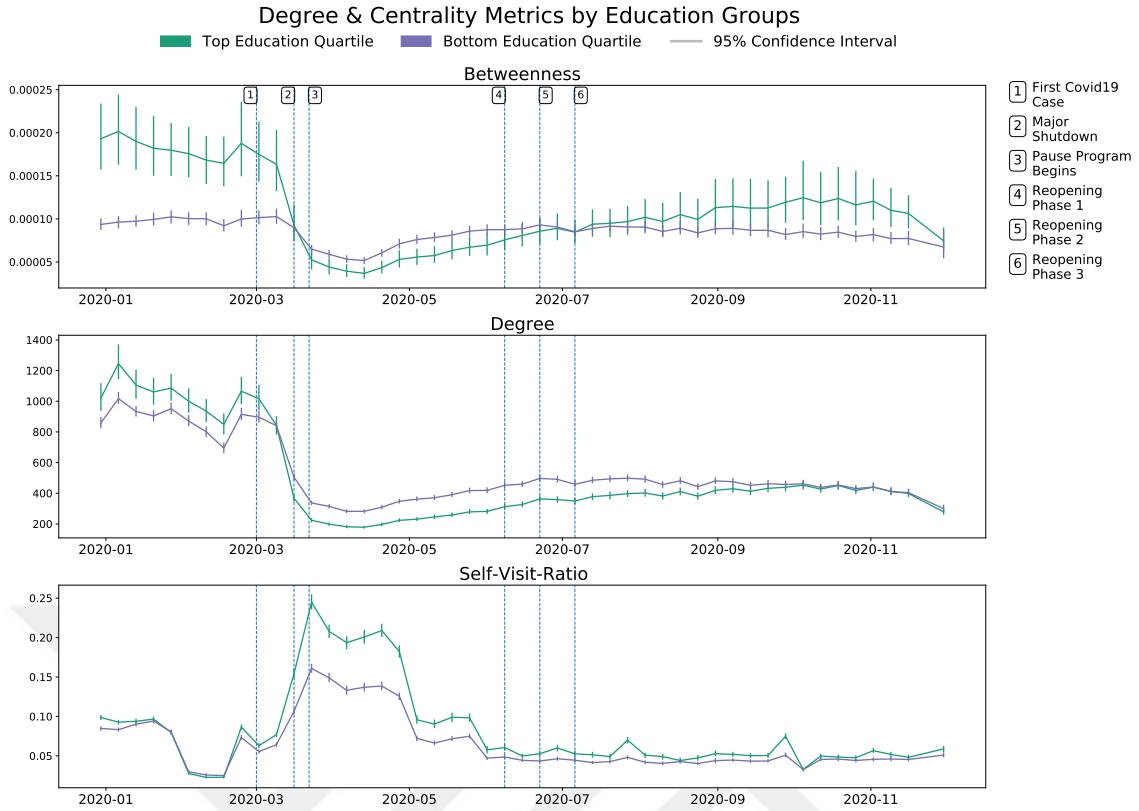

Figure 5.7: Analyzing the temporal shifts of three key centrality metrics—(A) betweenness, (B) total-degree, and (C) self-visit ratio—in both the top and bottom education quartiles. The vertical line segments represented in the graphs depict a 95% confidence interval.

each timestep, we first extract the CBGs with the highest new cases and name them as hotspots. Lastly, we analyze the visitation patterns of CBGs to the hotspots based on their frequencies. Analyzing the CBGs identified as COVID-19 bridge areas could offer effective insights for policymakers and public administration offices aiming to curb the spread of new infections and develop cities resilient to future pandemics.

In order to determine the bridge CBGs, we first isolate the CBGs in the top new cases quartile for each time step t as the hotspots. Next, over the course of the observed time frame of the pandemic, the CBGs that visited, i.e., had an outgoing edge in the mobility network, the hotspots in time step $t-2$ (considering the virus incubation period [31]) are compiled in a pool that stores the candidate bridges.

Finally, the CBGs in the resulting pool of candidate bridges are ranked based on their frequencies, i.e., how often they paid a visit to the hotspots in the observed time frame. Based on an empirical threshold value, the final bridge CBGs are identified as those surpassing the 75^{th} occurrence percentile.

Figure 5.8 highlights the geographic and sociodemographic distribution of the resulting bridge CBGs. The plot demonstrates that a significant percentage of the

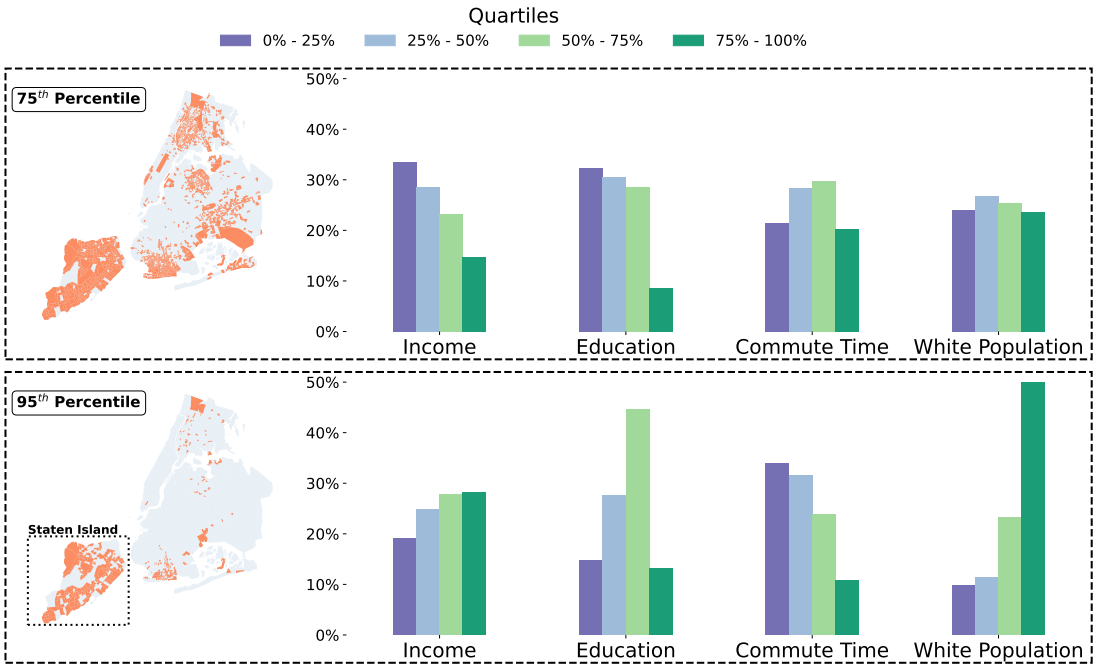

Figure 5.8: The geographic and demographic patterns of CBGs in the 75^{th} and 95^{th} frequency percentiles, identified as COVID-19 bridges, highlight Staten Island's distinct prominence.

bridge CBGs consist of low-income and education quartiles with a high commuting time. However, the geographic distribution of the bridge CBGs reveals the unique case of Staten Island (highlighted in Figure 5.8), where 85% of the Staten Island CBGs are considered bridges. In order to get a picture of the distribution with a higher threshold value, we extracted the CBGs above the 95^{th} percentile, with which the distributions begin to demonstrate the sociodemographic characteristics of Staten Island. Figure 5.9 displays the COVID-19 bridge CBG distributions of the NYC boroughs as a boxplot. Moreover, we also conduct an OLS regression analysis to delineate the relationship between NYC boroughs and their CBGs' bridge occurrences. In this setting, bridge occurrences are considered as the dependent variable while their corresponding borough codes are employed as the independent variables as demonstrated in Table B.1, which shows that boroughs are able to explain the occurrences of bridge CBGs. This finding is counter-intuitive and contradicts previous observations, as 48% of the CBGs in Staten Island belong to high-income quartiles, and their residents are predominantly white. In other words, Staten Island CBGs exhibit a distinctive behavior compared to CBGs with similar sociodemographic characteristics, i.e., high-income and high-white populations. This finding is notable as Staten Island is an island with limited connectivity compared to the rest of the boroughs and also has the lowest POI density as displayed in Figure 5.2. Given the stark mobility differences between Staten Island and its sociodemographic counterparts, Staten Island would have been more protected from the effects

of the pandemic. In light of the obtained observations and findings, we conduct an additional borough-level analysis on Staten Island.

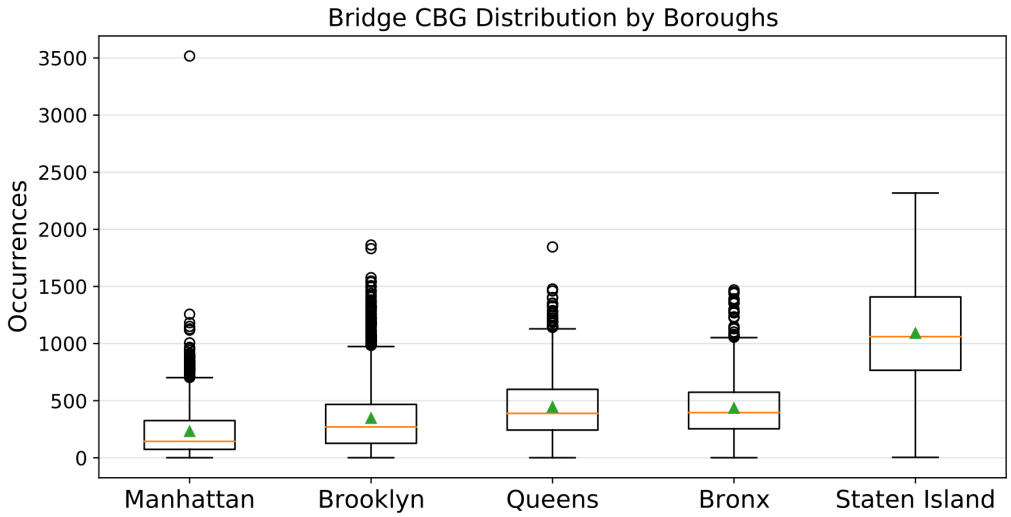

Figure 5.9: The distribution of bridge occurrences among CBGs across NYC boroughs throughout the observed time frame of the pandemic.

5.3.2.1 Borough-Level Analysis

Due to the limited workplace and office coverage of Safegraph POI data, we employ the COVID-19 Community Mobility Reports [2] provided by Google to analyze the mobility patterns targeting the workplaces at the borough level. Staten Island possesses the lowest relative change in visits to workplaces, which reveals that Staten Island residents were not as successful as other boroughs in lowering their mobility patterns as demonstrated in Figure 5.10.

Staten Island hosts the lowest number of POIs in addition to the lowest diversity. Moreover, the visits originating from Staten Island are significantly made to Brooklyn and Manhattan, the neighboring boroughs of Staten Island. This finding aligns with a report from the NYC government's planning department [114], which states that 24% of workers residing in Staten Island have their workplaces situated in Manhattan.

To summarize, the low POI density in Staten Island, coupled with its relatively isolated geographic position (e.g., just one automobile bridge to Brooklyn, one free ferry to Manhattan, no subway), which encourages the use of personal cars [113, 114], necessitated residents to visit POIs in other boroughs to fulfill their needs. As a result, they traveled longer distances to workplaces and the majority of POI categories. We hypothesize that this is likely the reason why a distinct response behavior among the CBGs in Staten Island is observed compared to their demographic counterparts, i.e., CBGs with relatively high income and percentage of white population.

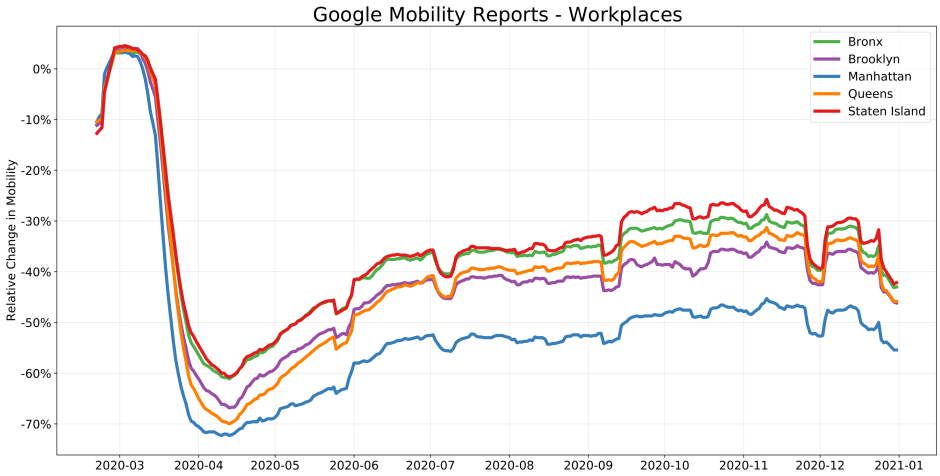

Figure 5.10: Relative change in mobility visits, with respect to the first week of the year, to the places of work by boroughs.

5.3.3 Hypothetical Scenario Analysis

Considering the low POI density of Staten Island, we conduct a hypothetical scenario analysis concerning the mobility patterns of Staten Island under differing POI densities. In this setting, the Huff Gravity Model [75] is employed to create synthetic visits from CBG to POIs. The Huff model incorporates the distance (between CBGs and POIs) and the POI attraction indicators into consideration. In the employed analytical setting, the POI attraction is modeled with respect to the store area in square feet. In addition, we only consider grocery stores, which sustain one of the most essential needs of human daily lives, for the simulations to narrow down the scope of the analysis.

Borough Name	**Grocery Stores per 1K Residents**	**Median Distance Travelled (in km)**
Manhattan	0.582	0.90
Brooklyn	0.470	1.35
Bronx	0.435	1.26
Queens	0.414	1.71
Staten Island	0.332	2.66

Table 5.2: The density of grocery stores per 1,000 residents and the median distance traveled by residents to reach grocery stores in kilometers, broken down by NYC boroughs.

Among all the boroughs in New York City, Staten Island has the lowest number of grocery stores per resident and the highest median distance traveled to grocery stores, as demonstrated in Table 5.2. The primary goal of the hypothetical scenario analysis is to depict the mobility patterns of Staten Island, the borough with the highest bridge CBGs, under different POI densities, in particular, the POI densities

of the boroughs with similar sociodemographic characteristics. The simulations consider the first wave of the pandemic between March 22nd (the pause program begins) and June 8th (the first phase of the reopenings). During the first wave of the pandemic, a strict lockdown policy was enforced, which led to business and workplace closures. Only essential businesses such as grocery stores were allowed to operate. Hence, we argue that the majority of trips to grocery stores during that time frame could be characterized as single-purpose trips rather than multi-purpose ones [99], which allows us to employ the distance between POIs and CBGs in the model. Furthermore, because of enforced lockdowns, the observed mobility patterns during the first wave of the pandemic were significantly lowered at the CBG level. In order to have more balanced mobility distributions, we focus on census tract-level mobility patterns and treat them as individual mobility centers. For each census tract, we estimate its own distance and attraction exponents, i.e., α and β, which in turn produces the probability of the residents visiting a particular POI. We employ the PSO optimization method to estimate the α and β exponents that yield the highest correlation with Safegraph mobility patterns.

In order to match the POI densities of Manhattan and Queens, the boroughs with similar sociodemographic characteristics, we introduce randomly generated hypothetical grocery stores inside Staten Island until the number of grocery stores per 1,000 residents matches the densities of the aforementioned boroughs. Once the α and β exponents are obtained, the model can be employed to generate hypothetical visit probabilities for each POI. In this setting, to obtain the number of visits from census tracts to POIs, we revert to the Safegraph mobility data. The total visits originating from a census tract are used in tandem with the computed probabilities to produce the hypothetical visits. The resulting hypothetical visits are then analyzed considering the visits originating from Staten Island census tracts to hotspot CBGs, which would demonstrate the visitation patterns of Staten Island with a higher POI density. Our results demonstrate that the visits to hotspots would have decreased by 47% and 23% by Manhattan and Queens' grocery store densities, respectively, which implies reduced exposure to potential COVID-19 spreaders, lowering the risk of contamination and ultimately reduced mortality rates. A detail view of the simulation is highlighted in Figure B.2.

5.4 Discussion

In this study, we extend the existing research on utilizing network structures to elucidate human behavior in socioeconomic contexts by examining the intricate relationship between human mobility, socioeconomic outcomes, and demographic attributes amidst the COVID-19 pandemic. Specifically, we adopt a network analysis

approach to explore the impact of the pandemic on the mobility patterns of NYC residents across its five boroughs and 6,493 CBGs throughout 2020. CBGs serve as network nodes, while links denote visits between pairs of CBGs by residents to POIs. We analyze node-specific and ego-network-based structural features to calculate dissimilarity scores between weekly networks in 2019 and 2020, aiming to comprehend the extent of change in the network structure over the years.

Furthermore, we examine the temporal evolution of node and degree centrality metrics across various socioeconomic groups. Our methodology and results unveil that while the COVID-19 response measures induced structural alterations in the mobility network, CBGs exhibiting minimal changes in their ego-network structure displayed elevated COVID-19 infection rates. Predominantly, these nodes originate from neighborhoods characterized by low income and education levels, hosting a higher proportion of frontline workers [115], such as those in healthcare, grocery, convenience and drug stores, childcare, food and family services, public transport, trucking, warehouse, and postal services. These individuals faced challenges in reducing their mobility due to the nature of their occupations, which entail working outside the home and frequent commuting.

The CBG-level analysis reveals a distinct demographic profile of residents residing in top dissimilarity quartile CBGs, which emerged in more than 60% of the weekly patterns scrutinized. Such CBGs tend to have higher income, education levels, and predominantly white populations. We hypothesize that these residents are more likely to be employed in job sectors conducive to adapting to shelter-in-place and physical/social distancing mandates through remote work, unlike residents in neighborhoods with different occupational structures. Consequently, in alignment with findings from other studies [31, 34, 76], we ascertain that these top-dissimilarity CBGs and their residents demonstrate greater resilience in the face of pandemic conditions such as COVID-19.

5.4.1 Two Faces of a City

Using a network setting, the conducted study shows that the adaptability of mobility patterns also varies based on the geographical attributes of neighborhoods, in addition to the sociodemographic characteristics. The network dissimilarity analysis demonstrates that less affluent and less educated neighborhoods exhibit lower adaptability to policy interventions aimed at reducing mobility levels. Additionally, we observe that neighborhoods with higher income and education levels can exhibit similar behavior to less affluent and less educated neighborhoods if they have limited access to public transportation options, workplaces, and other assorted amenities, i.e., the case of Staten Island. Despite the limited physical connections between

Staten Island and other boroughs of NYC, we find that the network changes were minimal, indicating relative fragility to the COVID-19 pandemic in terms of infections. Focusing on similar neighborhoods or isolated geographical units in other urban settlements could be the focus of future direction to aid policymakers in crafting policies to mitigate the impact of a pandemic in such areas and enhance their resilience.

5.4.2 Implications for Urban Planning

Our research offers insights into urban planning and policymaking. Both socioeconomic and geographic attributes of neighborhoods are crucial factors in enhancing neighborhoods' resilience to exogenous shocks. Regarding the latter, and drawing from our simulation findings, convenient access rates to POIs providing daily essential needs (e.g., grocery stores), workplaces, and hubs of attraction offering a range of amenities through a diverse set of POIs is expected to mitigate the necessity for extensive travel distances. Based on the Huff Gravity Model, we conducted a hypothetical scenario analysis on generating the synthetic number of visits to COVID-19 hotspot areas with the inclusion of hypothetically added grocery stores in Staten Island. Despite our model solely incorporating the distance and POI floor area, the results offer insights into simulated mobility. The results suggest that enhanced access to essential POIs leads to reduced exposure to hotspot districts for residents, enabling them to fulfill their needs without having to travel longer distances. In situations like the COVID-19 pandemic, improved access to POIs has the potential to lower infection rates and save lives.

5.4.3 Limitations

The utilized mobility data has several limitations. The employed mobility patterns are aggregated at the CBG level in a weekly manner. Although this approach respects the privacy of smartphone users, the resulting data granularity does not provide any clue regarding the trip purposes. The additional knowledge on trip purposes would have enhanced our studies to distinguish the essential and non-essential trips made by the residents and identify their sociodemographic traits. Moreover, the utilized mobility patterns dataset does not consider the mode of transportation considering the visits from CBGs to POIs. Although it is possible to estimate the most used transportation modes and routes between CBGs based on utility datasets, the granularity provided by Safegraph does not allow us to perform an analysis on determining the transportation modes taken by the residents of CBGs with differing sociodemographic characteristics to analyze the spread of the virus.

Moreover, Safegraph POI data mostly provides POI where financial transactions

take place, such as grocery stores, restaurants, and department stores, due to the nature of the data collection procedure. As a result, the POI data provides a low coverage of workplaces and offices, which would have allowed us to pinpoint the essential trips made by the residents and better fit the mobility model.

In conclusion, despite its limitations, the conducted research in this study offers two significant contributions. First, the findings provide insights into the diversity of mobility patterns across various neighborhoods during the COVID-19 pandemic by employing network science methodologies and hypothetical scenario analysis. Second, the research sheds light on the factors influencing a neighborhood's resilience and adaptability, which may assist urban planners and public administration offices in recommending sustainable policies, making informed intervention decisions, and responding effectively to future exogenous shocks such as the COVID-19 pandemic, ultimately saving more lives.

Chapter 6

Conclusions

Given its intricate interplay with socioeconomic dynamics, the multifaceted nature of human mobility analysis emerges as a pivotal avenue for researchers to understand and address a spectrum of societal challenges from a holistic perspective. Big data sources, such as CDR, financial transactions, and online behavior data, enable researchers to analyze mobility patterns at unprecedented scales and granularities that were not possible with survey-based methodologies. With the appropriate set of statistical analysis tools and frameworks, human mobility data acts as a crucial intermediary, translating complex human behaviors into invaluable insights that drive data-driven policymaking strategies.

This dissertation presents two studies on the employment of human mobility analysis in addressing societal issues by constructing dynamic multivariate mobility networks in different urban scales. To predict the financial performance of local businesses, temporal customer co-location networks are created by harvesting credit card transaction data, in which network-based features constitute the backbone of the proposed method. Secondly, to depict the impact of the COVID-19 pandemic on different sociodemographic groups and analyze the visitation patterns under different hypothetical points of interest densities, another large-scale smartphone mobility data is used to construct dynamic mobility networks between neighborhoods (i.e., census block groups). The results of the conducted studies provide actionable insights for policymakers, public administrations, and governments as data-driven policymaking becomes increasingly more significant in our world.

6.1 Summary and Discussion of the Contributions

6.1.1 Customer Co-Location Networks for Financial Performance Prediction

Extracting economic and business insights with the help of human mobility data may have drastic implications for businesses in urban areas and hence their home societies. I studied the use of customer co-location networks between businesses in an urban area for their financial performance prediction. The results of the study mark the importance of customer mobility patterns for the financial performance of businesses. Here, I summarize the conducted analysis of business financial performance prediction with customer co-location networks.

- **Background:** Financial institutions require predictive models to periodically assess the financial well-being of businesses, considering their need for loans to sustain their economic vitality. Existing predictive methods rely on internal financial indicators, which are difficult to standardize and raise significant privacy concerns in terms of data sharing with third-party entities.
- **Co-Location Networks:** In this study, human mobility data harvested from credit card transactions from an OECD country are employed to construct customer co-location networks, in which customer mobility patterns are conceptualized as a proxy for economic activity.
- **Financial Performance Definition:** As a novel method, the financial performance of a business is defined as a function of three variables, namely the revenue, the number of unique customers, and the number of total transactions in the observed time frame.
- **Network-Based Features:** From the constructed customer co-location networks, I first construct network-based features by combining the degree, betweenness, closeness, and eigenvector centrality metrics. In addition, businesses' node embeddings from the co-location networks are extracted with node2vec.
- **Privacy Implications:** The results of the predictive analysis highlight that the proposed network-based and node2vec features perform on par with revenue and demographic-based features. However, the proposed set of features prioritizes privacy, providing a higher level of safeguard for the sensitive financial data belonging to both businesses and customers.

The conducted study mainly targets economic & business insights based on large-scale human mobility data. I studied the role of customer visitation patterns between businesses in a certain urban area to predict future financial performance. Human mobility-based financial performance prediction studies exist in the literature, in which individual-level spatial mobility metrics constitute the backbone of the predictive models. In the proposed study, I explore the use of collective customer mobility patterns to unveil the future financial performance of businesses in a mobility network setting, which in turn enables a privacy enhancement for the resulting financial performance features.

This emphasis on privacy enhancement allows for a secure and effective information-sharing scheme with third-party organizations. Considering the rapid development of financial systems, data sharing between financial organizations fosters the evolution of a more interconnected and responsive ecosystem. Based on the evaluation results of the proposed network-based and node2vec features, a higher level of privacy-preserving data sharing between organizations is enabled.

6.1.1.1 Limitations

However, our approach has several limitations concerning the utilized data source and the target businesses. Below, I summarize these limitations regarding the conducted study.

- The employed credit card transaction data is obtained from a single financial institution that operates in one of the OECD countries. The customer co-location networks are constructed based on credit card transactions that take place inside the businesses. However, the employed data does not capture cash transactions, which could present challenges, particularly in countries with lower financial transparency and for businesses heavily reliant on cash transactions. Considering the scale of the data that presents over 2 million transactions, it is arguable that the amount of cash transactions is negligible. Moreover, a significant correlation between cash and credit card transactions is observed in developed economies [34].

- The scope of the study does not consider the size of businesses. Small and medium-sized businesses contribute substantially to local economies. In the conducted study, the business selection procedure does not consider the size of the businesses as a filtering criterion. To this end, a model specifically tailored for small and medium-sized businesses would probably enhance the financial performance prediction.

As a future direction, smartphone data serves as a promising data source as an alternative to financial transaction data when constructing mobility networks due to their rich and nuanced insights into human movement patterns. While financial data primarily tracks monetary transactions, smartphone data offers a more direct lens into individual mobility by recording nuanced information that is not limited by business visits.

The privacy implications of the conducted study require further validation. Network-based and node2vec features provide a higher level of safeguard against network reconstruction attacks that may pose a serious threat to financial institutions considering the gravity of financial records. In future work, the relationship between predictive performance and increased privacy levels can be analyzed.

6.1.2 Using Mobility Networks to Unravel the Effects of COVID-19 Pandemic

Considering the monumental impact of the COVID-19 pandemic, prioritizing research efforts for informed policymaking has become crucial. Public administrations, decision-makers, and governments around the world turn their faces to data-driven policymaking to enhance the efficacy of their strategies and tailor interventions and policies to diverse sociodemographic groups. To this end, I used dynamic mobility networks to analyze the impact of the COVID-19 pandemic on different sociodemographic groups in New York City, US and performed mobility simulations under hypothetical POI densities. Here, I summarize the main points of the conducted study.

- **Constructing Dynamic Mobility Networks:** Using large-scale smartphone mobility data, I constructed dynamic mobility networks between census block groups in NYC that capture the flows of people in a weekly manner. Each node in the network, i.e., census block groups, is enriched with a set of sociodemographic features obtained from US census data.

- **Analyzing the Topological Changes:** To depict the impact of the imposed NPIs on different socioeconomic groups, I focused on the well-established network centrality metrics and analyzed how these metrics changed in different socioeconomic groups represented by their quartiles, which concludes that census block groups that changed their ego-network structure the least had higher COVID-19 infection rates.

- **Dissimilarity Analysis:** To quantify the mobility change observed in sociodemographic groups, I computed node-level dissimilarity scores between

aligned weekly networks and analyzed how frequently census block groups appear in the top and bottom dissimilarity quartiles. The sociodemographic traits that frequently appear in the top and bottom dissimilarity quartiles are analyzed, highlighting the adaptability of high-income and high-white subpopulation groups.

- **Bridge Neighborhoods:** In order to identify the neighborhoods, i.e., census block groups, that act as a mediator for the infection to be spread across New York City, I define *bridge* neighborhoods. In this context, bridge neighborhoods are identified as the nodes in the network that frequently appear in the neighborhood of COVID-19 hotpots during the virus incubation period. This analysis highlights the significance of Staten Island in New York City, where distinct mobility patterns are observed in contrast to the boroughs with similar sociodemographic traits.

- **Hypothetical Scenario Analysis:** Considering the number of available POIs in urban areas as one of the drivers for human mobility, I perform a simple simulation for Staten Island neighborhoods and unravel how mobility flows would look under varying POI densities. The results obtained from the hypothetical scenario analysis highlight a clear correlation stating that improved access to essential amenities lowers infection rates.

6.1.2.1 Limitations

The study sheds light on the determinants, such as geographic constraints, that contribute to both the resilience and adaptability of neighborhoods. By analyzing these factors, urban planners and public administration offices can formulate more effective and sustainable policies, enabling better-informed intervention strategies. This understanding prepares them to react more adeptly to unforeseen crises, such as the COVID-19 pandemic, potentially saving more lives through proactive measures and timely interventions.

However, there exist several limitations concerning the conducted study. Firstly, Safegraph only provides visit counts between census block groups and POIs but does not address the mode of transportation. With such information at hand, further analyses could have been possible to discern the role of transportation in the spread of the virus.

A significant portion of the utilized POIs consist of stores such as grocery stores and restaurants. Due to the low coverage of workplaces and offices, the identification of essential trips, e.g., commuting, is out of the reach of the employed gravity model.

Finally, a significant implication of the study is that enhanced access to essential amenities serves as a preventive measure in curbing the spread of the virus. However,

the realization of this utopian vision hinges on private developers stepping in to offer these essential services within economically disadvantaged urban areas. While various solutions exist to address this issue, including governmental policies such as tax exemption zones aimed at incentivizing public sector involvement in commercial activities, my primary focus here is to conduct a direct assessment of hypothetical scenarios. This evaluation aims to gauge the potential scale of their impact on residents' mobility patterns without diversions into broader policy discussions.

Appendix A

Appendix: Local Business Performance Prediction with Co-Location Networks

MCC	Transaction Count	Business Count	Description
5411	691,957	10,376	Grocery Stores, Supermarkets
5541	382,316	4,413	Service Stations (with or without ancillary services)
5691	331,304	8,619	Men's and Women's Clothing Stores
5812	30,727	3,529	Eating places and Restaurants, and Fast Food places
5499	62,846	3,445	Misc. Food Stores, Convenience Stores, Specialty Markets
5045	68,482	548	Computers, Computer Peripheral Equipment, Software
5732	68,372	1,512	Electronic Equipment Sales
5977	42,249	1,691	Cosmetic Stores
5200	30,360	772	Home Supply Warehouse Stores
5661	25,533	2,957	Shoe Stores
5941	25,214	1,614	Sporting Goods Stores
5712	17,041	2,384	Furniture, Home Furnishings, and Equipment Stores
5942	14,046	550	Book Stores, Books, Periodicals, and Newspapers
5641	12,799	924	Children's and Infant's Wear Stores
5945	7,246	367	Hobby, Toy, and Game Shops
5992	5,889	88	Florists
5999	5,543	718	Miscellaneous and Specialty Retail Stores
5722	3,848	3,654	Household Appliance Stores
5621	3,671	1,558	Women's Ready-to-Wear Stores
5950	3,150	1,034	Glassware/Crystal Stores

Table A.1: Top Twenty Most Visited Business Categories Ranked by Transaction Counts in the Dataset.

Attribute	**Categories and their distribution in data**	**Type**
Customer ID	Unique hashed ID for each customer	Categorical ID number
Age	min = 19, max = 85, mean = 38.5, median = 37, IQR = 14	Numerical
Gender	Female (31.3%), Male (58.7%)	Categorical
Marital status	Single (21.4%), Married (70.8%), Divorced (4.3%), Dul (0.5%), Unknown (2.9%)	Categorical
Education level	Unknown (0.004%), Uneducated (1.2%), Elementary school (6.8%), Middle school (8.3%), High school (8.4%), Associate Degree (45.2%), Bachelor's (29.8%), Master's (2.9%), Doctoral degree (0.2%)	Categorical
Employment status	Unknown (0.1%), Unemployed (0.8%), Under work age (0.001%), Abroad (0.001%), Student (0.05%), Housewife (1.16%), Public sector (6.4%), Private sector (73.6%), Freelancer (9.8%), Retired (public sector) (4.9%), Retired (private sector) (1.93%), Retired (freelancer) (0.5%)	Categorical
Income	Monthly income as estimated by the bank	Numerical
Home district ID	Home location district identification number	Categorical ID number
Work district ID	Workplace district identification number	Categorical ID number

Table A.2: Summary of information for the customer attributes.

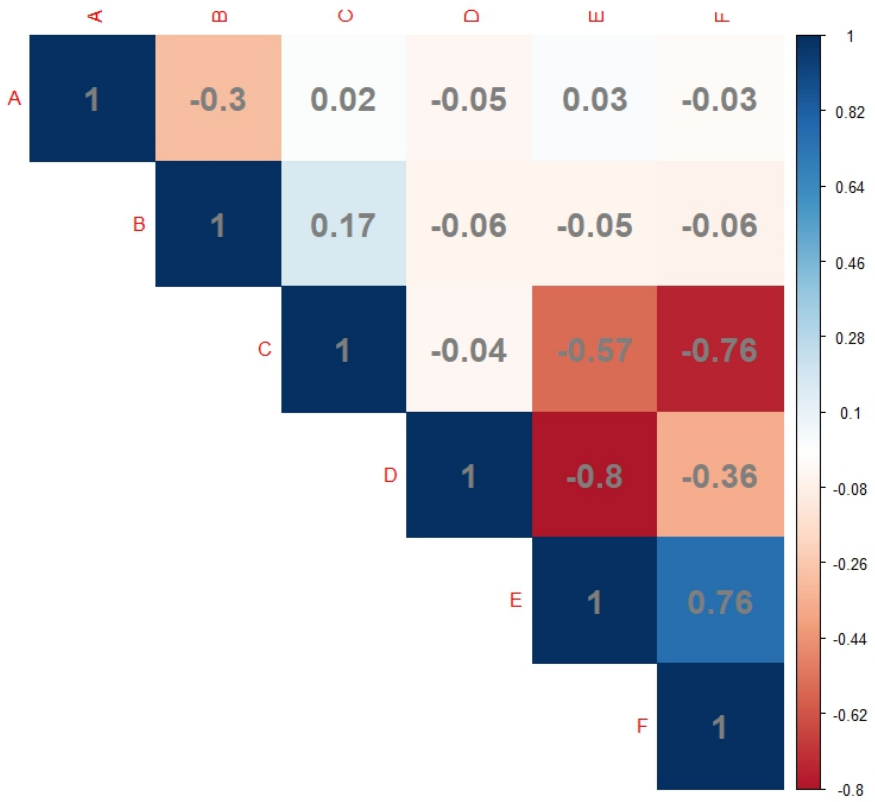

Figure A.1: Correlation coefficients between label percentages and their ratio with district population and monthly average household income.

Feature Name	**Alphabet Code**
business district population	A
business district average household income	B
percentage labeled as poorly-performing	C
percentage labeled as medium-performing	D
percentage labeled as well-performing	E
well-performing to poorly-performing class percentage ratio	F

Table A.3: Alphabet codes and corresponding feature names in correlation table of Figure A.1

Customer Feature	Group	Percentage of All	Percentage Edges Created	Median Edge Created
Income	Above median	54.2%	60.0%	6
	Below median	45.8%	40.0%	6
Age	Above median	53.6%	56.2%	6
	Below median	46.4%	43.8%	6
Number of transactions	Above median	51.1%	85.0%	15
	Below median	48.9%	15.0%	3
Average spent	Above median	50.0%	42.5%	6
	Below median	50.0%	57.5%	10
Education	University degree	29.7%	37.1%	10
	Below	70.3%	62.9%	6
Gender	Female	32.2%	34.0%	6
	Male	67.8%	66.0%	6
Marital status	Married	72.9%	76.0%	6
	Not married	27.1%	24.0%	6

Table A.4: Percentage of the co-location network edges created by each feature group of customers, along with the median edge created by each member within each group.

	Dependent variable:
	Num. of Edges Created
Education	0.410^{***}
	(0.003)
Gender	0.060^{***}
	(0.003)
log(Income)	-0.004^{***}
	(0.0005)
Age	0.006^{***}
	(0.0002)
Marital Status	0.281^{***}
	(0.003)
log(Mean Transaction Amount)	-0.308^{***}
	(0.002)
Constant	3.428^{***}
	(0.010)
Observations	38,192
Log Likelihood	−638,292.000
Akaike Inf. Crit.	1,276,598.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A.5: Poisson regression analysis on the number of edges created by each customer.

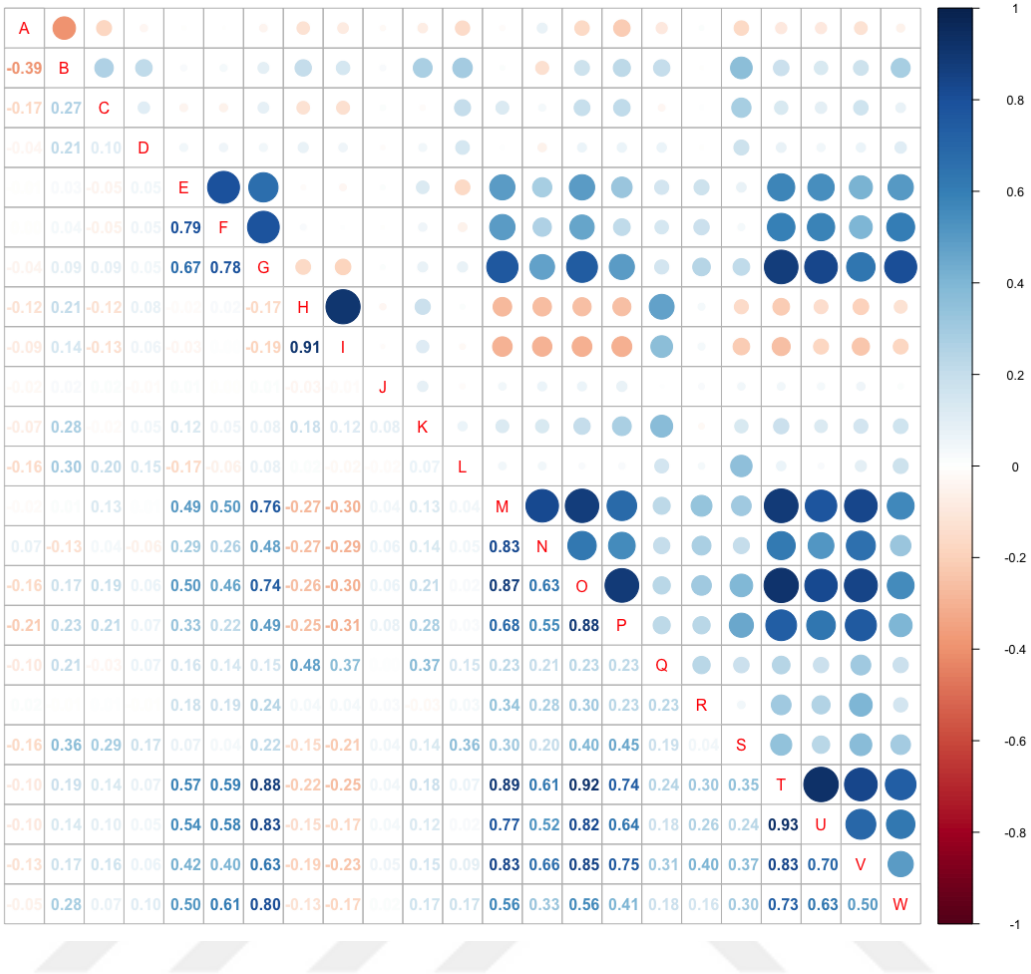

Figure A.2: Correlation coefficient values for all pairs of features computed in the conducted study.

Feature Name	**Alphabet code**	**Feature Name**	**Alphabet code**
Business district population	A	customer age mean	H
Business district average household income	B	customer age median	I
Business in buffer POI count	C	customer income mean	J
business in buffer POI diversity	D	customer income median	K
period revenue	E	customer gender entropy	L
period transaction count	F	number of customer workplace districts	M
period distinct customers count	G	customer workplace district entropy	N
customer home districts count	O	customer home district entropy	P
customer job entropy	Q	customer education entropy	R
customer marital status entropy	S	business node degree	T
business node betweenness	U	business node closeness	V
business node eigenvector centrality	W		

Table A.6: Alphabet codes and corresponding feature names in the correlation table of Supplementary Figure A.2.

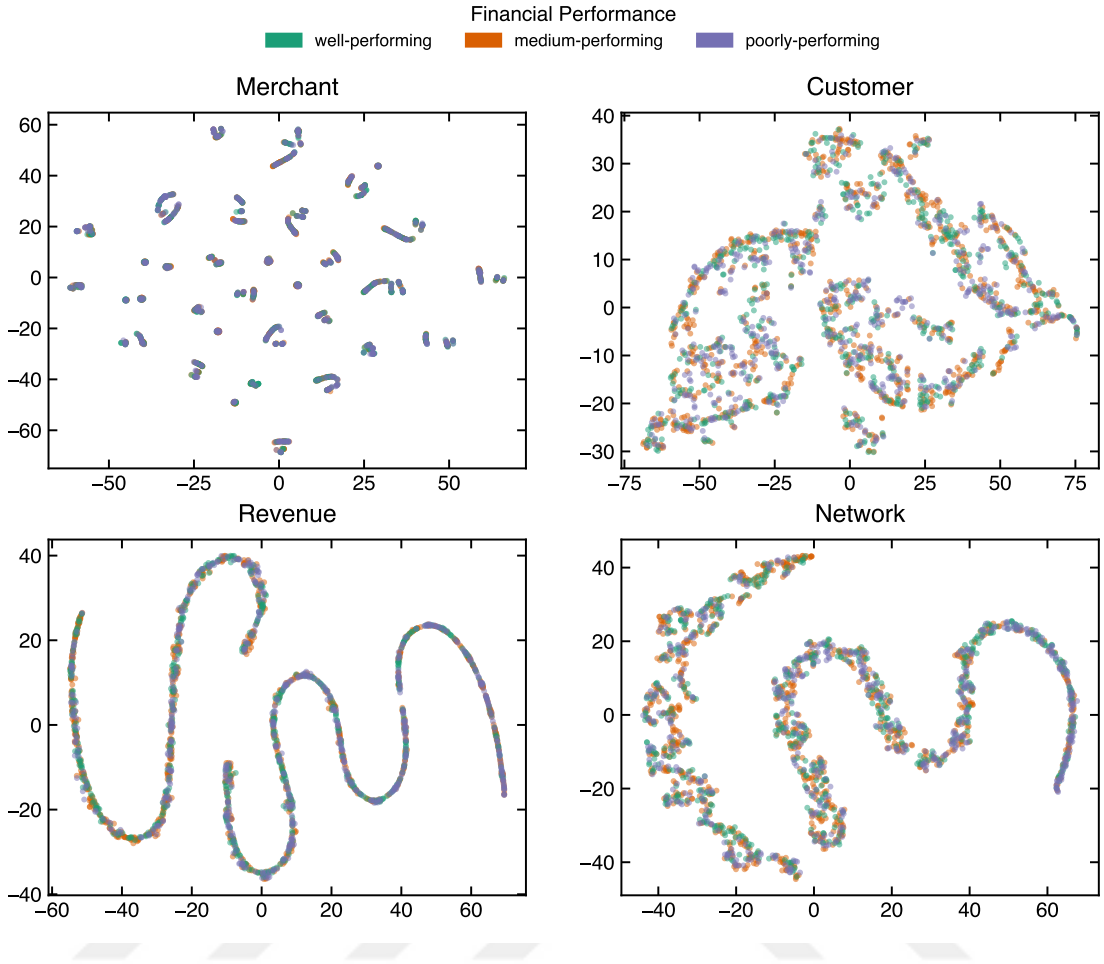

Figure A.3: t-SNE embeddings of the business, customer, revenue, and network features colored by financial performance labels.

	Dependent variable:
	Is business Well-Performing
log(Revenue)	-0.360^{***}
	(0.089)
log(POI Count)	0.040
	(0.056)
POI Diversity	0.065
	(0.162)
Customer Median Age	0.028
	(0.025)
log(Customer Median Income)	0.396
	(0.300)
Customer Gender Entropy	0.359
	(0.526)
Customer Job Type Entropy	−0.390
	(0.273)
Customer Education Entropy	0.311
	(0.340)
Customer Marital Status Entropy	-0.834^{***}
	(0.319)
Business Ego-Net MCC Entropy	0.016
	(0.202)
Business Ego-Net GEO Entropy	−0.001
	(0.062)
Centrality PCA	−2.047
	(2.620)
Observations	1,977
Log Likelihood	−1,144.396
Akaike Inf. Crit.	2,382.792
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A.7: Logistic Regression analysis on well-performing Businesses.

Appendix B

Appendix: Neighborhood Adaptability Indicators During the COVID-19 Pandemic

	Dependent variable: Occurrence Frequency
	(1)
Bronx	0.631***
	(0.008)
Brooklyn	0.447***
	(0.006)
Manhattan	0.275***
	(0.007)
Queens	0.555***
	(0.006)
Staten Island	0.901***
	(0.014)
Observations	6139
R^2	0.261
Adjusted R^2	0.261
Residual Std. Error	0.248 (df=6134)
F Statistic	542.452*** (df=4; 6134)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.1: Regression analysis on bridge CBG occurrences with respect to boroughs.

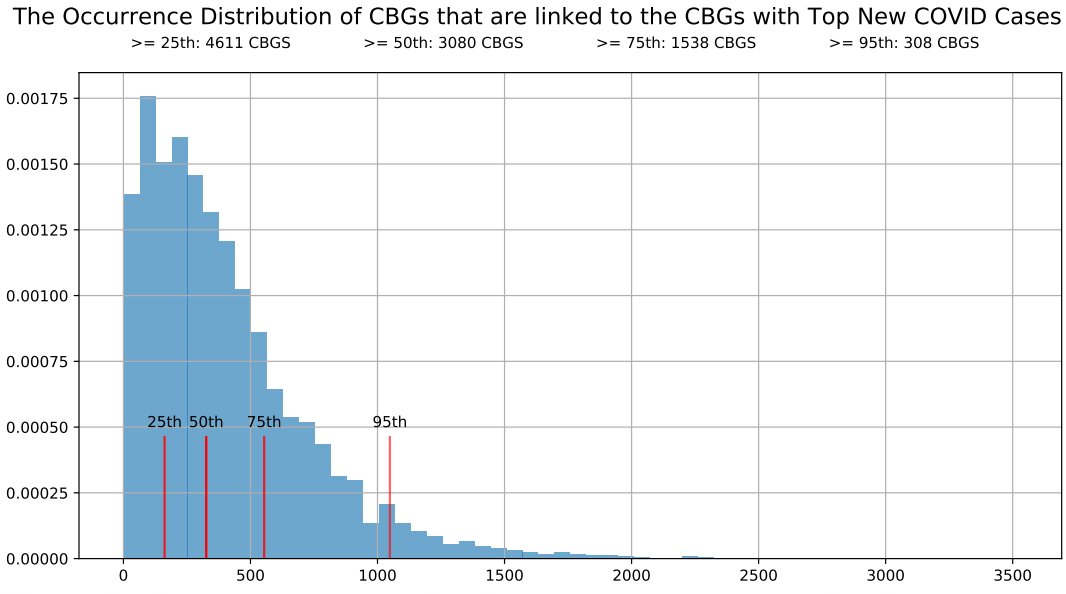

Figure B.1: Bridge occurrence distribution with percentiles. CBGs are ranked with respect to how frequently they appear in the neighborhood of COVID hotspots in a weekly manner. We use occurrence percentiles and consider the ones above the 75^{th} percentile as the final bridge CBGs group.

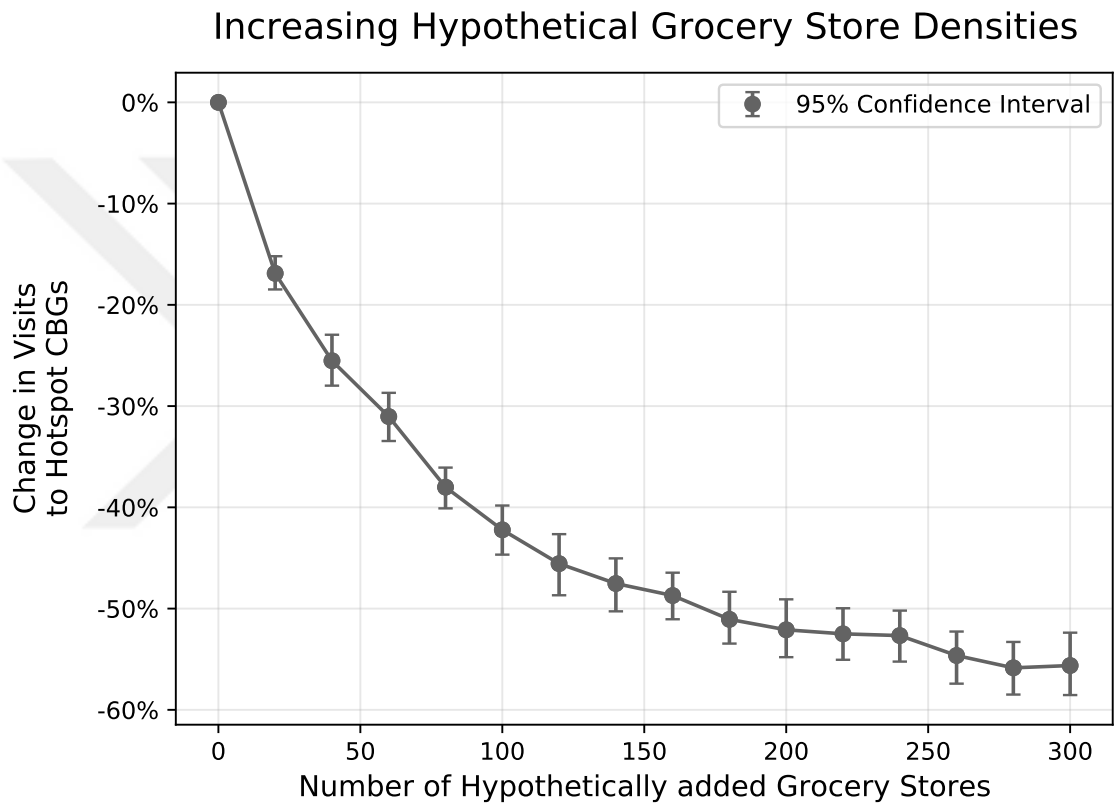

Figure B.2: Change in visits to hotspot CBGs in Staten Island with different number of hypothetical POI additions. In contrast to the POI area expansion, the addition of POIs displays a rapid decrease in visits to hotpot CBGs.

Bibliography

[1] 68% of the world population projected to live in urban areas by 2050, says un. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. Accessed: 2023-12-22.

[2] Google COVID-19 Community Mobility Reports. Available online at: https://www.google.com/covid19/mobility/, last accessed on 2023-11-23.

[3] ISO 18245 Merchant Codes. https://www.iso.org/standard/33365.html. Accessed: 2023-01-14.

[4] Organisation for Economic Co-operation and Development. https://www.oecd.org/. Accessed: 2023-01-14.

[5] Abideen, Z. U., Sun, H., Yang, Z., Ahmad, R. Z., Iftekhar, A., and Ali, A. Deep wide spatial-temporal based transformer networks modeling for the next destination according to the taxi driver behavior prediction. *Applied Sciences 11*, 1 (2020), 17.

[6] Agarwal, R. R., Lin, C.-C., Chen, K.-T., and Singh, V. K. Predicting financial trouble using call data—on social capital, phone logs, and financial trouble. *PloS one 13*, 2 (2018), e0191863.

[7] Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*, 7903 (2022), 864–870.

[8] Alessandretti, L., Sapiezynski, P., Lehmann, S., and Baronchelli, A. Multi-scale spatio-temporal analysis of human mobility. *PloS one 12*, 2 (2017), e0171686.

[9] Alis, C., Legara, E. F., and Monterola, C. Generalized radiation model for human migration. *Scientific reports 11*, 1 (2021), 22707.

[10] Alpaydin, E. *Introduction to machine learning*. MIT press, 2020.

[11] Alvarez-Rodriguez, U., Battiston, F., de Arruda, G. F., Moreno, Y., Perc, M., and Latora, V. Evolutionary dynamics of higher-order interactions in social networks. *Nature Human Behaviour 5*, 5 (2021), 586–595.

[12] Anderson, E., Lin, S., Simester, D., and Tucker, C. Harbingers of failure. *Journal of Marketing Research 52*, 5 (2015), 580–592.

[13] Badr, H. S., Zaitchik, B. F., Kerr, G. H., Nguyen, N.-L. H., Chen, Y.-T., Hinson, P., Colston, J. M., Kosek, M. N., Dong, E., Du, H., et al. Unified real-time environmental-epidemiological data for multiscale modeling of the covid-19 pandemic. *Scientific Data 10*, 1 (2023), 367.

[14] Bahrami, M., Boz, H. A., Suhara, Y., Balcisoy, S., Bozkaya, B., and Pentland, A. Predicting merchant future performance using privacy-safe network-based features. *Scientific Reports 13*, 1 (2023), 10073.

[15] Bahrami, M., Xu, Y., Tweed, M., Bozkaya, B., et al. Using gravity model to make store closing decisions: A data driven approach. *Expert systems with applications 205* (2022), 117703.

[16] Ball, P. *Why society is a complex matter: Meeting twenty-first century challenges with a new kind of science*. Springer Science & Business Media, 2012.

[17] Bao, Y., Huang, Z., Li, L., Wang, Y., and Liu, Y. A bilstm-cnn model for predicting users' next locations based on geotagged social media. *International Journal of Geographical Information Science 35*, 4 (2021), 639–660.

[18] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. Human mobility: Models and applications. *Physics Reports 734* (2018), 1–74.

[19] Barbosa, H., de Lima-Neto, F. B., Evsukoff, A., and Menezes, R. The effect of recency to human mobility. *EPJ Data Science 4* (2015), 1–14.

[20] Berger, A. N., and Frame, W. S. Small business credit scoring and credit availability. *Journal of small business management 45*, 1 (2007), 5–22.

[21] Berger, A. N., and Frame, W. S. Small business credit scoring and credit availability*. *Journal of Small Business Management 45*, 1 (2007), 5–22.

[22] Berger-Schmitt, R. *Social cohesion as an aspect of the quality of societies: Concept and measurement*, vol. 14. ZUMA Mannheim, 2000.

[23] Berke, A., Doorley, R., Larson, K., and Moro, E. Generating synthetic mobility data for a realistic population with rnns to improve utility and privacy. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (2022), pp. 964–967.

[24] Berlingerio, M., Koutra, D., Eliassi-Rad, T., and Faloutsos, C. Netsimile: A scalable approach to size-independent network similarity. *arXiv preprint arXiv:1209.2684* (2012).

[25] Bi, C., Pan, G., Yang, L., Lin, C.-C., Hou, M., and Huang, Y. Evacuation route recommendation using auto-encoder and markov decision process. *Applied Soft Computing 84* (2019), 105741.

[26] Bianconi, G., Pin, P., and Marsili, M. Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Sciences 106*, 28 (2009), 11433–11438.

[27] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. Complex networks: Structure and dynamics. *Physics reports 424*, 4-5 (2006), 175–308.

[28] Brockmann, D., Hufnagel, L., and Geisel, T. The scaling laws of human travel. *Nature 439*, 7075 (2006), 462–465.

[29] Caminha, C., Furtado, V., Pequeno, T. H., Ponte, C., Melo, H. P., Oliveira, E. A., and Andrade Jr, J. S. Human mobility in large cities as a proxy for crime. *PloS one 12*, 2 (2017), e0171609.

[30] Carroll Jr, J. D., and Bovls, H. W. Predicting local travel in urban regions. *Papers in Regional Science 3*, 1 (1957), 183–197.

[31] Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. Mobility network models of covid-19 explain inequities and inform reopening. *Nature 589*, 7840 (2021), 82–87.

[32] Chang, S., Vrabac, D., Leskovec, J., and Ugander, J. Estimating geographic spillover effects of covid-19 policies from large-scale mobility networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 14161–14169.

[33] Chen, Y., Long, C., Cong, G., and Li, C. Context-aware deep model for joint mobility and time prediction. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (2020), pp. 106–114.

[34] Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., et al. The economic impacts of COVID-19: Evidence from a new public database built using private sector data. *National Bureau of Economic Research* (2020).

[35] Chi, G., and Meng, B. Debt rating model based on default identification: Empirical evidence from chinese small industrial enterprises. *Management Decision* (2018).

[36] Chong, S. K., Bahrami, M., Chen, H., Balcisoy, S., Bozkaya, B., et al. Economic outcomes predicted by diversity in cities. *EPJ Data Science 9*, 1 (2020), 17.

[37] Christopoulos, A. G., Dokas, I. G., Kalantonis, P., and Koukkou, T. Investigation of financial distress with a dynamic logit based on the linkage between liquidity and profitability status of listed firms. *Journal of the Operational Research Society 70*, 10 (2019), 1817–1829.

[38] Ciampi, F., and Gordini, N. Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of italian small enterprises. *Journal of Small Business Management 51*, 1 (2013), 23–45.

[39] Cippà, P. E., Cugnata, F., Ferrari, P., Brombin, C., Ruinelli, L., Bianchi, G., Beria, N., Schulz, L., Bernasconi, E., Merlani, P., et al. A data-driven approach to identify risk profiles and protective drugs in covid-19. *Proceedings of the National Academy of Sciences 118*, 1 (2021), e2016877118.

[40] Colombo, G. B., Chorley, M. J., Williams, M. J., Allen, S. M., and Whitaker, R. M. You are where you eat: Foursquare checkins as indicators of human mobility and behaviour. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops* (2012), pp. 217–222.

[41] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[42] Cull, W. L., O'connor, K. G., Sharp, S., and Tang, S.-f. S. Response rates and response bias for 50 surveys of pediatricians. *Health services research 40*, 1 (2005), 213–226.

[43] Dai, G., Hu, X., Ge, Y., Ning, Z., and Liu, Y. Attention based simplified deep residual network for citywide crowd flows prediction. *Frontiers of Computer Science 15* (2021), 1–12.

[44] Dalziel, B. D., Pourbohloul, B., and Ellner, S. P. Human mobility patterns predict divergent epidemic dynamics among cities. *Proceedings of the Royal Society B: Biological Sciences 280*, 1766 (2013), 20130763.

[45] Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L., and Wang, D. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences 113*, 26 (2016), 7047–7052.

[46] Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., Vaitla, B., and González, M. C. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature communications 9*, 1 (2018), 3330.

[47] Dong, X., Suhara, Y., Bozkaya, B., Singh, V. K., Lepri, B., and Pentland, A. S. Social bridges in urban purchase behavior. *ACM Trans. Intell. Syst. Technol. 9*, 3 (2017).

[48] Doyle, D. P. Data-driven decision-making: Is it the mantra of the month or does it have staying power. *The Journal 30*, 10 (2003), 19–21.

[49] Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science 328*, 5981 (2010), 1029–1031.

[50] Fagiolo, G., and Santoni, G. Human-mobility networks, country income, and labor productivity. *Network Science 3*, 3 (2015), 377–407.

[51] Fan, J., and Stewart, K. Understanding collective human movement dynamics during large-scale events using big geosocial data analytics. *Computers, Environment and Urban Systems 87* (2021), 101605.

[52] Fan, Z., Su, T., Sun, M., Noyman, A., Zhang, F., Pentland, A., and Moro, E. Diversity beyond density: Experienced social mixing of urban streets. *PNAS nexus 2*, 4 (2023), pgad077.

[53] Fantazzini, D., and Figini, S. Random survival forests models for sme credit risk measurement. *Methodology and computing in applied probability 11*, 1 (2009), 29–45.

[54] Fernandes, G. B., and Artes, R. Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research 249*, 2 (2016), 517–524.

[55] Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research 210*, 2 (2011), 368–378.

[56] Florida, R. *The new urban crisis: How our cities are increasing inequality, deepening segregation, and failing the middle class-and what we can do about it*. Hachette UK, 2017.

[57] Forghani, M., Karimipour, F., and Claramunt, C. From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. *Transportation Research Part C: Emerging Technologies 117* (2020), 102666.

[58] Frias-Martinez, V., Virseda-Jerez, J., and Frias-Martinez, E. On the relation between socio-economic status and physical mobility. *Information Technology for Development 18*, 2 (2012), 91–106.

[59] Galeazzi, A., Cinelli, M., Bonaccorsi, G., Pierri, F., Schmidt, A. L., Scala, A., Pammolli, F., and Quattrociocchi, W. Human mobility in response to covid-19 in france, italy and uk. *Scientific reports 11*, 1 (2021), 13141.

[60] Gallucci, C., Santullli, R., Modina, M., and Formisano, V. Financial ratios, corporate governance and bank-firm information: a bayesian approach to predict smes' default. *Journal of Management and Governance* (2022), 1–20.

[61] Golbeck, J. *Analyzing the social web*. Newnes, 2013.

[62] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *nature 453*, 7196 (2008), 779–782.

[63] Granovetter, M. The impact of social structure on economic outcomes. *Journal of economic perspectives 19*, 1 (2005), 33–50.

[64] Grover, A., and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), pp. 855–864.

[65] Gu, W., Tandon, A., Ahn, Y.-Y., and Radicchi, F. Defining and identifying the optimal embedding dimension of networks. *Preprint at https://arxiv.org/abs/2004.09928* (2020).

[66] Han, S. Y., Tsou, M.-H., Knaap, E., Rey, S., and Cao, G. How do cities flow in an emergency? tracing human mobility patterns during a natural

disaster with big data and geospatial data science. *Urban Science 3*, 2 (2019), 51.

[67] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[68] Hillier, B., Turner, A., Yang, T., and Park, H.-T. Metric and topo-geometric properties of urban street networks: some convergences, divergences and new results. *Journal of Space Syntax Studies* (2009).

[69] Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (1995), vol. 1, IEEE, pp. 278–282.

[70] Hong, L., and Frias-Martinez, V. Modeling and predicting evacuation flows during hurricane irma. *EPJ Data Science 9*, 1 (2020), 29.

[71] Hossmann, T., Spyropoulos, T., and Legendre, F. A complex network analysis of human mobility. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)* (2011), pp. 876–881.

[72] Hu, T., Wang, S., She, B., Zhang, M., Huang, X., Cui, Y., Khuri, J., Hu, Y., Fu, X., Wang, X., et al. Human mobility data in the covid-19 pandemic: characteristics, applications, and challenges. *International Journal of Digital Earth 14*, 9 (2021), 1126–1147.

[73] Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems 37*, 4 (2004), 543–558.

[74] Huff, D., and McCALLUM, B. M. Calibrating the Huff model using arcgis business analyst. *ESRI White Paper* (2008), 1–33.

[75] Huff, D. L. Defining and estimating a trading area. *Journal of Marketing 28*, 3 (1964), 34–38.

[76] Hunter, R. F., Garcia, L., de Sa, T. H., Zapata-Diomedi, B., Millett, C., Woodcock, J., Pentland, A., and Moro, E. Effect of covid-19 response policies on walking behavior in us cities. *Nature Communications 12*, 1 (2021), 1–9.

[77] Iacus, S. M., Santamaria, C., Sermi, F., Spyratos, S., Tarchi, D., and Vespe, M. Human mobility and covid-19 initial dynamics. *Nonlinear Dynamics 101*, 3 (2020), 1901–1919.

[78] Jahromi, K. K., Zignani, M., Gaito, S., and Rossi, G. P. Simulating human mobility patterns in urban areas. *Simulation Modelling Practice and Theory 62* (2016), 137–156.

[79] Jiang, J., Chen, J., Tu, W., and Wang, C. A novel effective indicator of weighted inter-city human mobility networks to estimate economic development. *Sustainability 11*, 22 (2019), 6348.

[80] Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., and González, M. C. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences 113*, 37 (2016), E5370–E5378.

[81] Jones, S., Johnstone, D., and Wilson, R. Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting 44*, 1-2 (2017), 3–34.

[82] Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. Understanding human mobility from twitter. *PloS one 10*, 7 (2015), e0131469.

[83] Kaya, E., Alpan, E., Balcisoy, S., and Bozkaya, B. Quantifying insurance agency channel dynamics using premium sales big data and external factors. *Big Data 9*, 2 (2021), 116–131.

[84] Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., et al. Behavioral attributes and financial churn prediction. *EPJ Data Science 7*, 1 (2018), 41.

[85] Kim, S. Y., and Upneja, A. Majority voting ensemble with a decision trees for business failure prediction during economic downturns. *Journal of Innovation & Knowledge 6*, 2 (2021), 112–123.

[86] Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Group†, O. C.-. D. W., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., and Scarpino, S. V. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science 368*, 6490 (2020), 493–497.

[87] Kroll, C., and Delhey, J. A happy nation? opportunities and challenges of using subjective indicators in policymaking. *Social Indicators Research 114* (2013), 13–28.

[88] Lance, G. N., and Williams, W. T. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal 1*, 1 (1967), 15–20.

[89] Lazer, D., Brewer, D., Christakis, N., Fowler, J., and King, G. Life in the network: the coming age of computational social. *Science 323*, 5915 (2009), 721–723.

[90] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. Computational social science. *Science 323*, 5915 (2009), 721–723.

[91] Leković, B., and Marić, S. M. Measures of small business success/performance–importance, reliability and usability. *Industrija 43*, 2 (2015).

[92] Li, B., and Mostafavi, A. Location intelligence reveals the extent, timing, and spatial variation of hurricane preparedness. *Scientific reports 12*, 1 (2022), 16121.

[93] Li, D., and Liu, J. Uncovering the relationship between point-of-interests-related human mobility and socioeconomic status. *Telematics and Informatics 39* (2019), 49–63.

[94] Li, W., Tao, W., Qiu, J., Liu, X., Zhou, X., and Pan, Z. Densely connected convolutional networks with attention lstm for crowd flows prediction. *IEEE Access 7* (2019), 140488–140498.

[95] Liang, Y., Gao, S., Cai, Y., Foutz, N. Z., and Wu, L. Calibrating the dynamic Huff model for business analysis using location big data. *Transactions in GIS 24*, 3 (2020), 681–703.

[96] Liu, Y., Liu, C., Lu, X., Teng, M., Zhu, H., and Xiong, H. Point-of-interest demand modeling with human mobility patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (2017), pp. 947–955.

[97] Loaiza-Monsalve, D., and Riascos, A. Human mobility in bike-sharing systems: Structure of local and non-local dynamics. *PLoS One 14*, 3 (2019), e0213106.

[98] Long, X., Jin, L., and Joshi, J. Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (2012), pp. 927–934.

[99] Lucchini, L., Centellegher, S., Pappalardo, L., Gallotti, R., Privitera, F., Lepri, B., and De Nadai, M. Living in a pandemic: changes in mobility routines, social activity and adherence to covid-19 protective measures. *Scientific reports 11*, 1 (2021), 1–12.

[100] Mahmud, J., Nichols, J., and Drews, C. Home location identification of twitter users. *arXiv preprint arXiv:1403.2345* (2014).

[101] Mariooryad, S., and Busso, C. The cost of dichotomizing continuous labels for binary classification problems: Deriving a bayesian-optimal classifier. *IEEE Transactions on Affective Computing 8*, 1 (2015), 119–130.

[102] Miritello, G., Moro, E., and Lara, R. Dynamical strength of social ties in information spreading. *Physical Review E 83*, 4 (2011), 045102.

[103] Morales, A. J., Dong, X., Bar-Yam, Y., and 'Sandy'Pentland, A. Segregation and polarization in urban areas. *Royal Society Open Science 6*, 10 (2019), 190573.

[104] Moro, E., Calacci, D., Dong, X., and Pentland, A. Mobility patterns are associated with experienced income segregation in large us cities. *Nature communications 12*, 1 (2021), 4633.

[105] Morrison, P. S., Rossouw, S., and Greyling, T. The impact of exogenous shocks on national wellbeing. new zealanders' reaction to covid-19. *Applied Research in Quality of Life 17*, 3 (2022), 1787–1812.

[106] Muniz-Rodriguez, K., Ofori, S. K., Bayliss, L. C., Schwind, J. S., Diallo, K., Liu, M., Yin, J., Chowell, G., and Fung, I. C.-H. Social media use in emergency response to natural disasters: a systematic review with a public health perspective. *Disaster medicine and public health preparedness 14*, 1 (2020), 139–149.

[107] Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y., and Ryu, K. H. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability 11*, 3 (2019).

[108] Najarsadeghi, M., and Dorostkar, E. How do measure the triangle of human mobility in urban nightlife? *Cities 130* (2022), 103944.

[109] Netto, C. F. S., Bahrami, M., Brei, V. A., Bozkaya, B., Balcisoy, S., and Pentland, A. P. Disaggregating sales prediction: A gravitational approach. *Expert Systems with Applications* (2023), 119565.

[110] Nilforoshan, H., Looi, W., Pierson, E., Villanueva, B., Fishman, N., Chen, Y., Sholar, J., Redbird, B., Grusky, D., and Leskovec, J. Human mobility networks reveal increased segregation in large cities. *Nature* (2023), 1–7.

[111] Niu, H., Liu, J., Fu, Y., Liu, Y., and Lang, B. Exploiting human mobility patterns for gas station site selection. In *Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I 21* (2016), Springer, pp. 242–257.

[112] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one 7*, 5 (2012), e37027.

[113] NYC Department of Transportation. New York City Mobility Report 2019. Available online at: `http://www.nyc.gov/html/dot/downloads/pdf/mobility-report-2019-print.pdf`, last accessed on 2023-11-23.

[114] NYC Planning. The Ins and Outs of NYC Commuting. Available online at: `https://www1.nyc.gov/assets/planning/download/pdf/planning-level/housing-economy/nyc-ins-and-out-of-commuting.pdf`, last accessed on 2023-11-23.

[115] Office of the New York City Comptroller. New York City's Frontline Workers. Available online at: `https://comptroller.nyc.gov/reports/new-york-citys-frontline-workers/`, last accessed on 2023-11-23.

[116] Orman, G. K., Türe, N., Balcisoy, S., and Boz, H. A. Finding proper time intervals for dynamic network extraction. *Journal of Statistical Mechanics: Theory and Experiment 2021*, 3 (2021), 033414.

[117] Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. Using big data to study the link between human mobility and socio-economic development. In *2015 IEEE International Conference on Big Data (Big Data)* (2015), pp. 871–878.

[118] Pappalardo, L., and Simini, F. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery 32*, 3 (2018), 787–829.

[119] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. Returners and explorers dichotomy in human mobility. *Nature communications 6*, 1 (2015), 8166.

[120] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W., and Yu, S. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks 8*, 5 (2022), 745–762.

[121] Pentland, A. *Social Physics: How social networks can make us smarter*. Penguin, 2015.

[122] Perc, M. Diffusion dynamics and information spreading in multilayer networks: An overview. *The European Physical Journal Special Topics 228*, 11 (2019), 2351–2355.

[123] Perc, M. The social physics collective. sci rep 9: 16549, 2019.

[124] Peters, D. H., Garg, A., Bloom, G., Walker, D. G., Brieger, W. R., and Hafizur Rahman, M. Poverty and access to health care in developing countries. *Annals of the new York Academy of Sciences 1136*, 1 (2008), 161–171.

[125] Pillemer, K., Wells, N. M., Wagenet, L. P., Meador, R. H., and Parise, J. T. Environmental sustainability in an aging society: a research agenda. *Journal of Aging and Health 23*, 3 (2011), 433–453.

[126] Plattner, D. Why firms go bankrupt. the influence of key financial figures and other factors on the insolvency probability of small and medium sized enterprises. *KfWResearch 28* (2002), 37–51.

[127] Reagans, R., and Zuckerman, E. W. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science 12*, 4 (2001), 502–517.

[128] Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., and Chong, S. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking 19*, 3 (2011), 630–643.

[129] Riascos, A., and Mateos, J. L. Networks and long-range mobility in cities: A study of more than one billion taxi trips in new york city. *Scientific Reports 10*, 1 (2020), 4022.

[130] Roxas, B., and Lindsay, V. Social desirability bias in survey research on sustainable development in small firms: An exploratory analysis of survey mode effect. *Business Strategy and the Environment 21*, 4 (2012), 223–235.

[131] SafeGraph. Weekly Patterns. Available online at: `https://docs.safegraph.com/docs/weekly-patterns`, last accessed on 2023-11-23.

[132] Sapiezynski, P., Stopczynski, A., Gatej, R., and Lehmann, S. Tracking human mobility using wifi signals. *PloS one 10*, 7 (2015), e0130824.

[133] Schlosser, F., Maier, B. F., Jack, O., Hinrichs, D., Zachariae, A., and Brockmann, D. Covid-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences 117*, 52 (2020), 32883–32890.

[134] Shannon, C. E. A mathematical theory of communication. *Bell system technical journal 27*, 3 (1948), 379–423.

[135] Shen, Y., Han, Y., Zhang, Z., Chen, M., Yu, T., Backes, M., Zhang, Y., and Stringhini, G. Finding mnemon: Reviving memories of node embeddings. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022), pp. 2643–2657.

[136] Shibamoto, M., Hayaki, S., and Ogisu, Y. Covid-19 infection spread and human mobility. *Journal of the Japanese and international economies 64* (2022), 101195.

[137] Simester, D. I., Tucker, C. E., and Yang, C. The surprising breadth of harbingers of failure. *Journal of Marketing Research 56*, 6 (2019), 1034–1049.

[138] Simini, F., Barlacchi, G., Luca, M., and Pappalardo, L. A deep gravity model for mobility flows generation. *Nature communications 12*, 1 (2021), 6576.

[139] Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. A universal model for mobility and migration patterns. *Nature 484*, 7392 (2012), 96–100.

[140] Singh, V. K., Bozkaya, B., and Pentland, A. Money walks: Implicit mobility behavior and financial well-being. *PLOS ONE 10*, 8 (2015), 1–17.

[141] Sobolevsky, S., Sitko, I., Des Combes, R. T., Hawelka, B., Arias, J. M., and Ratti, C. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In *2014 IEEE international congress on big data* (2014), pp. 136–143.

[142] Solmaz, G., and Turgut, D. A survey of human mobility models. *IEEE Access 7* (2019), 125711–125731.

[143] Son, H., Hyun, C., Phan, D., and Hwang, H. Data analytic approach for bankruptcy prediction. *Expert Systems with Applications 138* (2019), 112816.

[144] Song, C., Koren, T., Wang, P., and Barabási, A.-L. Modelling the scaling properties of human mobility. *Nature physics 6*, 10 (2010), 818–823.

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

[146] Song, X., Kanasugi, H., and Shibasaki, R. Deeptransport: prediction and simulation of human mobility and transportation mode at a citywide level. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA, 2016), pp. 2618–2624.

[147] Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R., Yuan, N. J., and Xie, X. Prediction and simulation of human mobility following natural disasters. *ACM Transactions on Intelligent Systems and Technology (TIST) 8*, 2 (2016), 1–23.

[148] Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. Quantifying international human mobility patterns using facebook network data. *PloS one 14*, 10 (2019), e0224134.

[149] Statista. Gdp of the new york metro area from 2001 to 2020. Available online at: `https://www.statista.com/statistics/183815/gdp-of-the-new-york-metro-area/`, last accessed on 2023-11-23.

[150] Stouffer, S. A. Intervening opportunities: a theory relating mobility and distance. *American sociological review 5*, 6 (1940), 845–867.

[151] Suárez-Vega, R., Gutiérrez-Acuna, J. L., and Rodríguez-Díaz, M. Locating a supermarket using a locally calibrated huff model. *International Journal of Geographical Information Science 29*, 2 (2015), 217–233.

[152] Suhara, Y., Bahrami, M., Bozkaya, B., and Pentland, A. S. Validating gravity-based market share models using large-scale transactional data. *Big Data 9*, 3 (2021), 188–202.

[153] Sun, J., Zhang, J., Li, Q., Yi, X., Liang, Y., and Zheng, Y. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering 34*, 5 (2020), 2348–2359.

[154] Tang, J., Liang, J., Yu, T., Xiong, Y., and Zeng, G. Trip destination prediction based on a deep integration network by fusing multiple features from taxi trajectories. *IET Intelligent Transport Systems 15*, 9 (2021), 1131–1141.

[155] Tang, T. T. Information asymmetry and firms' credit market access: Evidence from moody's credit rating format refinement. *Journal of financial economics 93*, 2 (2009), 325–351.

[156] Tarozzi, A., and Deaton, A. Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *The Review of Economics and Statistics 91*, 4 (2009), 773–792.

[157] Te, Y.-F. *Predicting the Financial Growth of Small and Medium-Sized Enterprises using Web Mining*. Doctoral Thesis, ETH Zurich, 2018.

[158] Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., and Colizza, V. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology 10*, 7 (2014), e1003716.

[159] Traunmueller, M. W., Johnson, N., Malik, A., and Kontokosta, C. E. Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems 72* (2018), 4–12.

[160] Truscott, J., and Ferguson, N. M. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLOS Computational Biology 8*, 10 (2012), 1–12.

[161] U.S. Small Business Administration Office of Advocacy. 2018 Small Business Profile. `https://www.sba.gov/sites/default/files/advocacy/2018-Small-Business-Profiles-US.pdf`, 2018. Accessed: 2023-01-14.

[162] U.S. Small Business Administration Office of Advocacy. Frequently Asked Questions About Small Business, 2021. `https://advocacy.sba.gov/2021/11/03/frequently-asked-questions-about-small-business-2021/`, 2021. Accessed: 2023-01-14.

[163] Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. How correlated are network centrality measures? *Connections (Toronto, Ont.) 28*, 1 (2008), 16.

[164] Van der Maaten, L., and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research 9*, 11 (2008).

[165] van Veenstra, A. F., and Kotterink, B. Data-driven policy making: The policy lab approach. In *Electronic Participation: 9th IFIP WG 8.5 International Conference, ePart 2017, St. Petersburg, Russia, September 4-7, 2017, Proceedings 9* (2017), pp. 100–111.

[166] Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., and Stanley, H. E. Lévy flight search patterns of wandering albatrosses. *Nature 381*, 6581 (1996), 413–415.

[167] Wang, S., Cao, J., Chen, H., Peng, H., and Huang, Z. Seqst-gan: Seq2seq generative adversarial nets for multi-step urban crowd flow prediction. *ACM Transactions on Spatial Algorithms and Systems (TSAS) 6*, 4 (2020), 1–24.

[168] Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. Quantifying the impact of human mobility on malaria. *Science 338*, 6104 (2012), 267–270.

[169] Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., and Buckee, C. O. Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the national academy of sciences 112*, 38 (2015), 11887–11892.

[170] Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52.

[171] Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., and Pentland, A. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. *Available at SSRN 1130251* (2008).

[172] Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., and Liu, C. Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine 56*, 3 (2018), 142–149.

[173] Xiong, C., Hu, S., Yang, M., Luo, W., and Zhang, L. Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections. *Proceedings of the National Academy of Sciences 117*, 44 (2020), 27087–27089.

[174] Xu, Y., Belyi, A., Bojic, I., and Ratti, C. Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems 72* (2018), 51–67.

[175] Yabe, T., Bueno, B. G. B., Dong, X., Pentland, A., and Moro, E. Behavioral changes during the covid-19 pandemic decreased income diversity of urban encounters. *Nature Communications 14*, 1 (2023), 2310.

[176] Yabe, T., Bueno, B. G. B., Dong, X., Pentland, A., and Moro, E. Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters. *Nature Communications 14*, 1 (2023), 2310.

[177] Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., and Ukkusuri, S. V. Understanding post-disaster population recovery patterns. *Journal of the Royal Society Interface 17*, 163 (2020), 20190532.

[178] Yabe, T., Zhang, Y., and Ukkusuri, S. V. Quantifying the economic impact of disasters on businesses using human mobility data: a bayesian causal inference approach. *EPJ Data Science 9*, 1 (2020), 36.

[179] Yang, D., Fankhauser, B., Rosso, P., and Cudre-Mauroux, P. Location prediction over sparse user mobility traces using rnns. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (2020), pp. 2184–2190.

[180] Yoon, J. S., and Kwon, Y. S. A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information. *Expert systems with Applications 37*, 5 (2010), 3624–3629.

[181] Zhang, J., Feng, B., Wu, Y., Xu, P., Ke, R., and Dong, N. The effect of human mobility and control measures on traffic safety during covid-19 pandemic. *PLoS one 16*, 3 (2021), e0243263.

[182] Zhang, Z., Chen, M., Backes, M., Shen, Y., and Zhang, Y. Inference attacks against graph neural networks. In *31st USENIX Security Symposium (USENIX Security 22)* (2022), pp. 4543–4560.

[183] Zhu, B., Qian, C., Pan, X., and Chen, H. A trajectory-based deep sequential method for customer churn prediction. In *Proceedings of the 2020 5th International Conference on Machine Learning Technologies* (2020), pp. 114–118.

[184] Zipf, G. K. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review 11*, 6 (1946), 677–686.