Models for excess demand in urban environments

by

Xin Liu

B.S., Wuhan University, China, 2014

M.S., Western University, Canada, 2016

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation proposal was presented

by

Xin Liu

It was defended on

December 21, 2022

and approved by

Dr. Konstantinos Pelechrinis, School of Computing and Information, University of Pittsburgh

Dr. Prashant Krishnamurthy, School of Computing and Information, University of Pittsburgh

Dr. Hassan Karimi, School of Computing and Information, University of Pittsburgh

Dr. Alexandros Labrinidis, School of Computing and Information, University of Pittsburgh

Dissertation Advisor: Dr. Konstantinos Pelechrinis, School of Computing and Information, University of Pittsburgh

Models for excess demand in urban environments

Xin Liu, PhD

University of Pittsburgh, 2022

In urban lives, citizens are motivated to visit business venues by personal needs and venue attractiveness. This creates the demand from citizens on urban businesses. As citizens move around the city to visit multiple business venues, they rely on the urban transportation systems. This creates the demand from citizens on transportation systems. To provide decent service, business venues and transportation systems are designed to satisfy a specific demand level per the operator's expectation. However, the actual demand can exceed the operator's expected demand level due to external factors (e.g., peak hour, weather, special venues nearby). The portion of the demand exceeding the operator's expected demand level is identified as the excess demand. Generally, existing works did not consider excess demand since such demand can easily be unobserved and ignored; this leads to biased analysis and forecasting for the actual demand.

In this thesis, firstly, we use the real-world data to uncover the existence of excess demand. Next, we estimate the excess demand for the urban business. Particularly, we propose our approach, which is based on simulations and complementarity, to estimate the excess demand for urban business entities. For each urban business venue, we estimate every source of its excess demand. For urban areas, we reveal the excess demand patterns among different periods in a day, and find that the excess demand can be explained by the venue diversity, venue density, number of venues and inter-area distance of urban areas. We fetch the embeddings of urban areas via a graph neural network and reveal the inter-area relationship in the latent space. Then, we estimate the excess demand of the urban transportation systems. Particularly, we propose our approach to estimate the excess demand in an urban bike sharing system. To predict the net total demand (which includes the observed and excess demand), we build a Skellam regression model, which shows advantages over other alternative models, both in terms of predictive performance and interpretability. Moreover, our Skellam regression model, as a generalized linear model, allows us to get a better estimation of the uncertainty of our prediction. The estimated excess demand provides insights for business owners, transportation operators and urban planners to satisfy more demand, which increases the revenue for business and creates more convenience for citizens.

Keywords: Excess demand, Urban Business, Urban Transportation.

Table of Contents

List of Tables

List of Figures

Preface

I would like to express my appreciation to people who have been playing significant roles during my PhD study at Pitt.

First and foremost, I am deeply grateful for my advisor Prof. Konstantinos Pelechrinis. He has been guiding me with insightful directions and influencing me to an active and critical thinker, whenever I face research challenges. Specially in our remote meetings during COVID, I highly appreciate his kind and flexible support on both the research discussions and my mental health. He has set an example of excellence as a researcher, mentor and instructor.

I would like to express my gratitude to my committee members, Prof. Prashant Krishnamurthy, Prof. Hassan Karimi, and Prof. Alexandros Labrinidis, for their valuable suggestions and helpful guidance through my thesis study.

Great thanks also go to my fellow colleagues at Pitt, for all the happy times in and out of the lab pre-COVID and their lots of helps to me post-COVID. I am really grateful for their accompaniment through such a long journey.

I also have special thanks to my alumnus in Western University: Guanghui Song, Sicong Liu, Longyi Chen, Lei Shu. They have been constantly encouraging me and mentally supporting me during my PhD study, just like how they helped me during my master's.

Finally, I would like to give my endless gratitude to my parents whose love and encouragement have always been my support and inspiration to get over every difficulty.

1.0 Introduction

In this chapter, we will introduce the background of the topic of this thesis proposal, namely excess demand. Then we will provide an overview of research tasks to explore the excess demand.

1.1 The Concept of Excess Demand

The vitality of urban lives are highlighted by the demand from citizens to visit business venues. The intuitive motivation to visit a venue is typically a citizen's essential need. For example, a family needs to visit a grocery store weekly, a city dweller needs to visit a museum since this is her plan of relaxation. However, such basic need can possibly create temporary visits to other related and complementary venues. In the above examples, the family may take the opportunity during their trip to the grocery store and visit a gas station nearby to add some fuel in their vehicle even if this was not the initial purpose of the trip. Furthermore, the museum visitor may decide to visit a bar nearby after even if she has never been there before, simply because it is convenient and serves her current need of getting a refreshment (shown in Fig. 1). Since such demands are relatively temporary and perhaps not *primary*, the service capacity expected by a venue operator potentially fails to satisfy part of such demands. In other words, such demands are excess over a venue operator's expected service capacity; we refer to such demand as *excess demand*. Additionally, in the example of the bar above, the excess demand from the museum visitor at the bar is observed since she places an order at the bar. However, the excess demand in the business venues may or may not be observed. For example (shown in Fig. 2), a customer sees a long queue at Bar A and then she placed an order at a nearby Bar B instead. This customer creates unobserved part of excess demand on Bar A since she planned to order at Bar A if the queue was not overwhelming; but her original plan to order in Bar A cannot be observed/recorded in Bar A's transaction system.

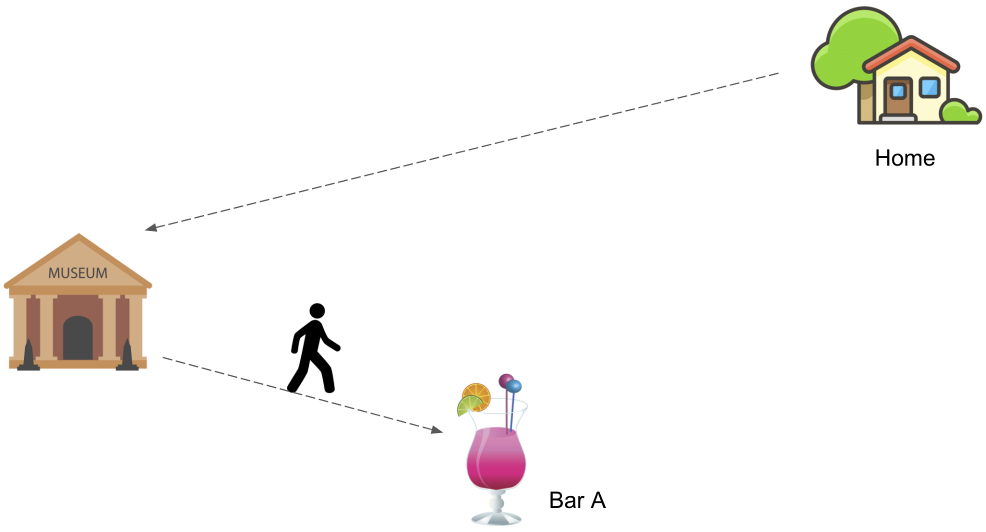


Figure 1: An example of excess demand for business venue Bar A due to complementarity with a museum.

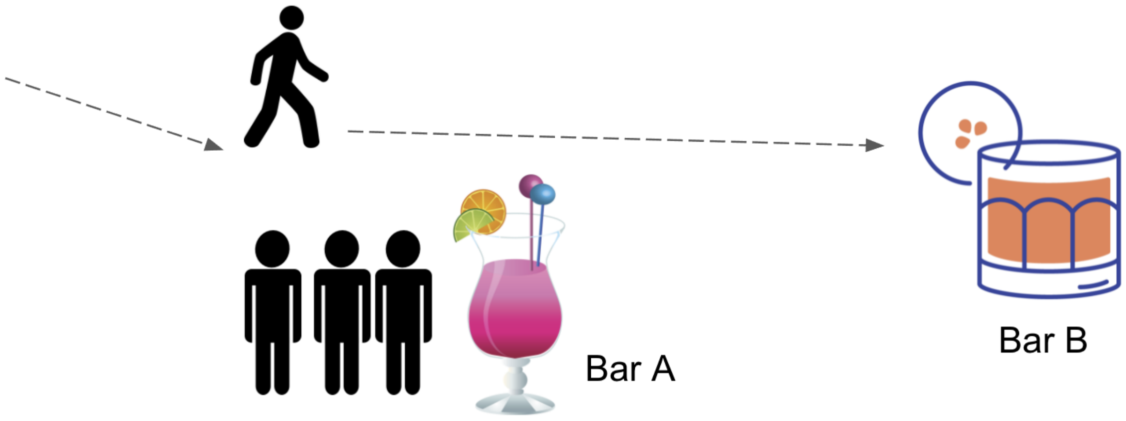


Figure 2: An example of the unobserved part of excess demand for business, where there is an unobserved part of excess demand at Bar A and observed demand at Bar B.

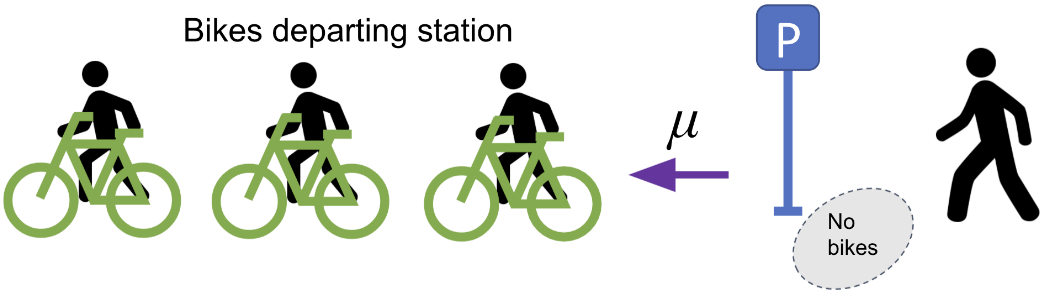


Figure 3: An example of excess demand for a bike station.

The excess demand not only exists at business venues but also in the urban transportation system, where multiple reasons can lead to the excess demand. Firstly, citizens rely on the transportation system to visit venues. The excess demand on venues is potentially propagated into the urban transportation system; this causes the demand on the transportation system to increase and probably higher than the operator's expected demand level. Moreover, the excess demand may come from relatively straightforward factors. For example, during weekday peak hours, the number of commuters can be a lot larger than the transportation system's capacity. When the weather is sunny, more commuters would choose to ride shared bikes for health and relaxation purposes, instead of other transportation approaches; this way, bike stations have excess demand and easily run out of bikes. Furthermore, when a special event happens at a venue, the transportation systems near that venue can suddenly experience excess demand, which is caused by a large number of participants and audiences coming for the event's scheduled hours. The excess demand on the transportation system is typically unobserved. For example, a customer arrives at a bike station without any available bikes (shown in Fig. 3). Then she will probably go to other stations or choose other transportation methods. Her original plan to rent a bike at this station cannot be observed or recorded in the bike rental logs; this is actually an excess demand at this bike station, which cannot be observed.

Due to the existence of excess demand, formally, we define the following types of demands, which is applied to both the urban business venues and the transportation stations:

- Capacity demand: the demand volume that a venue operator expects.
- Excess demand: the demand volume that exceeds a venue operator's expectation.
- Total demand: the sum of capacity demand and excess demand.

1.2 Importance of Excess Demand

The excess demand is easily ignored by researchers and business, transportation operators, since it may not be directly observed and quantified. Failing to consider excess demand by researchers leads to biased analysis and predictions of demands (to be elaborated in the next section). Operators can uncover benefits from researchers' results if excess demand is involved in researches, and consequently benefit the citizens/customers. That is, the operator can improve their service efficiency to satisfy the excess demand and then obtain more revenue; at the same time, since more customers' demands are satisfied, the urban life becomes more convenient for general citizens.

We elaborate the importance of excess demand for various types of operators in the following paragraphs.

For existing business owners: As mentioned in the previous section, excess demand of business venues can come from related and complementary venues. Bringing the example of a bar in the previous section, one of its complementary venue is a museum not far away. It may have other complementary venues such as restaurants and office buildings. If the bar owner can know the volume, reasons and patterns of the excess demand, she can try special strategies to increase the service efficiency. Let us assume she knows excess demand can be high during holiday peak hours. She can encourage customers to reserve online to avoid physically queuing for available seats. She can hire more staff to increase cooking and serving efficiency. She may also advertise a fast "take-out" option to customers who just need a drink (without sitting down and talking). In this way, she can serve more customers and increase the revenue. Otherwise, customers may need to queue for longer time, or go to alternative and competitive venues, such as other bars and convenient stores. It is a pity for the bar owner to lose these customers.

For new business owners: Excess demand of business venues provides insights of candidate locations to open new business venues. Let us assume that by demand analysis, we find that a specific neighborhood has high excess demand on bars. This means the current bars in this neighborhood cannot satisfy the demand from customers. Then a businessman can choose this neighborhood to open a bar.

For transportation operator: The excess demand analysis for the transportation system can help improve transportation services. For the bike sharing system, if the operator know that the excess demand of some stations is high, the operator can increase the bike fleet or relocate more bikes to those stations; the operator may also need to install more racks to handle these increased bikes. For bus or subway, if the excess demand is high in some neighborhood, the operator may consider increasing the service frequency; the operator can also consider revise the routes or the locations of stations to avoid the commuter gathering at few locations.

1.3 Existing works on Excess Demand

When analyzing the demands of business venues and transportation systems, most existing works did not consider the excess demand. They directly use historical transactions to represent all the demands, which leads to biased analysis and forecasting for the actual total demand. We categorize these literature into the aspects of urban business and the urban transportation.

1.3.1 Urban Business

As mentioned in the previous sections, excess demand of business venues (or urban areas) mainly comes from inter-venue (or inter-area) complementarity. The diversity of an urban area can potentially strengthen its complementarity with other areas and the work [78] examines such effect of urban area diversity. Particularly, using citizens' mobile signal data, this work uncovers that the number of visitors to an urban area is correlated with the venue diversity of this urban area. In contrast, competitiveness is contradictory to complementarity; so competitiveness may undermine the extent of excess demand. The work [20] aims at uncovering the effect of competitiveness. Using historical menu data and operation condition data of restaurants, it finds that existing restaurants in an urban area do not respond differentially to newly open restaurants in this urban area. Also, areas with lots of newly open restaurants (i.e. high-level competitiveness) have higher possibility for restaurants to exit. While the above two works [78, 20] elaborate the effect of complementarity/competitiveness, such effect is not connected to the concept of "excess demand" or similar concept. Furthermore, most other works on business demands only mention that complementarity/competitiveness can have non-trivial influence on demand; they did not explore and elaborate the extent of such influence. To quantify the complementarity and its influence, we may need data in finer granularity, such as the Foursquare Future Cities Challenge (FCC) dataset[1] which provides the movement records between venues. Calafiore *et al.* [11] use such data to have a detailed study of the human dynamics, where Pearson correlation coefficient is applied as a useful analytical metric.

As mentioned in previous sections, the excess demand can be unobserved. Only few works attempt to estimate such unobserved demand. This work [70] defines the unobserved demand as the number of customers visiting a store but finally did not buy any item from this store. Since the transaction history can only record the observed demand, the authors propose a method based on "multinomial logit (MNL) model" [71] to estimate the unobserved demand. Firstly, a lost-share ratio is pre-defined to express the percentage of unobserved demand in total demand. Then they use Poisson regression to model the total demand, i.e., the customer flow arriving at the store. They also use the linear regression to model the attraction level of the competitors; this is the reason of unobserved demand of this store since customers are attracted by competitive stores such that they do not buy any item in the current store. Then the Poisson regression and linear regression models are jointly trained by the data from transaction records. Finally, the whole model can output the volume of unobserved demand, i.e., the number of customers to visit a store and choose not to buy anything. The validity of this method is verified by simulation based on the idea of

[1] https://www.futurecitieschallenge.com/

Poisson distribution. Another works [47] uses a similar approach to reach a sightly different objective: to estimate the probability of a customer not buying a product from a store after visiting.

Multiple studies examine how the demand can be correlated with some metrics of venue locations, which is a signal of excess demand existence in good venue locations. In general, a retail store is expected to be more successful if it is located within a shopping center or a central business district (CBD), which provides convenient transportation access and attractiveness [45]. Given also the correlation between retail store density and street network centrality [61, 60], a central location will be preferable. Jensen [37, 38] also considers network effects in interactions between different types of venues, while Aboolian *et al.* [1] develop a spatial interaction model that seeks to simultaneously optimize location and design decisions for a set of new venues. However, the proposed model assumes a purely homogeneous customers base, that is, all customers are identical with respect to their venue preferences and expenditure decisions. Furthermore, the lack of detailed customer volume for the venues, creates issues for estimating the potential excess demand for a venue in an area. In a subsequent series of studies [2, 3] the authors assume that customers' demand for a venue v (i.e., the probability of visiting the venue) decreases with the distance from v and increases with the *attractiveness* of it. Based on this assumption, the authors provide a spatial interaction model for locating a set of new facilities that compete for market share. Their models are applied and evaluated on synthetic data, showing the efficiency of the proposed algorithmic solution to the discrete multi-venue competitive interaction optimization problem. However, it is not clear how they will perform in the real-world. Other approaches [7] identify the optimal location for a store by maximizing the number of customers expected to be covered taking mobility patterns into consideration. Bozkaya *et al.* [10] use a genetic algorithm to select a single site among several candidate locations to maximize the market share under budget constraints. The results are verified by a real-world dataset of a supermarket chain in City of Istanbul. Furthermore, Karamshuk *et al.* [42] study the predictive power of various geographic and mobility-related features on the popularity of retail stores in New York City using Foursquare data. Works [51, 62] also find that an area with multiple venue options will overall attract more people that are interested in exploring the area and hence, all the venues will potentially enjoy the benefits from the associated network effects.

Depending on the granularity and quality of the data, sometimes urban business demand analysis will be conducted in the level of urban area. Identifying urban areas and revealing their functionality will assist in analyzing excess demand in the level of urban area. Existing literature has attempted to identify the functionality of urban areas, and consequently, cluster areas based on their functionality. Topic modeling is the dominant techniques in this line of research (e.g., [22, 77]). Other studies have attempted to identify similar areas across cities mainly using the type of activities recorded in the different areas of the different cities (e.g., [46, 26]).

1.3.2 Urban transportation systems

There have been many existing works on demand analysis and forecasting for urban transportation systems. For these works on bus and subway services [39, 14, 48, 80, 72, 73], they consider that the actual total demand only includes the records in historical logs (e.g., transit smart card data, fare collection machine data). Gerte1 et al [31] point out that the recorded demand has an unobserved error from the approach of measurement; such unobserved error follows Gaussian distribution. Note that such unobserved error is by nature the irreducible error of the measurement system, which is not directly related to the excess demand; the excess demand is due to low transportation service capacity when the demand is very high.

Particularly, for bike sharing systems, there have been several studies on demand analysis and predictions, i.e., the expected number of bikes to be rented and returned at each station. Most of them only consider the observed demand, i.e., the demand reflected in the trip data logged by the system [41, 50, 75, 15, 34, 29]. However, the total demand includes also trips that were never realized due to empty docks. To reiterate, we refer to this part of the total demand as (unobserved part of) excess demand. Failing to involve the excess demand will essentially provide a model that only captures the observed demand of the system, essentially treating any period with zero observed rentals (or returns respectively) as periods of zero demand, which is not true in general.

In a slightly different, but relevant, problem formulation some studies focus on bike availability prediction, i.e, the expected number of bikes available for rental at a station [64, 28, 75, 76, 50, 30]. A variety of specifications have been used for the prediction models, including auto-regressive moving average, K Nearest Neighbors, random forest, gradient boosted tree, and neural networks. Hierarchical predictions [34, 53] have also been developed, where stations are firstly clustered into relevant groups (e.g., geographically close) and then, predictions happen at the cluster level.

Some of these studies, such as the one from Schlote *et al.* [64] point out that a popular station may run out of bike quickly if the demand is so high, while others [15] identify "over-demand" stations as those that are full or empty for more than 10 minutes. Then they propose algorithms to classify a station as an "over-demand" one. However, none of these studies attempts to estimate the volume of excess demand.

However, there are studies that attempt to estimate the volume of excess demand using a simple method based on the duration for a station being empty [56, 52, 49]. These methods assume that excess demand exists every time there are zero bikes available for rental. They further consider this excess demand to be equal to the observed demand in adjacent time periods. It should be evident that neither of these assumptions are very realistic. A station can be empty and no user is interested in renting a bike from that station, while the excess demand does not have to be equal to the observed demand in adjacent times.

The patent [5] raises a method to use the process in supply chain the estimate the unsatisfied demand; the unsatisfied demand is similar to excess demand and the rationale of [5] is related to satisfying more demands in bike sharing systems. The situation in [5] is that the supply chain is unable to satisfy the all the demand at a specific moment t_1. At a later time t_2 it replenishes extra supply as compensation. The objective is to estimate the portion of the demand which is unsatisfied at t_1. The proposed method starts by initializing a scaling parameter to scale the satisfied demand at t_1 to the total demand. Then through the supply chain process and observed demand history data from t_1 to t_2, the scaling parameter is optimized iteratively. Finally, the optimized parameter can directly be used to calculate the unsatisfied demand. In bike sharing system, the excess demand at a station is similar to the unsatisfied demand at t_1, and later, bikes rebalanced to this station is similar to extra supply replenishment at t_2. The difference is that in bike sharing system, the excess demand at t_1 may immediately disappear since the customer may choose alternative transportation methods and does not need a bike anymore.

1.4 Research tasks and contributions

This dissertation proposal has three major tasks: (1) Exploring the preliminary effect of complementarity on demand distribution; (2) Estimating excess demand for urban business; (3) Estimating excess demand for urban transportation.

Task (1) Complementarity on demand distribution

Task (1.1) Factors correlated with demands

`Contributions:` We find that complementarity can be a proxy for excess demand by examining the fast food restaurants near highway exits. The excess demand is correlated with complementarity and competitiveness among venues. We also find that the demand distribution is influenced by distances and venue ratings, which provides insights to explore demand patterns in more general urban areas.

Task (2) Excess demand for urban business

We explore the patterns and the volume of excess demand in the level of urban areas.

Task (2.1) Patterns of real-world demand

`Contributions:` We propose `hood2vec` to demonstrate the total demand pattern in latent space, which is very different from the pattern of venue categories. Similarities among urban areas can not only be quantified through Euclidean distances and correlation, but also be visualized through our implemented web APP.

Task (2.2) Excess demand quantification and patterns

`Contributions:` By choosing the complementarity as the proxy, we estimate the excess demand for the urban business. Particularly, we propose our approach, which is incorporates real-world and simulated data, to estimate the complementarity for urban business entities. For each urban business venue, we estimate every source of its complementarity. For urban areas, we reveal the complementarity patterns among different periods in a day, and find that the complementarity can be explained by the venue diversity, venue density, number of venues and inter-venue distance of urban areas. We fetch the embeddings of urban areas via a graph neural network and reveal the inter-area relationship in the latent space. Using these results, venue owners can improve their business strategy to satisfy more excess demand and increase their revenue.

Task (3) Excess demand for urban transportation

For urban transportation, we estimate and predict the excess demand in bike sharing systems.

Task (3.1) Excess demand quantification approach

`Contributions:` We design that the proxy of excess demand is a temporal segment in the bike availability data, that include changes in the availability from zero to non-zero. Assisted by this proxy, we propose our approach to estimate the excess demand based on queuing theory. We verify through simulations its ability to estimate the excess demand present in the bike-sharing system.

Task (3.2) Total demand prediction model

`Contributions:` We learn a Skellam regression model to predict the net total demand, which shows advantages over other alternative models, both in terms of predictive performance, as well as, interpretability. Using these results, the bike sharing operator can strategically rebalance bikes to satisfy more excess demand, which provides convenience to citizens and improves the city's transportation condition.

1.5 Chapters Overview

The rest of the dissertation is organized as follow: Chapter 2 explores the preliminary effect of complementarity on demand distribution. Then, in Chapter 3 for urban business, we present the patterns of real-world demands; then, we estimate the excess demand via the proxy - complementarity, and analyze its patterns. After that, in Chapter 4 for urban bike sharing system, we estimate the excess demand and build a prediction model for the net total demand. Finally, we conclude and present future directions in Chapter 6.

2.0 Complementarity on demand distribution

Generally, demand distribution patterns of a city or a state can help one understand citizens' needs in this area. Such patterns also provide insights for business owners to create more demands and attract more customers by cooperating with other related venues. In this chapter, we firstly explore the intuitive factors which affect the demand distribution. In order to focus on intuitive factors, we minimize the effect of non-intuitive factors, such as the user preferences. More specifically, we only choose one type of venues as our venues of interests: fast food restaurants near highway exits. By observing biased demand distribution of such restaurants among different highway areas, we find potential sources of excess demand using descriptive regressions and identify complementarity as a candidate proxy for excess demand. These potential sources also motivate us to further incorporate them in estimating the excess demand of general urban areas in the next chapter.

Notations used in describing our approaches and models through Chapter 2 are shown in Table 1.

Table 1: A list of notations used through Chapter 2.

Symbol	**Description**
C_F	Average single-restaurant check-in count in a large-extent cluster
N_F	Number of fast food restaurant in a cluster
N_G	Number of gas venues in a large-extent cluster
N_L	Number of lodging venues in a large-extent cluster
N_O	Number of venues in "others" category in a large-extent cluster
f	market share fairness in a cluster
d	Average pairwise distance of venues in a cluster
ρ	Coefficient of variation for venue reputation in a cluster
h	Coefficient of variation for hours of operations in a cluster

2.1 Dataset of high-way fast food restaurants

To analyze the patterns of demand distribution, we start from the intuitive factors which affects the demand distribution. This means we need to minimize the effect of non-intuitive factors, such as user preferences and product differences among venues. We focus on a specific type of venues, namely fast food restaurants, and on a particular environment, that is, highway exits. With this setting, our analysis will include venues that offer the same service, at very similar quality and price points. Particularly, we minimize the effect of these two non-intuitive factors:

- Customers' needs and preferences. Customers at these restaurants are mainly drivers on the highway. Typically, these customers share a common objective - resolving hunger. Fast food restaurants are cost-effective and fast for them to get rid of hunger. As drivers are busy heading to their final destinations via highway, they don't care very much about their preferences of the types and tastes of food.
- Restaurants' service. Fast food restaurants serve similar food; services are efficient in a similar style. Such similar services cannot easily shape a customer's preference towards a specific fast food restaurant.

For our study we focus on the state of Pennsylvania, and we use the iExit API[1] that provides information about points-of-interest at highway exists. Every point of interest is a tuple of the following form: <id, phone, latitude, longitude, address, name, category, rating, price Tier, brand name, exit ID>. We collect a total of 1,537 tuples that correspond to fast foods over the highway network in Pennsylvania over 482 exits. We also need data for customer visitation. We then query Foursquare's public venue API[2] to obtain information about the number of check-ins in each of these venues.

Locations of venues

Figure 4 depicts the locations of the fast food restaurants used in our analysis. As we can see the restaurants are clustered very closed to each other. We further annotate each

[1] https://iexit.readme.io/

[2] https://developer.foursquare.com/

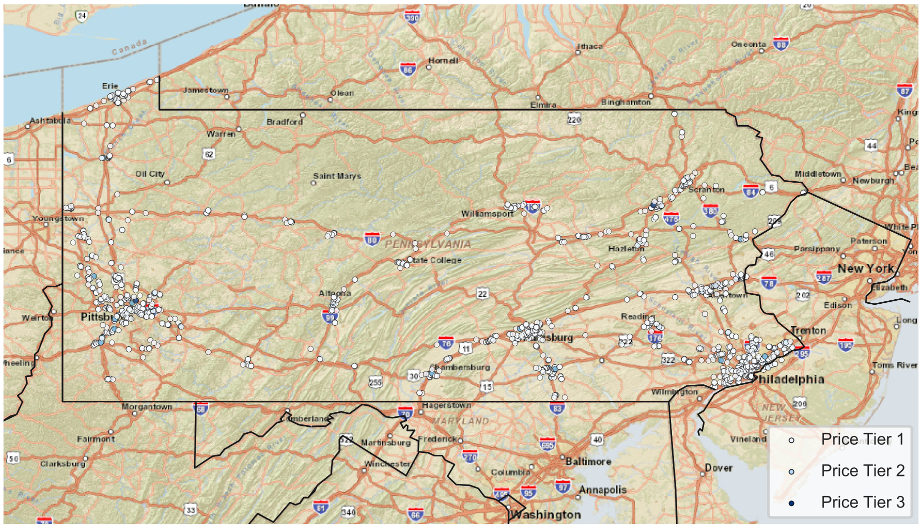


Figure 4: A map of the restaurants used in our study. The vast majority of them belong to the lowest price tier.

point with a color representing its price tier based on the iExit data (with 1 being the lowest - cheapest - and 3 being the highest). As we can see the vast majority of them (95.1% of them) belong to the lowest price tier, which means that the venues in our dataset have very similar price points.

Number of check-ins

We begin by calculating the average daily check-ins for every venue in our dataset. For this we use the number of days that each venue has been on Foursquare up to the day of data collection (i.e., 19/06/2018). By using the average daily number of check-ins we essentially alleviate problems associated with the fact that older venues might have higher number of total check-ins simply by virtue of being on the system for longer. We also want to filter out venues that did not have a check-in for an extended period of time, which can be a sign of a venue that has been closed, and hence, we remove all venues with average daily check-ins less than 0.1.

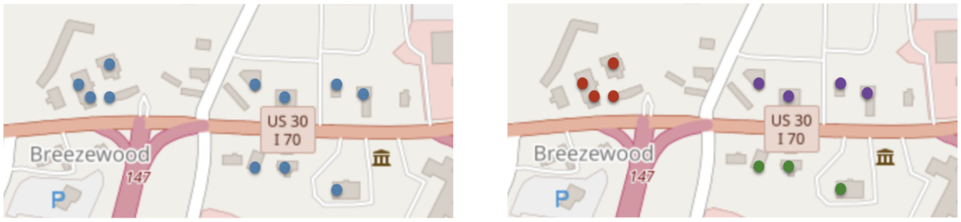


Figure 5: The different extents for our analysis. At a large extent (left) we consider a set of individual venues that are accessible from the same highway exits. After clustering the venues using HDBSCAN, at a small extent analysis we consider the set of venues within each one of these small clusters (right).

Geographic extent

The geographic extent will be initially defined through the area covered from the venues in the vicinity of each highway exit. Based on the data from the iExit API, a venue can be accessed from multiple exits[3]. Venues sharing the same group of highway exits through which they are accessible are all geographically close to each other and they can be considered as co-located at a large-extent. There is a total of 150 such clusters that will form our initial analysis unit and we will refer to them as *large* clusters. This setting is presented on the left part of Figure 5, where the blue circles correspond to a large cluster of fast food restaurants around a set of highway exits.

As we *zoom in* to smaller extents, we further divide the large clusters to smaller sub-clusters and will explore the demand patterns in these smaller extents. We identify sub-clusters based on their density using HDBSCAN with haversine distance [13]. HDBSCAN is a variation of DBSCAN that adaptively chooses the value of ϵ, that is, the maximum distance between two points to be considered in the same cluster. Therefore, the only parameter we need to specify is the minimum number of points `minPTS` that a cluster needs to include. We set `minPTS` = 2 since from our application point of view it is not meaningful to have only one venue to be identified as a cluster. HDBSCAN, similar to DBSCAN, will label points

[3] These exits typically correspond to different directions on the same highway, or exits that are located close enough to provide accessibility to the same venues.

that cannot be included to any cluster (due to distance greater than ϵ) as noise. These data points will be ignored for the subsequent analysis. We identify a total 256 sub-clusters, which we will refer to them as *small* clusters. We will refer to this case as the small extent setting (right part of Figure 5).

In order to obtain an idea of the actual length scales the different settings refer to we calculate the maximum pairwise distance between venues for all the clusters. The average of these maximum pairwise distances are: 2.44 miles and 0.51 miles for large and small extent respectively.

In the next section, we start to analyze the demand patterns in the level of large-extent clusters. This can give us a high-level overview and impression of demand patterns for our venues of interest: fast food restaurants near highway exits. The results in the overview can provide insights for the aspects we will explore in small-extent clusters.

2.2 Excess demand in large-extent clusters

To iterate, we start to analyze the demand patterns in the level of large-extent clusters, which can give us a high-level overview and impression of demand patterns. Another reason is that highway demands/checkins are relatively sparse. If we aggregate such check-in data in large-extent clusters, the results can be more statistically powerful.

The idea of aggregation directly raises our interest in the sum number of check-ins. Equivalently, we focus on average number of checkins for single restaurant in a large-extent cluster, denoted as C_F. We observe that some clusters have larger C_F while others have smaller C_F. Larger C_F indicates fast food restaurants in this cluster have extra customer visits while other clusters may not have such extra visits. Such visitation difference among clusters exists even if fast food restaurants near the highway are very similar. Therefore, we recognize such extra visits as a signal for excess demand, since this is the portion of the total demand which potentially exceeds the business owner's expectation.

We are interested in how intuitive factors can influence total demand; since excess demand is part of total demand, we can move on to inspect how these factors are potentially correlated with excess demand. We build a regression model where C_F is the dependent variable. Then we use the following factors as independent variables:

- N_F: Number of fast food restaurant in a large-extent cluster.
- N_G: Number of gas venues in a large-extent cluster.
- N_L: Number of lodging venues in a large-extent cluster.
- N_O: Number of venues in "others" category in a large-extent cluster. Venues in this category include: supermarkets, banks, medical venues, attraction and camping venues.

The result of our regression model is demonstrated in Table. 2 [4]. We analyze the effect of each independent variable as follows:

- N_F: The coefficient of N_F is positive and significant. Since more fast food restaurants show larger restaurant diversity for a cluster, this indicates that restaurant diversity potentially attracts more customers to this cluster. This reflects that venue diversity is a possible source of excess demand.
- N_G: The coefficient of N_G is negative and significant. This means more gas-related venues undermine the population of fast food restaurants. This is potentially because most gas stations have their own convenience stores. Customers going to gas stations may directly buy food and drinks in the corresponding convenience stores. In other words, gas venues and fast food restaurants are competing with each other. The excess demand of a fast food restaurant can be suppressed by its competitors.
- N_L: The coefficient of N_L is positive and significant. This indicates that customers in lodging venues find it convenient to go to restaurants nearby to have a meal. This means a restaurant can be complementary to a lodging venue. Therefore, another possible source of excess demand is complementarity.
- N_O: The coefficient of N_L is not significant. This potentially because customers among highway do not have very high demand on these venues.

In this section, we find that complementarity can be a proxy for excess demand by exploring the data in the level of large-extent clusters. We also find that the extent of

[4] There are a total of 150 large clusters. However, two of them does not have gas or lodging venues. We exclude those two clusters in this regression model.

Table 2: Our regression model results where the dependent variable is C_F.

variable	**coefficient**
intercept	27.0360***
	(1.211)
N_F	0.3122***
	(0.098)
N_G	-0.4837**
	(0.228)
N_L	0.7076***
	(0.240)
N_O	0.1685
	(0.147)
N	148
R^2	0.137

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

diversity and complementarity are positively correlated with the volume of excess demand, and the extent of competition is negatively correlated with the volume of excess demand.

2.3 Demand distribution within a cluster

In this section, we will analyze more detailed patterns of demands. Firstly, we will analyze how the demands of restaurants are distributed within the same cluster. Secondly, we not only inspect the single-restaurant demand distribution in large-extent clusters but also small-extent clusters. We will explore the effect of other factors, which are different from the previous section.

We start by introducing the dependent variable in this section.

Market Share Fairness f: In this section, we define and use Market Share Fairness as the metric to identify the patterns of the demand distribution among venues. Every set of venues A (let us assume a large-extent cluster WLOG) can be described through a vector $C_A = [c_1, c_2, \ldots, c_N]$, where c_i is the average daily check-ins in venue i of cluster A. Vector C_A should exhibit *fairness*, that is, every venue in A obtains their fair share of the market. To quantify the market share fairness f_A in the set of venues A we are going to use the coefficient of variation of C_A [67]:

$$f_A = \frac{\text{std}(C_A)}{\text{mean}(C_A)}. \tag{1}$$

When the total market within cluster A is allocated fairly across the venues in A, f_A will be 0. Hence, the smaller the value of f_A, the more fair the allocation of the market share within the cluster venues. (In the remaining part of this section, f_A will be written as f for simplicity.)

Then we introduce the independent variables as follows.

Average pairwise distance of venues d: In order to inspect how distance influences the demand distribution, we also calculate the average pairwise haversine distance of venues within the area of interest. In particular, in the case of the large extent this corresponds to the average pairwise distance of all the venues in a large cluster. Formally, with $d(i, j)$ being the distance between venues i and j, that belong to cluster A, the average pairwise distance of venues d_A, in cluster A, is given by:

$$d_A = \frac{\sum_{i,j=1,i\neq j}^{N} d(i,j)}{N_d} \tag{2}$$

where N is the number of venues in cluster A and $N_d = N(N-1)/2$ is the number of venue pairs within A. Similarly, for the case of the small extent setting the average pairwise distance is calculated in the same way, using only the venues within the corresponding small cluster. (In the remaining part of this chapter, d_A will be written as d for simplicity.)

Venue reputation: Even though one expects that fast food restaurants offer a similar quality of service/food, the *reputation* of specific brands might impact the market share they get. To get an estimate for the reputation of a venue we use the average Foursquare rating of all the brand's venues in the 10 largest cities in Pennsylvania. We then use the coefficient of variation for the reputation ρ_A of the venues within a cluster A as an independent variable in our regression. (In the remaining part of this chapter, ρ_A will be written as ρ for simplicity.)

Hours of operations: If a venue within a cluster has significantly different hours of operations (e.g., shorter hours of operations), then this will potentially affect the market share it obtains. Hence, we collected hours of operation for every venue in our dataset and calculated for every cluster the coefficient of variation (similar to Equation (1)) for the weekly hours of operations h_A for the venues in each cluster A. (In the remaining part of this chapter, h_A will be written as h for simplicity.)

Next, we present the results of our analysis in the different extents examined (i.e., the large extent and the small extent). We start by building a regression model where our dependent variable is the market share fairness f and our independent variables are the aforementioned variables in this section, which can have an impact on f. Figure 6 presents the correlations between the independent variables and the market share fairness for both extents examined, while Tables 3 present the results from our regression models for the different geographic extents.

We start by analyzing the effect of N_F, since N_F is also discussed in the previous section. From Tables 3, we can observe that, in both large and small extents, the lower N_F is, the

Table 3: Our regression model results where the dependent variable is f.

	large extent		**small extent**	
variable	d **only**	**all features**	d **only**	**all features**
intercept	0.381***	0.166***	0.340***	0.111***
	(0.025)	(0.043)	(0.017)	(0.030)
d	0.0402***	0.0170*	0.0337*	0.0092
	(0.010)	(0.010)	(0.019)	(0.017)
ρ		1.7273***		1.0881***
		(0.432)		(0.299)
h		0.3870***		0.0524
		(0.115)		(0.067)
N_F		0.0060***		0.0565***
		(0.002)		(0.010)
N	150	150	256	256
R^2	0.108	0.345	0.012	0.250

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

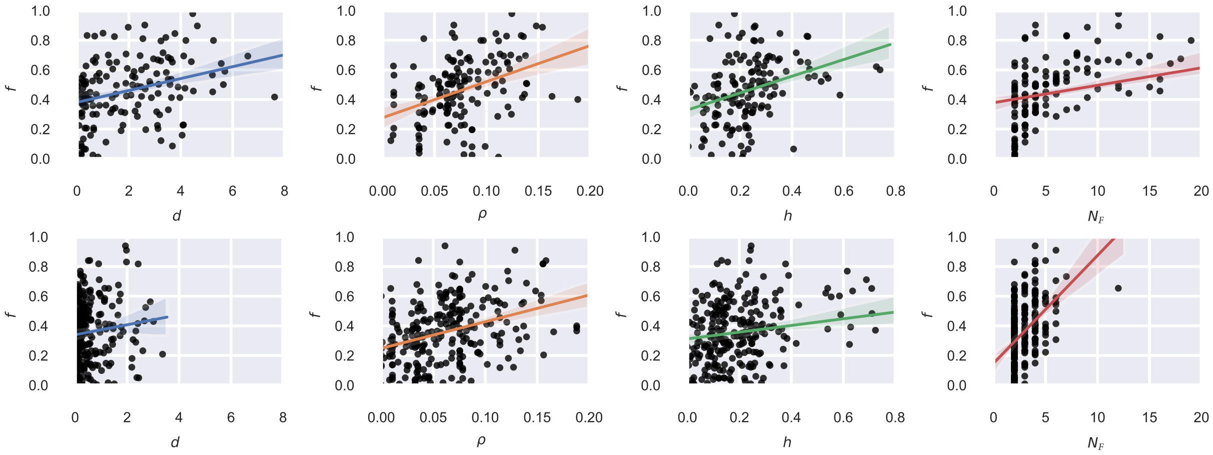


Figure 6: The relationship between pairwise venue distance d_A and market share fairness is strong in the large extent environment (top row) as compared to the small extent environment (bottom row).

more even check-ins are distributed. A possible reason is that when there are only two restaurants, they tend to be similar to each other in order not to “lose” in any aspects. This way, they become more and more similar, which causes their check-ins to become similar. In contrast, if there are many restaurants in a cluster, a restaurant may feel more confident to provide its special service in order to have more market share by being different and special, which may exactly match the preference of certain customers. Customer preferences are not evenly distributed, so such biased distribution of preferences is propagated to restaurant check-ins. In other words, restaurant check-ins are not evenly distributed since customers are able to choose restaurants based on their preferences on the special service in a specific restaurant. Here, we find that large N_F leads to the service variety of restaurants in a cluster; this is consistent with the phenomenon in the regression in Table 2 where large N_F provides restaurant diversity in a cluster.

The idea that a restaurant owner “does not want to lose” in the analysis of N_F (previous paragraph) is also applicable to d, ρ, h in the large extent. For example, we observed that the more closer restaurants are (lower d), the more even check-ins are distributed. This is because when the restaurants are very close, they tend to be similar to each other in order to not “lose” in any aspects. This way, they become more and more similar, which causes their check-ins to become similar. Their competition is so intense that potentially they may not have excess demand. This further motivates us to explore whether in a larger dataset describing a whole city, citizens only pick up one restaurant to visit from multiple restaurants that are geographically close.

Then, we will have a more detailed inspection of d, i.e., the impact from d to f in large and small extents. In the large extent, after controlling for hours of operations and venue reputation, the distance between the venues is still significantly and positively correlated with the market share fairness. However, in the small extent the relationship is less strong and not significant. Consequently, it has very limited explanatory power. In particular, while in the large extent setting the average venue pairwise distance explains about 11% of the total variance in the market share fairness, in small extent it merely explains 1% of it.

Part for this difference could be attributed to the much smaller variability of the pairwise distance in the small extent setting (as it is evident from the x-axis range in Figure 6). In particular, the variance of d_L for the large extent clusters L is $\sigma^2_{d_L} = 3$, while for the small extent clusters S is $\sigma^2_{d_S} = 0.45$. With a small variability in the regressor it is extremely difficulty to identify any meaningful relationship even if one exists. However, apart from that one of the key ideas is that the distance to the venue is an important factor in the decision making process. For the small extent setting, since all venues are extremely close to each others the market share fairness can be very sensitive to other parameters that we have assumed are similar among venues in our setting (e.g., pricing and service quality), while factors such as the venue reputation are more important for the customer’s decision (thus, if there is a larger skew in the reputation of the cluster’s venues this translates to a skew at the market share). In contrast, at the large extent, while venues are relatively close to each other as well, they are also reachable from many different highway exits. This means that specific venues might be preferable to others purely based on the *direction* of arrival in the large cluster, leading to a fair share of the market (when the pairwise distance for the venues d_A is relatively small). I.e., the relative co-location of venues attracts drivers from many different exits but then their relatively *larger* pairwise distance - as compared to that between venues

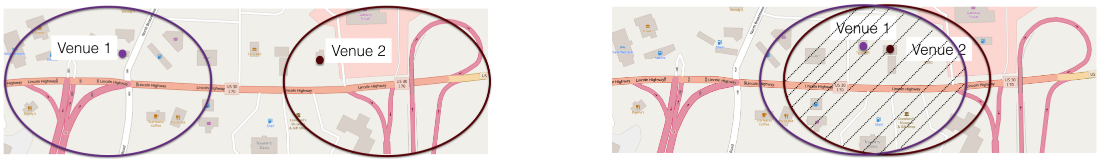


Figure 7: In a large extent cluster (left), two venues will be accessible from different exits leading to a more fair allocation of the customers, while in a small extent cluster (right) there will be significant overlap in the *service areas* of the venues and hence, the market share can be extremely sensitive to other factors (e.g., small differences in pricing, reputation etc.).

in the small clusters - can be the deciding factor for the customer's choice. We visualize this idea in Figure 7, where on the left we have two venues belonging to the same large cluster, while on the right we have two venues belonging to the same small cluster. In the former case, the venues are accessible from different exits and the circled areas include the ingress points from the highway. Customers within these areas will prefer the corresponding venue. However, in the small extent cluster, these areas have significant overlap, which means that now customers from this area might use other criteria to choose between these venues.

2.4 Summary

In this section, we find that complementarity can be a proxy for excess demand. Moreover, excess demand is correlated with diversity of venues in an urban area, the complementarity and competitiveness among venues. We also find that the demand distribution is highly influenced by distances and venue ratings. We will further explore these factors in the next section, where we will identify patterns of excess demands in more general urban areas (rather than highway areas in this section).

3.0 Excess demand of the urban business

In this chapter, we explore the extent and the pattern of excess demand in urban business entities. Firstly, by mapping urban areas into the latent spaces, we demonstrate how the excess demand can potentially influence the similarity among urban areas. Next, we propose our simulation approach to estimate the capacity demand where demands are only affected by distances and venue ratings. We consider the real-world check-in data as the total demand. By choosing the complementarity as the proxy for excess demand, we estimate the extent of complementarity (i.e., excess demand) using the difference between total demand and capacity demand. Next, we provide examples of complementarity in the venue level and area level. Finally, we analyze the statistical patterns of area-level complementarity, which provides insights for venue owners and city planners to improve the business.

3.1 Total demand from real-world data

3.1.1 Aggregate the data

In this chapter, we will use the mobility patterns of Foursquare users in the three US cities included in the Future Cities Challenge (FCC) dataset, namely, New York, Los Angeles and Chicago. The FCC dataset provides information about the mobility patterns of Foursquare users. Each data point has the following tuple format: `<start venue, end venue, trip year and month, trip period in a day, number of checkins>`. The *number of check-ins* captures the number of times that the specific movements were observed in the dataset. These check-ins occurred between April 2017 and March 2019. The dataset also provides information about the name, geographic coordinates and category for each venue.

The majority of the movements recorded in the dataset are observed only one time. In particular, 95% of the movements are observed less than 3 times.

Practically, there should be more movements than the number of movements in the dataset. In other words, the dataset only contains random samples of real-world citizen movements. If we directly use such data to estimate the demands of each venue, the results may not statistically powerful; the noise may also have high influence on the results.

We decide to aggregate the movements over a wider geographical extent. Depending on objectives and contexts of specific problems, we can aggregate the movements in one of the following level of urban area: census block group, postal area, neighborhood. Then we can transform the original data to the following format per period: <start urban area, end urban area, trip year and month, number of checkins>.

It is important to note that due to the decision to aggregate the data, we specific that our objective of this chapter is to estimate the excess demand of specific urban areas (instead of specific venues).

3.1.2 Total demand from the data

To estimate the excess demand, we need to find the best proxies to estimate the total demand and capacity demand. The excess demand of a business venue contains an "unobserved" part which cannot be recorded in transaction history. For example (shown in Fig. 2), a customer sees a long queue at Bar A and then she chooses a nearby Bar B instead. This customer creates excess demand on Bar A since she planned to order here if the queue was not overwhelming; but her original plan to order in Bar A cannot be observed and recorded in Bar A's transaction system. Here, her demand on Bar A is unobserved.

Since the total demand includes the excess demand, the proxy to represent the total demand should be able the capture the unobserved demand. We choose the number of check-ins in an urban area as the proxy of total demand for that urban area. This proxy can capture the unobserved part of excess demand, since the unobserved demand of venue A can easily become an observed demand (a check-in) at another venue nearby, i.e., venue B. Because these two geographically close venues are (very possibly) in the same urban area, the unobserved part of excess demand of venue A is successfully recorded in the same urban area as a check-in at venue B. Bringing the example of the previous paragraph, the Bar A and Bar B are in the same urban area. The total demand of this urban area should record the unobserved part of excess demand for that customer to order Bar A. Practically that demand is successfully recorded in this urban area, since her original plan to order on Bar A becomes a practical order on Bar B in the same urban area.

Discussion: Our chosen proxy for total demand in an urban area cannot 100% match the practical total demand, but given the currently available data, this proxy is the best choice. An edge case is that in the example in the previous paragraph, the customer who planned to order at Bar A may choose to go to Bar C that not in the same urban area as Bar A. This way, the unobserved demand of Bar A is not recorded as a check-in (our proxy of total demand) in the correct urban area. However, this edge case does not happen a lot, since it is much more possible for the customer to visit nearby Bar B in the same urban area rather than Bar C in another urban area. Overall, our chosen proxy is able to capture most of the unobserved demand of an urban area.

Our selected proxy of total demand in this section will be applied to the remaining sections of this chapter. We will elaborate the proxy for capacity demand in the corresponding sections of this chapter.

3.2 Demand patterns in latent space

3.2.1 Proxy for capacity demand

It is non-trivial to identify a proxy to estimate the capacity demand; the source of capacity demand should only involve essential factors which are used by business owners to conduct the demand expectation. Since our objective is to examine the demand patterns in the urban area level, one baseline method is to use the fraction of venue categories to represent an urban area. To some extent, this method can serve as a proxy to demonstrate the patterns of capacity demand of an urban area.

Venue category is an important motivation for a citizen to visit a specific location within the city. For a venue, a citizen chooses to visit it probably because the venue category meets her essential need. For an urban area, the fraction and structure of venue categories can easily reflect this area's function and signature; a citizen chooses to visit this urban area probably because the area's function meets her essential need. Capacity demand reflects a business owner's expectation on the demand volume, which involves the consideration of the venue categories. More specifically, a venue's owner knows that capacity demand is constrained by the number of people who potentially need that venue's service. For example, if area A (dominated by restaurants) and area B (dominated by luxury stores) are close to each other, the capacity demand of the area A is possible to be higher than area B. This is because everyone needs to eat and has more potential to be a customer on area A.

Therefore, though the fraction of venue categories may not be the best proxy for capacity demand, as a baseline approach, it is efficient to reveal the pattern difference from the total demand. In this section, we treat the fraction of venue categories as the proxy for capacity demand. (In the next section, we will use a more detailed approach as the proxy to reflect the influences of essential factors on capacity demand.)

3.2.2 Map demands to latent space

As mentioned above, the proxy of capacity demand in this section is the fraction of venue categories in an urban area. In other words, it is a vector where each element is the percentage of a specific category. To make the total demand and capacity demand to be comparable, we also convert the total demand (movements recorded in the dataset) of each urban area to vectors; we propose `hood2vec` as the approach for such conversion. Also, we choose the zip code level as the level of urban area in this section. We elaborate the vectorization process for both the total demand and capacity demand as follows.

To iterate, the majority of the transitions recorded in the FCC dataset are observed only one time. In particular, 95% of the transitions are observed less than 3 times. In order to avoid fitting the noise, we aggregate the transitions (movements) over a wider geographical extent. We also separate the movements according to the time period of movement occurrence according to the data - i.e., overnight (00:00 to 05:59), morning (06:00 to 09:59), midday (10:00 to 14:59), afternoon (15:00 to 18:59), night (19:00 to 23:59). Using MapQuest's

Geocoding API[1] we obtain the zip code for each venue and we aggregate the movements at the zip code level (the wider extent). More specifically, we transform the original data to the following format per period: <start zip code, end zip code, trip year and month, number of checkins>. At zip code level, only 10% of the movements have less than 2 observations. However, 20% of the zip codes contain fewer than 10 venues and hence, we filter them out from our analysis. While this might sound a large number to ignore, the checkins within these zip codes cover only 0.5% of the total checkins in the dataset.

Then, for each city $f \in \mathcal{F} = \{\text{New York, Los Angeles, Chicago}\}$ we define its directed urban flow network $\mathcal{G}_{f,p}$ per period $p \in \mathcal{P} = \{\text{overnight, morning, midday, afternoon, night}\}$ at the zip-code level as follows: $\mathcal{G}_{f,p} = (\mathcal{U}, \mathcal{E})$, where the set of nodes $\mathcal{U}$ is the set of zip code areas in city f. A directed edge $e_{ij} \in \mathcal{E}$ exists between two zip codes $u_i, u_j \in \mathcal{U}$ if there has been observed at least one movement from a venue in u_i to a venue in u_j during period p. We also annotate every edge e_{ij} with a weight $w(e_{ij})$, which captures the number of checkins of such movements observed.

We would like to note here that while we have chosen the zip codes as our unit, one can define an urban area in other levels, such as census tracts and any other levels/definitions [21].

hood2vec: Vector Representation by node2vec

In order to obtain a vector representation for the nodes of $\mathcal{G}_{f,p}$, i.e., the zip codes at f in period p, we will rely on learning a network embedding. There are several ways to learn a node embedding for a network but in this work we make use of node2vec [32]. Briefly, node2vec utilizes second order random walks to learn a vector representation for the network nodes that optimizes an urban area preserving objective function. The framework is flexible enough to accommodate various definitions of network urban areas and facilitate the projections of the network nodes in the latent space according to different *similarity* definitions. Here, we are interested in the structural equivalence of the urban areas, so we pick the parameters of node2vec accordingly ($p = 1$ and $q = 2$ [32]). We also utilize 1,000 random walks for the sampling process, while we set the dimensionality of the latent space

[1] https://developer.mapquest.com/documentation/geocoding-api/

to $d = 10$. This is consistent with the dimensionality of another vector representation to be introduced in the next heading **Vector Representation utilizing Venue Categories**. `node2vec` finally provides us with a vector $\mathbf{v}_i \in \mathbb{R}^d, \forall u_i \in \mathcal{U}$, that we can then use to identify the similarity between two urban areas. Such similarity is one important metric the reflect the demand patterns.

We refer to our proposed approach of generating vector representation by `node2vec` as `hood2vec`.

Vector Representation utilizing Venue Categories

As alluded to above, we design the proxy of capacity demand to be the fraction of venue categories to reflect the type of venues that an urban hosts. More specifically we can define a vector $\mathbf{z}_i$ for each urban area node, such that its k^{th} element $z_{ik} = \frac{n_{ik}}{N_i}$, where n_{ik} is the number of venues of type k within area i and N_i is the total number of venues within i. For defining vectors $\mathbf{z}_i$ we use the 10 top-level venue categories in Foursquare (thus, $\mathbf{z}_i \in \mathbb{R}^{10}$) : Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. Similar to `hood2vec`, we can now define the similarity between two urban areas i and j using the distance between vectors $\mathbf{z}_i$ and $\mathbf{z}_j$.

Similar to number of venues, the number of checkins in venues of different types can also be used as the vector representation of an urban area. In particular, we define a vector $\mathbf{z}_i^{\text{check}}$ for each urban area node, such that its k^{th} element $z_{ik}^{\text{check}} = \frac{n_{ik}}{C_i}$, where c_{ik} is the number of checkins of venues of type k within area i and C_i is the total number of checkins of venues within i. We follow the same 10 top-level venue categories for $\mathbf{z}_i^{\text{check}}$ (i.e., $\mathbf{z}_i^{\text{check}} \in \mathbb{R}^{10}$). Then we can also define the similarity between two areas i and j by the distance between vectors $\mathbf{z}_i^{\text{check}}$ and $\mathbf{z}_j^{\text{check}}$.

3.2.3 Urban Area Similarity

To iterate, in this section, the representation by `hood2vec` is the proxy for total demand, and the representation by venue categories is the proxy for capacity demand. Our objective is to explore the pattern difference between them, which may give us insights on the pattern of excess demand. To achieve this, we will calculate the pairwise similarities using the network embedding learnt from `hood2vec` and compare them with the corresponding pairwise similarities obtained from a simple venue-based representation of urban areas (see heading **Vector Representation utilizing Venue Categories**). Formally, the similarity of two areas i and j, with vector representations $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively, is defined as:

$$\sigma_{ij} = \texttt{dist}(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

where $\texttt{dist}(\mathbf{x}_i, \mathbf{x}_j)$ is the (Euclidean) distance between the representations of i and j.

We can now examine whether different representations for the urban areas provide different views for their similarity. In particular, if σ_{ij} and $\sigma^{'}_{ij}$ are the similarities between areas i and j using different vector representations, their Pearson correlation coefficient $\rho_{\sigma,\sigma'}$ will be high if the two representations provide similar information, and low otherwise. We can further compare in the same way the similarity of two areas for the same vector representation over different time periods.

3.2.4 Experiments and Results

In this section, we will present the results of our analysis and compare the pairwise similarities obtained from `hood2vec` and a simple venue category-based representation.

Movement and Venue Categories

We calculate the correlation between two representations, $\mathbf{v}$ and $\mathbf{z}$, (by the method in Section 3.2.3) in three cities: New York City, Los Angeles and Chicago. There is a total of 141 zip codes u_i (9870 pairs) in New York city, 111 zip codes (6105 pairs) in Los Angeles, and, 59 zip codes (1711 pairs) in Chicago. We further extend our comparisons to each time period provided in the data. The results are presented in Table 4. Note that we use the following notation for the five time periods - O: overnight; MO: morning; MI: midday; A: afternoon; N: night. As we can see all the correlations are positive, albeit, small, pointing to the two representations capturing different types of information. We also calculate the

Table 4: Correlation between movement and category representations.

Period	**O**	**MO**	**MI**	**A**	**N**
New York City	0.116^{***}	0.152^{***}	0.147^{***}	0.152^{***}	0.144^{***}
Los Angeles	0.184^{***}	0.290^{***}	0.229^{***}	0.219^{***}	0.142^{***}
Chicago	0.284^{***}	0.316^{***}	0.327^{***}	0.336^{***}	0.323^{***}

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

correlation between $\mathbf{z}$ and $\mathbf{z}^{\text{check}}$ in three cities. The correlations for these three cities are 0.839, 0.930, 0.936, respectively. I.e., the representations of venue category based on number of venues and checkins are highly correlated. This indicates low correlation of representations between `hood2vec` and checkin-based venue category.

We further inspect the relationship between the two approaches from the perspective of the top-k neighbors for each zip codes. In particular, for each zip code u_i we find the $k = 5$ closest zip codes to i based on their `hood2vec` representation ($\mathbf{v}$), $\mathcal{N}_{5,i,\texttt{hood2vec}}$. Similarly, we calculate the top-5 neighbors of zip code u_i based on their venue category representation ($\mathbf{z}$), $\mathcal{N}_{5,i,cat}$. We then calculate the Jaccard index of the two sets:

$$J(\mathcal{N}_{5,i,\texttt{hood2vec}}, \mathcal{N}_{5,i,cat}) = \frac{|\mathcal{N}_{5,i,\texttt{hood2vec}} \cap \mathcal{N}_{5,i,cat}|}{|\mathcal{N}_{5,i,\texttt{hood2vec}} \cup \mathcal{N}_{5,i,cat}|} \tag{4}$$

Table 5 presents the average Jaccard index for every city and time period. Furthermore, Figure 8 presents the Jaccard index as a function of the number of neighbors k considered for every city, averaged over different time periods and zip codes. As one might have expected from the earlier results presented, in general, under different k, there are few shared neighbors when using the two different representations for the zip codes. This strengthens our hypothesis that these two types of representations capture different information for the areas.

Moreover, Figures 9, 10, 11 illustrate the Jaccard index for every zip code per city, averaged over the different time periods. As we can see most of the zip codes in all cities have a fairy low Jaccard index. New York City's zip codes exhibit overall lower Jaccard

Table 5: Jaccard index ($k = 5$) for the three cities for the different time periods averaged over the corresponding zip codes.

Period	**O**	**MO**	**MI**	**A**	**N**
New York City	0.036	0.065	0.062	0.029	0.016
Los Angeles	0.098	0.254	0.150	0.104	0.015
Chicago	0.139	0.136	0.170	0.164	0.129

index compared to Chicago and Los Angeles (in accordance to the results in Table 4, 5). Zip codes with high Jaccard index are essentially urban areas for which the two different representations examined identify a high overlap on areas similar to them. This happens to a larger extend in Los Angeles and Chicago compared to New York City. This can potentially be due to (a) the compact nature of NYC that allows people to explore several different areas and hence, geographically remote zip codes are close in the `hood2vec` latent space, and/or, (b) the different geographic distribution of venues in the three different cities. More specifically, the compact nature may cause venues in New York city more evenly distributed, since they are easily accessible by dwellers. In contrast, scattered nature of Los Angeles may lead to biased venue distribution due to various accessibility of different regions; this could be the reason for slightly high Jaccard indices in some areas. Chicago has fewer zip code areas such that an area can has higher probability of sharing the same closest area(s) in two representations; this can cause slightly high Jaccard indices in some areas. Nevertheless, regardless of the reasons for the differences across the cities examined, in all cases the Jaccard index does not go beyond 0.4. Simply put, there is no zip-code in these three cities, for which the overlap between the top-5 neighbors identified by `hood2vec` and a simple venue-based vector representation is more than 40%, supporting our hypothesis that these two different approaches capture different information with respect to the similarity of the areas.

`hood2vec` representation across time

We further explore how the representation obtained for a zip code through `hood2vec`

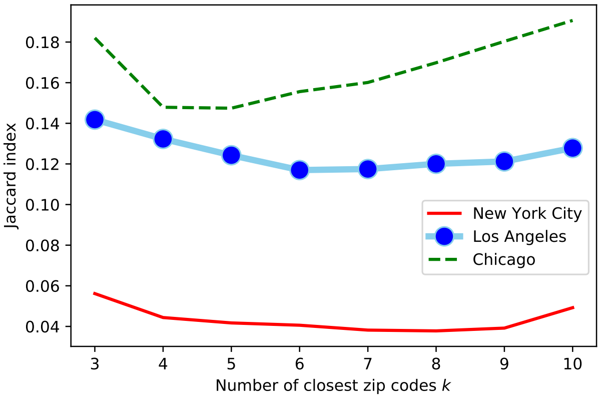


Figure 8: Average Jaccard index as function of number of closest neighbors k.

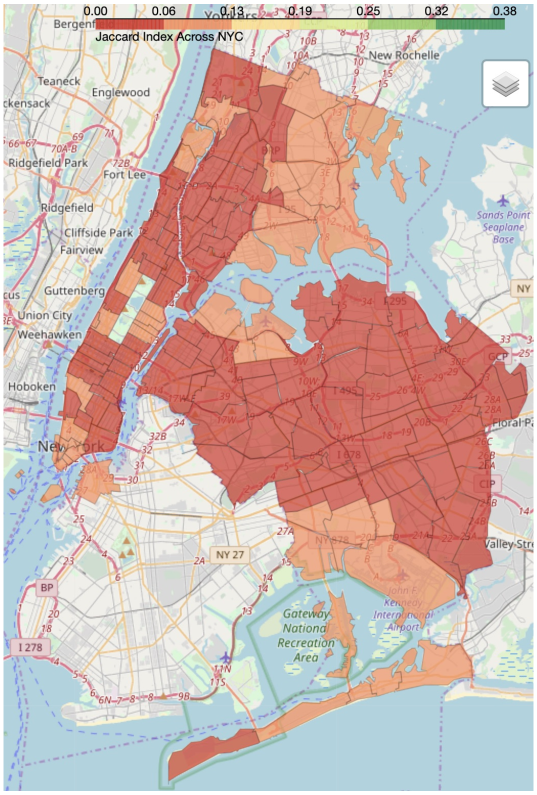


Figure 9: Average Jaccard index over the different periods in New York City.

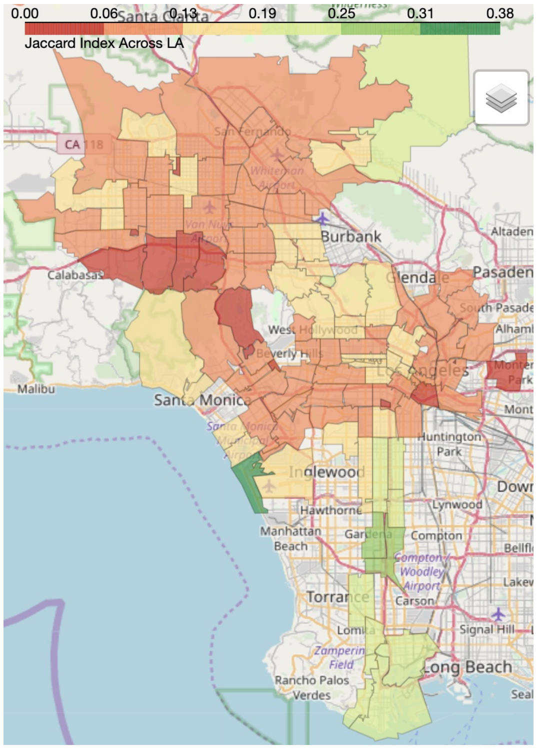


Figure 10: Average Jaccard index over the different periods in Los Angeles.

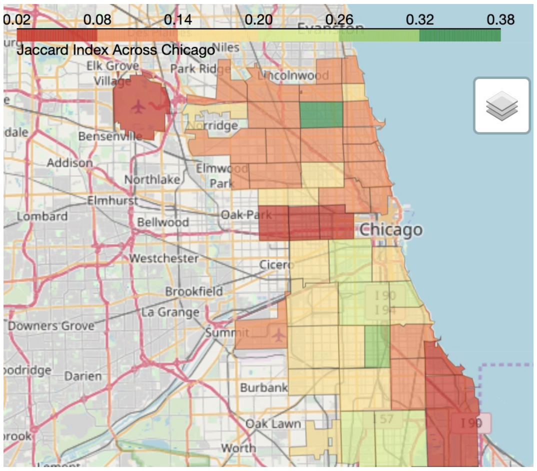


Figure 11: Average Jaccard index over the different periods in Chicago.

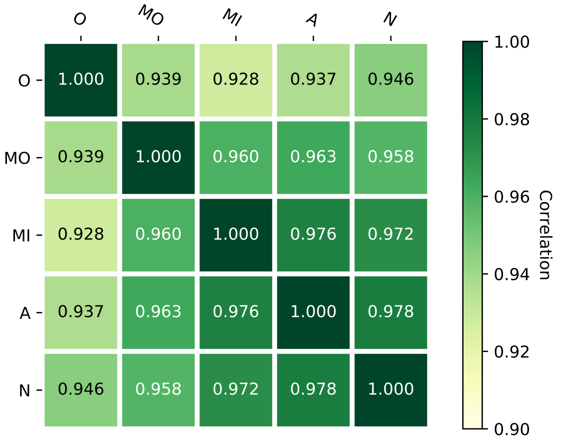


Figure 12: Correlation among representations of different periods in New York City.

changes over time (i.e., over the different time-periods in the dataset). Let us assume the two periods p_1 and p_2, and the corresponding `hood2vec` representation vectors $\mathbf{v}_{p_1}$ and $\mathbf{v}_{p_2}$ respectively. Then following similar steps as the ones described in Section 3.2.3, we can obtain the pairwise correlation of the between periods p_1 and p_2 for the same city. The correlations of each city are shown in Fig. 12, 13, 14.

One can observe that for these three cities, the correlations between any pair of periods are very high, all over 0.9. This means that, the patterns of movements are similar regardless of the time of a day (based on the `hood2vec` representation). Since New York City and Chicago are more geographically compact, it is easier for dwellers to move within the city for any purpose at any time. This could be the reason that the overall movement patterns within a day are similar. Los Angeles is geographically scattered, which limits the convenience of movements; dwellers tend to move within nearby areas at any time of the day. This may cause similar movement patterns of all day. Readers can use our implemented web APP explore the different urban area representations at: `http://www.pitt.edu/~xil178/hood2vec.html`; a screenshot for this APP is shown in Fig. 15.

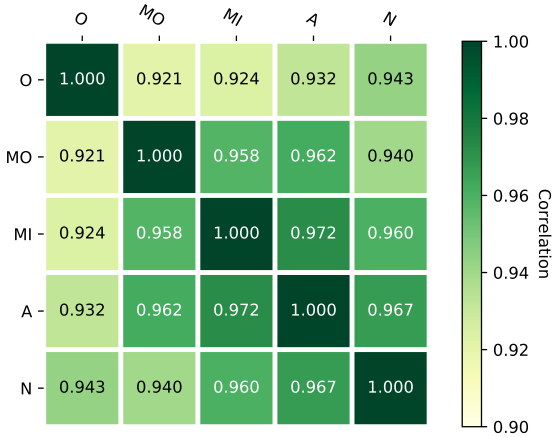


Figure 13: Correlation among representations of different periods in Los Angeles.

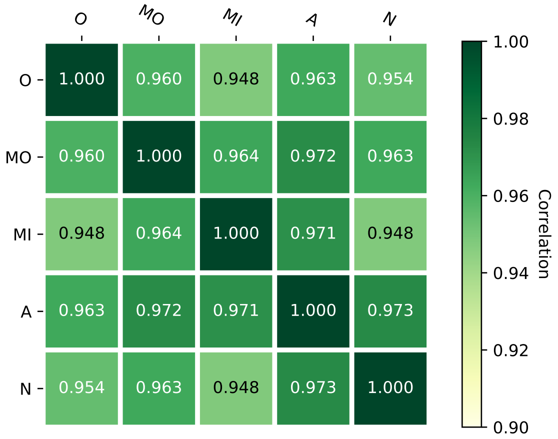
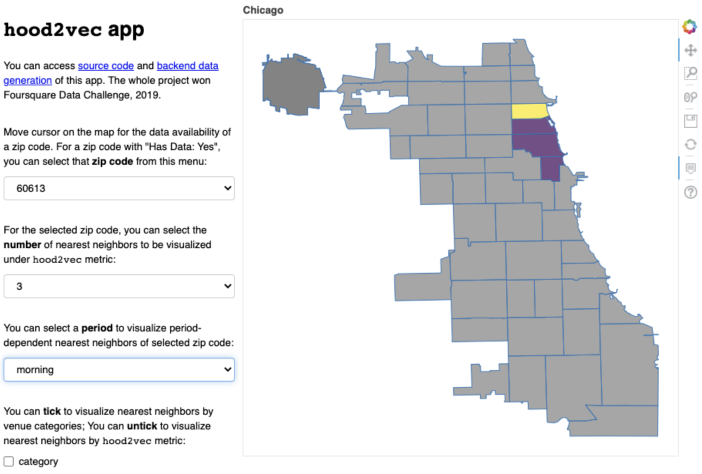


Figure 14: Correlation among representations of different periods in Chicago.



hood2vec app

You can access source code and backend data generation of this app. The whole project won Foursquare Data Challenge, 2019.

Move cursor on the map for the data availability of a zip code. For a zip code with "Has Data: Yes", you can select that **zip code** from this menu:

60613

For the selected zip code, you can select the **number** of nearest neighbors to be visualized under `hood2vec` metric:

3

You can select a **period** to visualize period-dependent nearest neighbors of selected zip code:

morning

You can **tick** to visualize nearest neighbors by venue categories; You can **untick** to visualize nearest neighbors by `hood2vec` metric:

☐ category

Figure 15: Screenshot of our implemented web APP exploring the different urban area representations.

3.2.5 Summary

In this section, we compare the similarity among urban areas in the latent space. Similarity is generated from two perspectives: venue category and citizen movement. As a baseline and efficient method, venue types reflects capacity demand patterns, since capacity demand is constrained by the number of people who potentially need that venue's service. Citizen movement reflects the total demand patterns.

Our contributions of this section are as follows. Firstly, in order to represent the total demand of an urban area based on actual citizen movement, we propose `hood2vec` to generate an embedding for each urban area. Secondly, we uncover the difference between capacity demand and total demand patterns reflected by venue category and hood2vec embeddings, respectively. Since venue category embedding has only limited information (i.e., fraction of venue categories), such pattern difference can be caused by a citizen's intuition to visit a venue nearby with high venue rating. Such difference can also correspond to excess demand; in other words, a citizen visit a venue because of inter-venue or inter-area complementarity. This motivates us to incorporate distances and venue ratings into the proxy for capacity demand in the next section. Based on the improved proxy for capacity demand, we expect to estimate the excess demand more reasonably and accurately. Lastly, we create an APP to customize the visualization of the aforementioned embedding difference in this section.

3.3 Estimate demands by simulation

In this section, based on the findings of the previous section, we will choose complementarity as the proxy for excess demand. We will also propose our simulation approach as the proxy for capacity demand, where the essential factors to drive the simulation are distances and venue ratings. The simulation results will be used to estimate the volume of excess demand for urban entities for sections afterwards. Some important notations used in describing our approaches and models from Section 3.3, 3.4, 3.5 are shown in Table 6.

3.3.1 Demand volume of urban entity

Based on the previous section, the essential factors to motivate citizens' visits are distance and venue rating. The inter-venue or inter-area complementarity is the factor to create citizens' demands/visits beyond venue owners' expectations. Therefore, we choose the complementarity as the proxy to estimate the excess demand for urban entities.

As introduced in Section 1, complementarity means the complementary relationship between two urban entities. As complementarity is the proxy for excess demand, the excess demand will also represent a relationship between two urban entities. More formally, we would like to emphasize that for the rest of this chapter (Section 3.3, 3.4, 3.5), the demand is defined as a *directed* manner from an urban entity to another urban entity; such manner is applied to capacity demand, excess demand and total demand. For example, we can say that the complementarity (excess demand) from venue v_i to v_j is 4.

Someone may argue that the calculated demand in a directed manner may not be straightforward for venue owners and city planners to understand. In fact, the directed manner is in the finest granularity. Once we obtain the demands in the directed manner, we can sum them up as the demand for a specific urban entity. We can also use the directed manner to

Table 6: A list of important notations used from Section 3.3, 3.4, 3.5.

Symbol	**Description**
v_i	a venue i
$\mathbb{B}$	a set containing business venues
$\mathbb{NB}$	a set containing non-business venues
$N_{NB \to B}$	number of movements from non-business to business venues
$d_c + \Delta r$	a ring area where d_c, Δr are the radius, width
$c^q_{i,j}$	complementarity from venue i to j
$a_{i,j}$	actual number of movements from venue i to j
$s^q_{i,j}$	number of simulated movements from venue i to j in the q-th simulation
$c^q_{I,J}$	complementarity from area I to J
$a_{I,J}$	actual number of movements from area I to J
$s^q_{I,J}$	number of simulated movements from area I to J in the q-th simulation
ψ	venue diversity from one area to another area
ϕ	number of venues from one area to another area
ρ	venue densify from one area to another area
d_{cen}	the distance between two areas' centroids
$CSR_{i,j}$	"complementarity strength ratio" from area I to J
$\mathbb{E}_s$	a set containing edges with strong complementarity. $\mathbb{E}^O_s$, $\mathbb{E}^{MO}_s$, $\mathbb{E}^{MI}_s$, $\mathbb{E}^A_s$, $\mathbb{E}^N_s$ are the sets for overnight, morning, midday, afternoon, night periods.

explain the source of demands of an urban entity. For simplicity, in the rest of this chapter (Section 3.3, 3.4, 3.5), if we mention the capacity (or total, or excess) demand of a specific venue v_j, it means the sum of capacity (or total, or excess) demand from other venues to v_j.

To iterate, in order to calculate the excess demand, we need to estimate the capacity demand and the total demand. As the proxy for capacity demand, we will use simulation to generate citizen movement records, which have the same format as the real-world dataset for total demand. The format is in a directed manner as already mentioned. Therefore, we elaborate our method of using movement data to calculate the demand volume from an urban entity to another. Such method is applied to both capacity demand and total demand.

3.3.2 Rationale of simulation

Proxies for capacity demand and excess demand

The objective of the simulation is to serve as the proxy to estimate the *capacity* demand for business. Thus, we only use essential factors to drive the simulation. As aforementioned, these essential factors are distances and venue ratings. More specifically, the capacity demand going from any other venue to venue v_1 is based on two essential factors: venues surrounding v_1 and v_1's quality. The effects of these two factors are:

- If there are lots of venues surrounding v_1, it is highly possible that existing residents and visitors in those venues will soon visit v_1 simply because v_1 is very close to them.
- If v_1 has high reputation, v_1 is very attractive to citizens, leading to lots of incoming visits.

In the real-world data, it is very hard to identify what part of the actual visits to v_1 belongs to capacity demand, which is only caused by the two essential factors (distances and venue ratings). Thus, in order to estimate capacity demand, we use our proposed simulation approach as the proxy. In the simulation, citizens are only driven by distances and venue ratings to move around the city.

Note that our defined capacity demand of venue v_j does not necessarily mean the actual physical venue capacity of v_j. For example, a businessman would like to open a store at a specific plaza; per his expectation, the capacity demand of the store in his mind is 30. But in the plaza, there is only one available venue v_j with actual capacity 40 for the him to rent. He has to rent v_j even if v_j's capacity exceeds the capacity demand in his mind. Simply put, as previously mentioned, our defined capacity demand of venue v_j is the expected demand of the venue owner on his venue v_j.

Complementarity from venue to venue is the proxy we use to estimate the excess demand. In other words, complementarity is the most possible and dominant reason for the total demand to exceed the venue owner's expected capacity demand[2]. For example, a citizen is currently in a museum because she is interested in museums. After this museum visit, she would like a short break and then visits a bar nearby for a break. In this situation, the bar has excess demand from this citizen because of the museum and the bar are complementary, not only due to the previously mentioned essential factors: distances and venue ratings. On the contrary, the objective of our simulation is to obtain the capacity demand based on essential factors. Thus, we do not simulate the movements caused by complementarity.

Complementarity based on time period

The demands vary among different periods within a day, causing the complementarity to be period-specific. Recall that we have five periods within a day. Thus, we will calculate the period-specific complementarity, in order to help venue owners deal with demands in different time periods. To achieve this, we split the real-world movements by the time period they occurred. Then in each period, we conduct the simulations and calculate the complementarity for that specific period. The diagram illustrating this idea is shown in Fig. 16.

Preprocess non-business venues

Since our research objective is to analyze the demand of business venues and help business owners, we rely on the simulation of inter-venue interactions to quantify the capacity demand of business. Before simulation begins, we need to pre-process non-business venues, which are transport venues and venues without ratings in the Foursquare venue dataset. Transport

[2] Note that part of the demand due to complementarity may have been considered by the business owner in the capacity demand, but under the currently available data, complementarity is the best proxy we have found to estimate the excess demand.

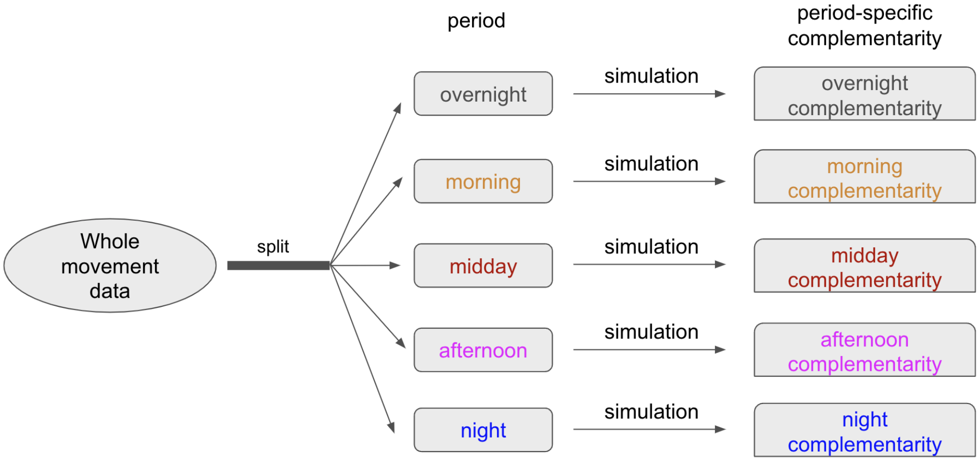


Figure 16: Diagram for period-specific simulations and complementarity.

venues serve as platforms to help citizens to reach business venues. Examples of transport venues are: bus stops, subway stations. Moreover, FourSquare has predefined certain venues not to have the attribute "rating" in the venue dataset. Examples of such venues are: Urgent Care Center, Catholic Church. Such venues are often considered non-profit. Thus, for transport venues and venues without ratings, we assign them to the "non-business venues" set (denoted as $\mathbb{NB}$) in our simulation. Otherwise, venues belong to "business venues" set (denoted as $\mathbb{B}$). For non-business venues, we assign that they all have exactly the same rating, which is the average of the ratings of business venues.

Recall that our research objective is to analyze the demand of business venues. Thus, we should control the business related statistics between the simulated and the actual movements. Only in this way, the comparison between simulated and the actual movements are meaningful in order the quantify the complementarity and help business venues. More specifically, for the simulation and the actual dataset, we split all the movements into 4 subsets based on how they are related to business. Particularly, for the actual data, we have these following 4 subsets of movements (illustrated in Fig. 17):

- From Business venues to Business venues. We have a total of $N_{B\to B}$ movements in this subset.

- From Non-Business venues to Business venues. We have a total of $N_{NB \to B}$ movements in this subset.
- From Business venues to Non-Business venues. We have a total of $N_{B \to NB}$ movements in this subset.
- From Non-Business venues to Non-Business venues. We have a total of $N_{NB \to NB}$ movements in this subset.

Next, we plan to let our simulations to match the number of movements in these 4 subsets. That is, we will do 4 sets of simulations separately. These 4 sets are:

- From Business venues to Business venues. We will simulate a total of $N_{B \to B}$ movements in this subset.
- From Non-Business venues to Business venues. We will simulate a total of $N_{NB \to B}$ movements in this subset.
- From Business venues to Non-Business venues. We will simulate a total of $N_{B \to NB}$ movements in this subset.
- From Non-Business venues to Non-Business venues. We will simulate a total of $N_{NB \to NB}$ movements in this subset.

3.3.3 Simulation Steps

We use an agent-based simulation approach where agents move inside a city to simulation citizen movements. A naive idea is to let multiple agents move simultaneously, which, however, makes the simulation become complicated and computational expensive. Instead, our approach is to simulate one agent's movement first and then switch to another. In this way, since we only collect the number of movements as the capacity demand, simulating one agent's movement after another is equivalent to movements of multiple agents simultaneously. Additionally, as aforementioned, we will have 4 sets of simulations. We will firstly propose our approach of simulating the set "from Business venues to Business venues". Then we will explain how to revise this approach to conduct the simulations for 3 other sets.

Simulating "from Business venues to Business venues"

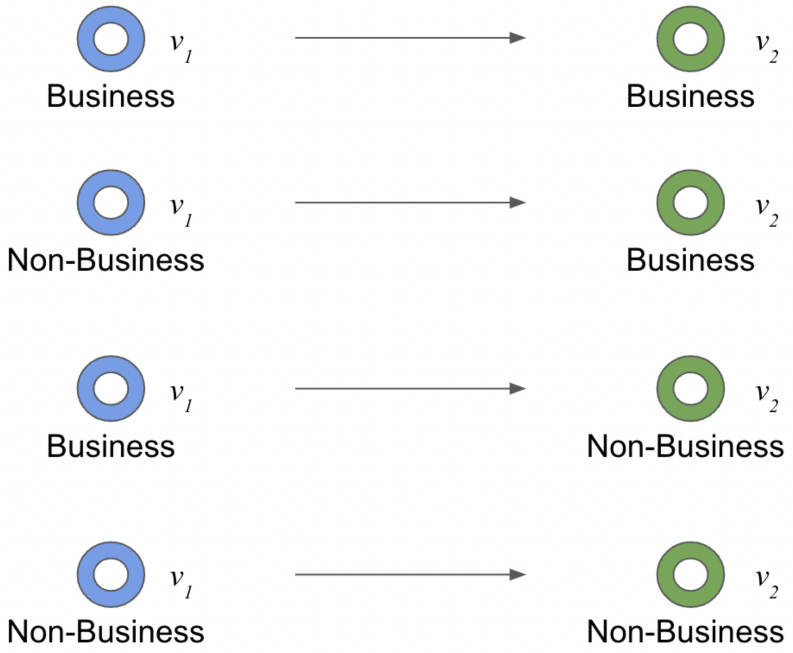


Figure 17: Illustration of business (or non-business) related movements.

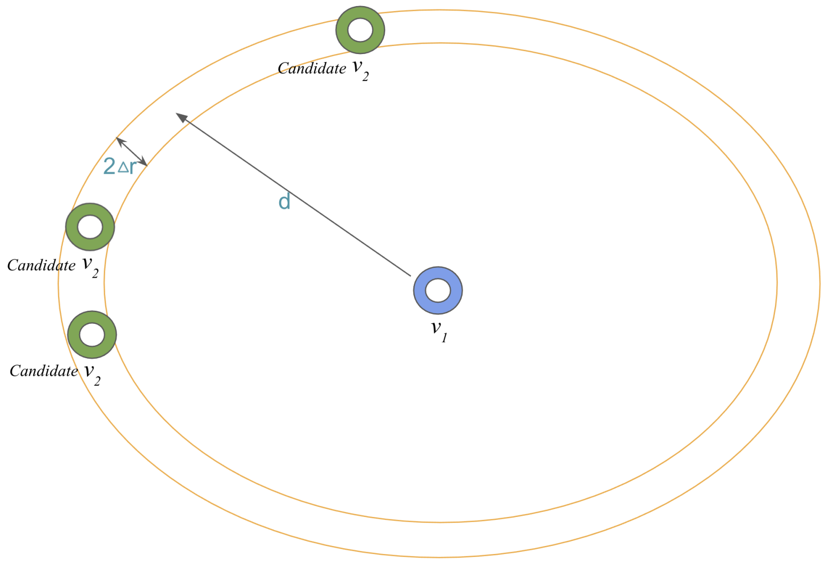


Figure 18: Illustration of choosing the next venue in a movement.

Here, we give our approach of simulating "from Business venues to Business venues", where we have a total of $N_{B \to B}$ movements. To simulate one movement, an agent (citizen) starts from a specific venue v_1, which is randomly selected from the business venue set $\mathbb{B}$. Then, two essential factors will influence an agent's choice of the destination venue: distance from the start venue v_1 and venue rating. Typically, consideration of the distance is prior to the rating since citizens usually put movement convenience and time saving in priority. Therefore, to identify the destination venue of the agent, in the first stage, we select a set of candidate venues V_c based on distances and, in the second stage, select a destination venue from V_c based on the venue rating. These two stages are illustrated in Fig. 18 and elaborated as follows.

More specifically, in the first stage, our idea is to select a set of candidate venues V_c which have similar distances to the start venue v_1. In order to achieve this, we need to firstly assume that all candidate venues have exact the same distance d_c to v_1, and then we relax this assumptions to be less strict to accommodate "similar" distances. To decide d_c, instead of randomly assigning a value, we use the distance distribution of movement distance in the real-world data as a reference. We find such a distribution has a form of $p(d) = 1/d^{\alpha}$ where d is the distance and α is a parameter with $\alpha > 1$. This expresses the tendency for citizens to visit venues within short distances, which is consistent with the essential factor of our simulation. Thus, from the distribution $p(d)$, we sample a distance d, which becomes our decided d_c. Then, to relax the strict d_c to "similar" distances, we use v_1 as the centroid to draw a circle line with radius $d_c + \Delta r$ and another circle line with radius $d_c - \Delta r$. This ring area $d_c \pm \Delta r$ contains venues with similar distances to v_1. For the simulation of each specific city, we select a suitable value of Δr to let the ring area $d_c \pm \Delta r$ have multiple venues but not too many. The appropriateness of our selected Δr will be indirectly justified by the k-s test; we will elaborate this later.

In the second stage, in the ring area $d_c \pm \Delta r$ between two circle lines, only the business venues belong to V_c. This is because in this example, we already assume that the destination should be a business venue. Within V_c, we select a venue based on the probability distribution proportional to the venue ratings within this ring area. This selected venue becomes the destination v_2 of the currently simulated movement for the current agent.

In fact, if we follow the exact steps in the first and the second stage, the computational complexity of the simulation is very high. Thus, we implement a sub-optimal approach of simulation based on the concept "geohash", which is elaborated in Appendix B. The appropriateness of this approach will be justified later by Section 3.3.4.

Then, since there are a total of $N_{B \to B}$ movements, we iterate the same process of simulating a single movement to simulate all $N_{B \to B}$ movements.

Simulating "from a type of venue to another"

Recall that we have 4 sets of simulations:

- From Business venues to Business venues
- From Non-Business venues to Business venues
- From Business venues to Non-Business venues
- From Non-Business venues to Non-Business venues

We already provide the approach of "from Business venues to Business venues". We only need to revise some details of this approach to simulate "from non-Business venues to Business venues". Here are notable revisions:

- There should be a total of $N_{NB \to B}$ movements.
- The start venue v_1 should be randomly selected from the non-business venue set $\mathbb{NB}$.
- During the second stage, only business venues (from set $\mathbb{B}$) can belong to the candidate set V_c, since the destination in this set of simulation should be a business venue.

Following the same rationale, the simulation approach can be revised to simulate "From Business venues to Non-Business venues" and "From Non-Business venues to Non-Business venues". These revisions and differences are summarized in Table 7.

We have explained the approaches to conduct 4 sets of simulations. In fact, our objective is to analysis demands going to business. This way, we only need to analyze 2 sets where the destinations are business venues, i.e., $\mathbb{B} \to \mathbb{B}$ and $\mathbb{NB} \to \mathbb{B}$. In the rest of this chapter, we only analyze and generate demand results from these 2 sets.

Table 7: Differences of 4 simulation sets.

simulation set	**total movements**	**start venue**	**end venue candidate set V_c**
$\mathbb{B} \to \mathbb{B}$	$N_{B\to B}$	business	only contains business
$\mathbb{NB} \to \mathbb{B}$	$N_{NB\to B}$	non-business	only contains business
$\mathbb{B} \to \mathbb{NB}$	$N_{B\to NB}$	business	only contains non-business
$\mathbb{NB} \to \mathbb{NB}$	$N_{NB\to NB}$	non-business	only contains non-business

3.3.4 Simulation process analysis

As aforementioned, in our simulation, citizens follow these two intuitions based on essential factors:

1. Citizens tend to visit venues within short distances of their current location.
2. Citizens tend to visit venues with high rating.

Here, we will conduct statistical analysis of our simulation to show that these two intuitions are achieved.

Simulated distances

Recall that in order to simulate citizens' tendency to visit nearby locations, our strategy is to let the distances of simulated movements follow the distribution of the distances of actual movements, since these actual distances also express citizens' tendency to visit nearby locations. Here, as we have completed the simulations, we need to verify that the simulated movements has similar distance distribution to the actual movements. The idea of such verification is to compare the simulated distances with actual distances via two-sample Kolmogorov–Smirnov (K-S) test.

Firstly, the input to the two-sample K-S test has two arrays: an array of simulated distances and an array of actual distances. In each array, the number of elements (distances) is $(N_{B\to B} + N_{NB\to B})$. This is because we use the movements whose destinations are business venues to calculate and analyze demands. Note that in NYC, Chicago, LA, this number

$(N_{B\to B} + N_{NB\to B})$ is 2271978, 2287247, 2606944, respectively; this means the input arrays are very long.

However, such long input arrays are not suitable to be the direct inputs to the K-S test. The reason is that K-S test is very strict and sensitive to the difference of two very long input arrays [54]. In other words, even if these two arrays practically follow the same distribution, the statistical result from the K-S test is very likely to express that they do not follow the same distributions. To resolve this, we involve extra steps to assist the K-S test. The idea of the steps is to randomly sample a short array from the long array of simulated distances and another short array from the long array of actual distances; then these two short arrays are input to the K-S test. This way, the K-S test becomes more reasonably strict with the short input arrays, which generates more meaningful p-values as the results of the tests. However, sampling a short array only once is not sufficient to represent the whole long array. Thus, we iterate this process of sampling short arrays and conduct K-S tests for multiple (15,000) times. Then we take the average of the p-values from these k-s tests to represent the statistical characteristics of the whole long array. Additionally, since there is no empirical length of the short arrays, we apply this process under different lengths of short arrays to provide more information. We illustrate the results of such process in Fig. 19, 20, 21 for NYC, Chicago, LA; the x-axis is the length of short arrays, and the y-axis is the average p-value by multiple experiments under the same short length. As observed in Fig. 19, 20, 21, for any short length presented in the x-axis, the p-values are larger than 0.01. This means the K-S test cannot reject the hypothesis that the simulated distances follow the same distribution as the actual distances.

Venue ratings

In the actual data, due to the effect of complementarity, citizens do not necessarily go to venues with high ratings. But in our simulations, citizens tend to visit venues with high ratings. Thus, we expect the average ratings of the actual end (destination) venues and simulated end venues are different, which we would like to verify. The idea of such verification is to compare the ratings of simulated end venue and actual end venues via two-sample t-test.

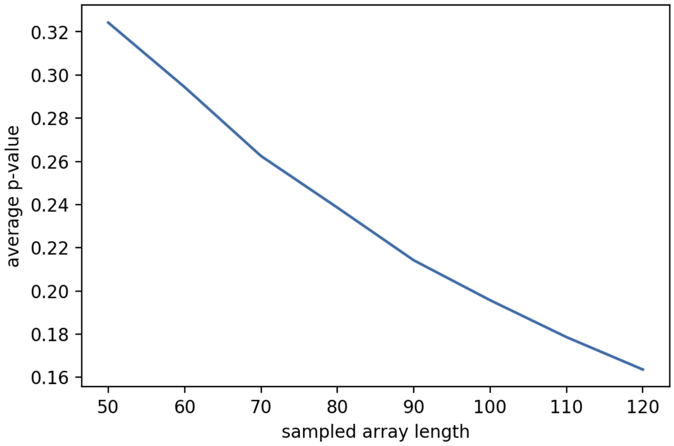


Figure 19: K-S test p-values for midday period in New York City.

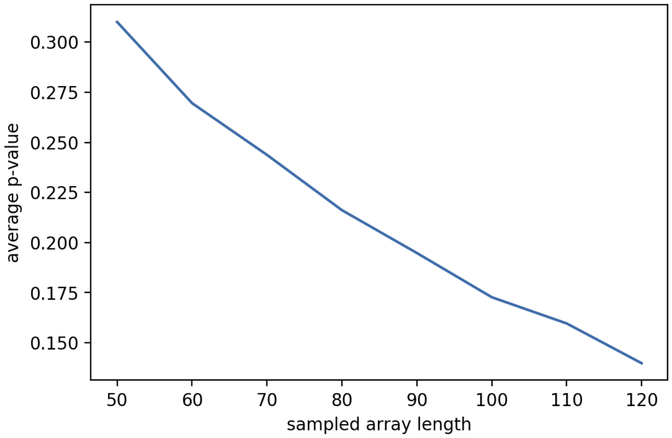


Figure 20: K-S test p-values for midday period in Chicago.

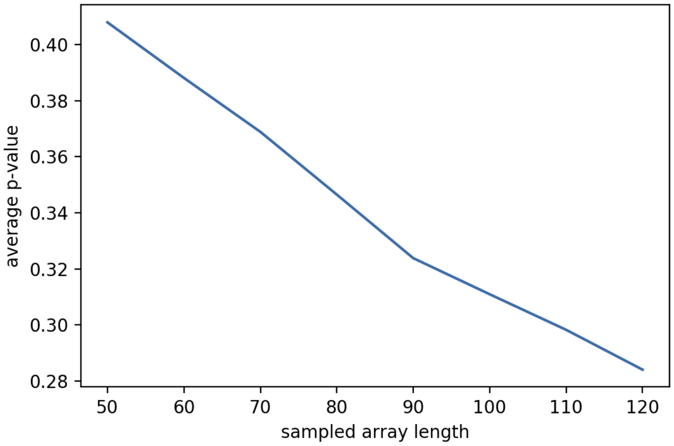


Figure 21: K-S test p-values for midday period in Los Angeles.

Particularly, the input to the t-test includes two arrays: an array of ratings of actual end venues and an array of ratings of simulated end venues. In each array, the number of elements (venues) is $(N_{B \rightarrow B} + N_{NB \rightarrow B})$, which is the same as the aforementioned K-S test. The results of the t-test are shown in Table 8. As observed, the p-values for all cities are 0.000, which verifies that the average ratings are sigfinicantly different between the actual end venues and the simulated end venues.

Note that in the current section (Section 3.3.4), for simplicity, we only show the analysis on the simulated movements in the midday period. In fact, for simulations of other periods, our results of the K-S test for distances and t-test for the ratings have the same pattern as the

Table 8: p-values w.r.t. end venue ratings.

City	**number of end venues**	**p-value**
New York City	2271978	0.000
Chicago	2287247	0.000
Los Angeles	2606944	0.000

midday period presented in the current section, which further justifies the appropriateness of our simulation approach.

3.4 Complementarity estimation in venue level and area level

3.4.1 Complementarity estimation and interpretation

Excess demand for urban entities is the focus on this chapter. In order to obtain it, we will calculate its proxy, i.e, complementarity. To iterate, the high-level idea of calculating complementarity is:

$$\text{simulated demand} + \text{complementarity} = \text{actual demand} \tag{5}$$

As follows, we generalize the meaning of annotations used in Section 3.2.2 to formulate the complementarity. For each city $f \in \mathcal{F} = \{\text{New York, Los Angeles, Chicago}\}$ we define its directed network (graph) $\mathcal{G}_{f,p}$ per period $p \in \mathcal{P} = \{\text{overnight, morning, midday, afternoon, night}\}$ as such: $\mathcal{G}_{f,p} = (\mathcal{U}, \mathcal{E})$, where the set of nodes $\mathcal{U}$ is the set of urban entities in city f. A directed edge $e_{ij} \in \mathcal{E}$ exists from entity $u_i \in \mathcal{U}$ to entity $u_j \in \mathcal{U}$. Here, the edge e_{ij} represents the complementarity from entity u_i to entity u_j. We also annotate every edge e_{ij} with a weight c_{ij}, which captures the value/extent of complementarity. Here are notable details of complementarity:

- The level of urban entities: Depending on the objective of a study, the complementarity can be analyzed in different levels. In the venue level, u_i is a specific venue i and u_j is another specific venue j. In the area level, u_i is a specific area I and u_j is another specific zip code area J.
- An edge's weight $c_{i,j}$. This weight value is the extent of complementarity. Its range is $(-\infty, \infty)$. We will elaborate the interpretation of $c_{i,j}$ later, especially when $c_{i,j} < 0$.

Note that in our simulation, we simulate the venue-to-venue movements one by one. Thus, the simulation is always conducted in the venue level. Then, the calculation of complementarity can be in venue level or area level.

Venue-level complementarity calculation

For each edge, per the actual dataset, we have the actual number of movements $a_{i,j}$, which denotes the number of movements coming from venue i to j. Our objective is to estimate the value of complementarity $c_{i,j}$, which cannot be directly obtained. Thus, we use multiple simulations in order to estimate it more precisely. Particularly, we do the simulation $Q = 20$ times; so for each edge, we have an array of number of simulated movements $\{s^1_{i,j}, s^2_{i,j}, ..., s^Q_{i,j}\}$. For each $q = 1, 2, ..., Q$, we have

$$c^q_{i,j} = a_{i,j} - s^q_{i,j} \tag{6}$$

where $c^q_{i,j}$ is the complementarity calculated using the q-th simulation. This way, we have an array of complementarity values by all simulations

$$\{c^1_{i,j}, c^2_{i,j}, ..., c^Q_{i,j}\}, \tag{7}$$

which can be considered as Q samples of complementarity. Finally, our estimated complementarity for the venue level is the average of these Q samples, which is

$$c_{i,j} \approx \hat{c}_{i,j} = \frac{1}{Q}\sum_{q=1}^{Q} c^q_{i,j}. \tag{8}$$

Area-level complementarity calculation

We will use the aforementioned a total of Q simulations to calculate complementarity in the area level. There are different extends of urban areas, such as a postal-code area, a neighborhood, a census block. In this paper, we will study the extent of postal area, but the same approach of estimating complementarity can be applied to other extends of urban areas.

To estimate the area-level complementarity, the idea is to firstly map a venue to its corresponding area code (postal code, in our situation). Then we sum up the statistics from the venue level to the area level. We illustrate this process in Fig. 22. We assume that venue i_1 and i_2 belong to area I; venue j belongs to area J. The number of actual movements from i_1 to j is $a_{i_1,j}$, and this number from i_2 to j is $a_{i_2,j}$. Since venue i_1 and i_2 belong to the same area I, we can sum up the actual movements from area I to J by: $a_{I,J} = a_{i_1,j} + a_{i_2,j}$.

Recall that we have done simulations Q times. For the q-th simulation, we use the same method to calculate the number of simulated movements from I to J. As illustrated in in Fig. 23, in the q-th simulation, the number of simulated movements from i_1 to j is $s^q_{i_1,j}$, and this number from i_2 to j is $s^q_{i_2,j}$. Since venue i_1 and i_2 belong to the same area I, we can sum up the simulated movements from area I to J by: $s^q_{I,J} = s^q_{i_1,j} + s^q_{i_2,j}$.

So, for each edge, we have an array of number of simulated movements $\{s^1_{I,J}, s^2_{I,J}, ..., s^Q_{I,J}\}$. For each $q = 1, 2, ..., Q$, we have

$$c^q_{I,J} = a_{I,J} - s^q_{I,J} \tag{9}$$

This way, we have an array of complementarity values by all simulations

$$\{c^1_{I,J}, c^2_{I,J}, ..., c^Q_{I,J}\}, \tag{10}$$

which can be considered as Q samples of complementarity. Finally, our estimated complementarity for the area level is the average of these Q samples, which is

$$c_{I,J} \approx \hat{c}_{I,J} = \frac{1}{Q} \sum_{q=1}^{Q} c^q_{I,J}. \tag{11}$$

Interpretation

There are multiple perspectives to interpret the complementarity to business owners or city planners. For convenience of explanation, we generalize the symbol of complementarity as c_{u_i,u_j}, in order to interpret venue-level and area-level complementarity together as follows:

- **Qualitative perspective:** Complementarity evaluates the complementary relationship from urban entity u_i to u_j.
- **Quantitative perspective:** Based on Eq. (6), complementarity c_{u_i,u_j} quantifies: from urban entity u_i to u_j, how many actual movements there are beyond expectation. In other words, it is the residual between the number of actual and expected movements.
  - If $c_{u_i,u_j} > 0$, then from urban entity u_i to u_j, there are c_{u_i,u_j} complementarity movements.
  - If $c_{u_i,u_j} < 0$, it is weird to say there are c_{u_i,u_j} complementarity movements. Instead, we would say that from urban entity u_i to u_j, the extent of repellence is $|c_{u_i,u_j}|$.

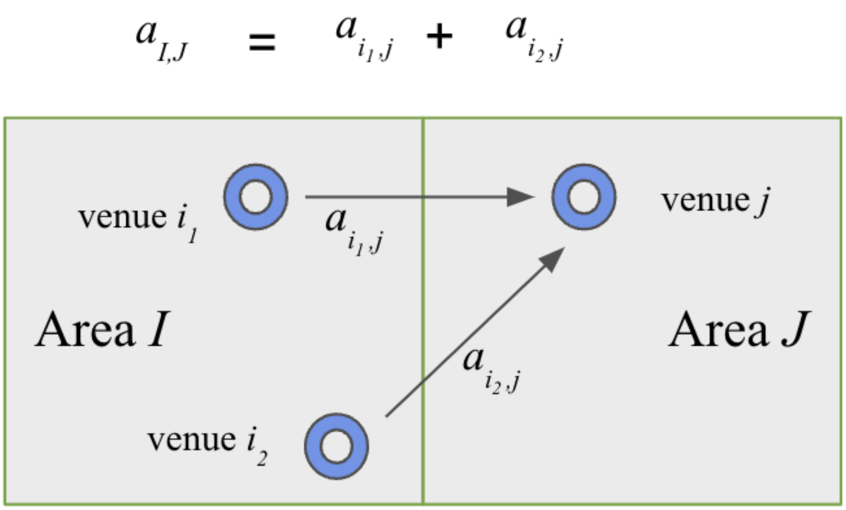


Figure 22: Illustration of area level movements.

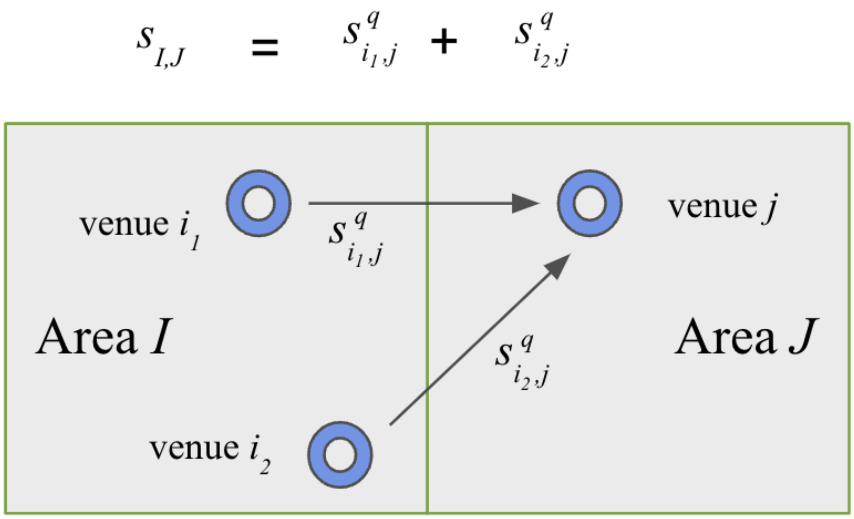


Figure 23: Illustration of area level movements by the q-th simulation.

- **Probabilistic perspective:** From Central Limit Theorem, we can assume that c_{u_i,u_j} follows normal distribution. Then, from the complementarity samples Eq. (7), we can calculate the probability of the complementarity being larger than 0, denoted as $P\{c_{u_i,u_j} > 0\}$.
  - If $P\{c_{u_i,u_j} > 0\} = 1$, this means we are confident to say that the complementarity is much larger than 0. This also means the traffic from entity u_i to u_j is higher than expected. Then we can provide the calculated c_{u_i,u_j} from Eq. (8) (this is certainly a positive value) to business owners. Then they can respond accordingly to improve the business.
  - If $0 < P\{c_{u_i,u_j} > 0\} < 1$, this means we are not 100% confident that there is complementary effect from entity u_i to u_j. Then the business owner can choose not to respond to the potential complementarity from u_i to u_j.
  - If $P\{c_{u_i,u_j} > 0\} = 0$, this means we are confident that c_{u_i,u_j} is smaller than 0. In this way, there is no complementary effect from entity u_i to u_j. Moreover, the absolute value $|c_{u_i,u_j}|$ means the extent of repellence from entity u_i to u_j.
- **Confidence Interval:** A 95% confidence interval can be inferred from Eq. (7), which describes the uncertainty surrounding the estimated complementarity value $\hat{c}_{u_i,u_j}$.

3.4.2 Venue-level complementarity examples

Following the complementarity calculation and interpretation methods, we provide examples of complementarity occurring during midday period for venues near Wrigley Filed, Chicago. For the purpose of convenience, we abbreviate the venue "Slugger's World Class Sports Bar and Grill" as "Slugger", "Merkle's Bar & Grill" as "Merkle". Their locations are marked in Fig. 24. The complementarity values existing between them are:

⋆ from Wrigley Field to Slugger: 50.00

⋆ from Wrigley Field to Merkle: 35.50

⋆ from Slugger to Merkle: -3.15

We provide interpretation in detail for some of these complementarity values. For example, for "from Wrigley Field to Slugger: 50.00":

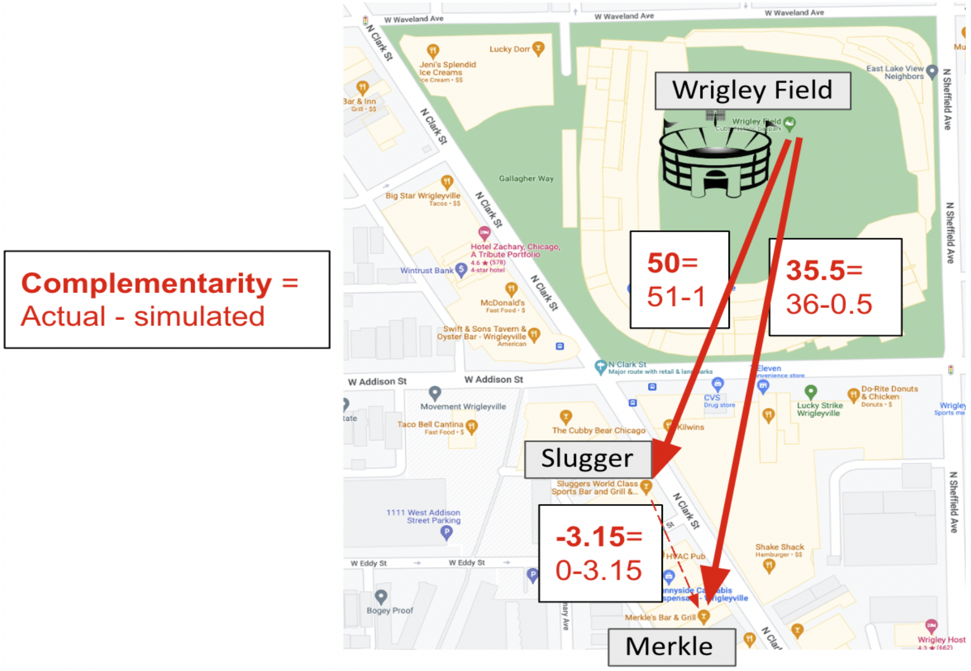


Figure 24: Examples of venue-level complementarity in Chicago during midday period.

- ⋆ **Quantitative perspective:** There are 50.00 complementary movements from Wrigley Field to Slugger, i.e., 50 movements beyond expectation.
- ⋆ **Probabilistic perspective:** From Central limit theorem, we can obtain that $P\{c_{i,j} > 0\} = 1$. This means we are confident to state that the complementarity is larger than 0. Once venue j's owner knows $c_{i,j}$, she may choose to keep the current business strategy and improve the serving efficiency to satisfy these complementary movements from venue i.
- ⋆ **Confidence Interval:** We can obtain that the 95% confidence interval is [49.57, 50.43]. Per the our estimated complementarity value 50.00, confidence interval can be written as [50.00±0.43]. The extent of uncertainty is a small number 0.43. When the venue owner knows this, she can choose to focus on the estimated value 50.00 and ignore this small uncertainty.

Then, we also provide interpretation in detail for a negative complementarity example "from Merkel to Slugger: -3.15".

- ⋆ **Quantitative perspective:** There is repellence from Merkel to Slugger; the extent of repellence is 3.15.
- ⋆ **Probabilistic perspective:** From Central limit theorem, we can obtain that $P\{c_{i,j} > 0\} = 0$. This means we are confident to say that the complementarity is smaller than 0. This also means the traffic from venue i to j is lower than expected. Once venue j's owner knows this, she may strategically choose to attract or give up movements from venue i.
- ⋆ **Confidence Interval:** the 95% confidence interval is [-3.86, -2.44]. Per the our estimated complementarity value -3.15, confidence interval can be written as [-3.15±0.71]. The extent of uncertainty is 0.71, which is not very small compared to the estimated complementarity -3.15. This means, the venue owner may need to pay attention to the complementarity interval [-3.86, -2.44] rather than a single value -3.15.

It is worth mentioning our venue-level complementarity is in a very fine granularity. This provides every source of excess demand for a specific venue and then the venue owner can take advantage of. In the above examples, let's focus on Merkle as the end venue. One source of complementarity is Wrigley Field, which provides 35.50. Another source of complementarity is Slugger, which provides -3.15. Once Merkle's owner knows Wrigley Field provides high complementarity, she may choose to change Merkle's environment to be more friendly to sports fans to improve Merkle's venue reputation. She may also choose to advertise Merkle near Wrigley Field since audience there are probably very interested in relaxing in a bar after the game. On the contrary, as Slugger provides negative complementarity to Merkle, Merkle's owner may choose to ignore this source. Simply put, in order to improve the business, the venue owner can choose and prioritize the measures to be taken based on each source of complementarity.

Discussion: Negative complementarity

Per the aforementioned negative complementarity example "from Merkel to Slugger: -3.15", our interpretation is there is repellence from Merkel to Slugger; the extent of repellence is 3.15. One may ask: does this also mean the extent of competitiveness is 3.15 from Merkel to Slugger? Our answer is: it depends on the context.

As complementarity defines, a negative complementarity value means the traffic from the start entity to the end entity is lower than expected. There are two possible relationships between the two entities to cause this: (i) they are competing. (ii) they may not be competing since their business focuses are so different.

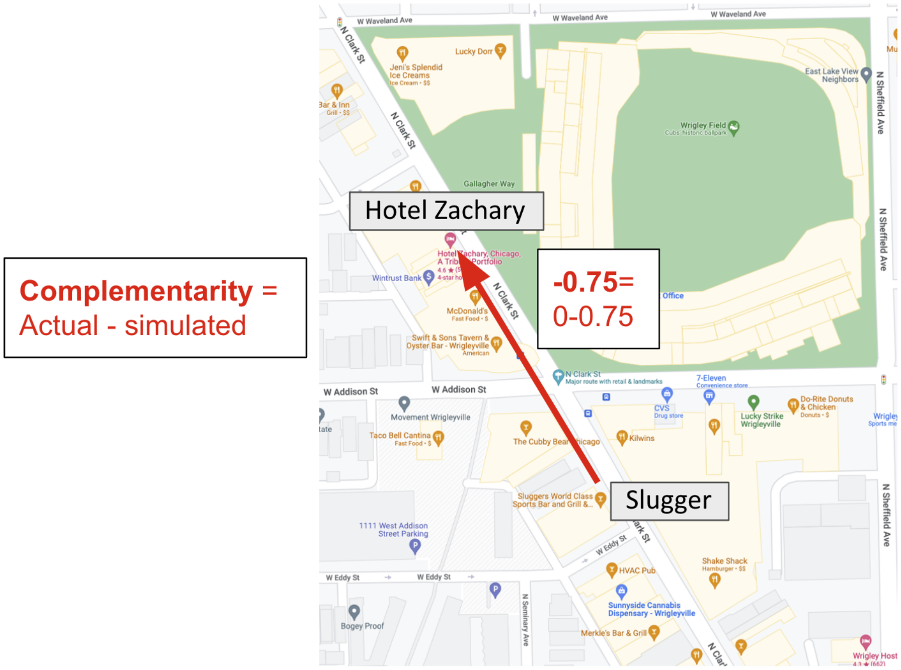
Depending on the context of the example "from Merkel to Slugger: -3.15", it belongs to the competing relationship. There are multiple reasons in this context to justify their competing relationship. Firstly, both of the two bars have much complementarity from the same start venue, Wrigley Field. This indicates they compete for the same traffic. Secondly, these two bars are very similar: they both belong "Bars" category in Foursquare dataset; they are geographically close, i.e., they are at almost the same location. Thus, it is possible for these two bars to be competitors since they are too similar. Thirdly, the complementarity value from Merkel to Slugger is negative, which means they are not complementary. This further justifies they are competitors.

Following the process of identifying Merkel and Slugger as competitors, we find that we need multiple factors to identify two urban entities as competitors. Factors include but are not limited to:

- Each of the two entities are complementary to the same third entity.
- There is repellence between the two entities.
- The two entities belong to the same category.
- The two entities are geographically close.
- The total number of nearby entities may have impacts on them.

Thus, it is non-trivial to define the competitiveness metric, which is out of the scope of this paper.

Per our calculated complementarity, we have another example "from Slugger to Hotel Zachary: -0.75", illustrated in Fig. 25. This edge has negative complementarity value but Slugger and Hotel Zachary may not be competing. The key reason is that they do not belong to the same category: Slugger is a bar while Hotel Zachary is a hotel. Slugger's



Complementarity = Actual - simulated

-0.75= 0-0.75

Figure 25: Examples of venue-level “no competition” in Chicago during midday period.

business focus is to offer alcohol drinks and nightlife. Hotel Zachary’s business focus is to provide a place to sleep, especially for travellers. They also don’t have explicit reasons to be complementary, so even if they are geographically close, citizens are not motivated to move from Slugger to Hotel Zachary. That is why the traffic in this edge is lower than expected, which causes a negative complementarity value.

3.4.3 Area-level complementarity examples

For the area-level complementarity, we provide the examples among these zip codes in Chicago: `60611` (downtown), `60613` (has Wrigley field), `60618` (does not have very special venues).

The complementarity values among these zip codes are as follows.

- from zip code `60611` to `60613`: 1124.5
- from zip code `60611` to `60618`: -690.8

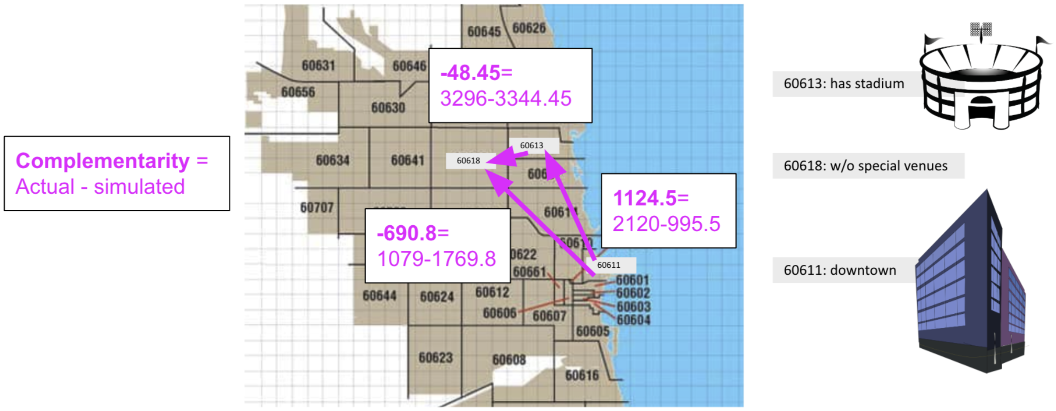


Figure 26: Examples of area-level (zip code) complementarity in Chicago during midday period.

- from zip code 60613 to 60618: -48.45

We provide interpretation in detail for some of these complementarity values. For “from zip code 60611 to 60613: 1124.5”:

- **Quantitative perspective:** There are 1124.5 complementary movements from zip code 60611 to 60613, which also means this edge has 1124.5 movements beyond expectation.
- **Probabilistic perspective:** From Central limit theorem, we can obtain that $P\{c_{i,j} > 0\} = 1$. This means we are confident to say that the complementarity is larger than 0. The city planners can take measures on zip code 60613 to satisfy these excess visits from 60611.
- **Confidence Interval:** The 95% confidence interval is [1109.8, 1139.2]. Per the our estimated complementarity value 1124.5, confidence interval can be written as [1124.5±14.7]. The extent of uncertainty is 14.7, which is a small number compared to the estimated complementarity 1124.5. Thus, the city planner can choose to focus on the estimated complementarity 1124.5 and ignore the uncertainty.

We also interpret this negative example “from zip code 60611 to 60618: -690.8” as follows:

- **Quantitative perspective:** There is repellence from zip code 60611 to 60618; the extent of repellence is 690.8. This is probably due to no attractions in 60618.
- **Probabilistic perspective:** From Central limit theorem, we can obtain that $P\{c_{i,j} > 0\} = 0$. This means we are confident to say that the complementarity is smaller than 0. This also means the traffic from zip code 60611 to 60618 is lower than expected. City planners may choose not to take any measures. This is because if zip code 60618 is a residential area, then it makes sense for 60618 not to attract traffic from downtown area 60611. It is unnecessary to improve this complementarity.
- **Confidence Interval:** the 95% confidence interval is [-704.71, -676.89]. Per the our estimated complementarity value -690.8, confidence interval can be written as [-690.8±13.91]. That is, the extent of uncertainty is 13.91, which is a small number compared to the estimated complementarity -690.8. Thus, the city planner can choose to focus on the estimated complementarity -690.8 and ignore the uncertainty.

The complementarity in area level also provides the every source of complementarity for a specific end area. In order to improve the business for a specific area, the city planner can choose and prioritize the measures to be taken based on each source of complementarity.

3.5 Complementarity Patterns in area level

We have calculated the complementarity in the area level, which is basically the summation aggregation of venue-level complementarity. Thus, the area-level complementarity, as an aggregation, is relatively statistically powerful and may reveal multiple statistical patterns. In this section, we will explore the factors causing such statistical patterns of complementarity and how different these patterns are among periods.

3.5.1 Compare complementarity values among periods

Since we calculate the period-specific complementarity values, we are interested how different they are among periods. Our method is that for a specific city we calculate the coefficient between the complementarity of two periods.

The results coefficients are shown in Fig. 27. For example, the value is 1.282 located at the first row and the second column of the results for Chicago. It is calculated by the following steps:

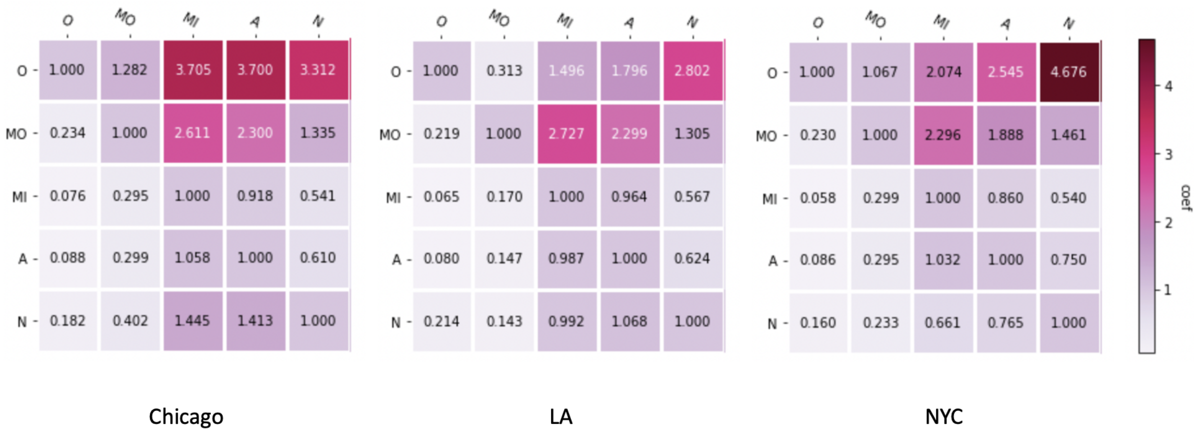
1. We build a linear model with only one independent variable. The independent variable is the complementarity in one period while the dependent variable is the complementarity in another period. We directly put the calculated complementarity values into the model, even if they have negative signs.
2. For the value 1.282, the row label (overnight) corresponds to independent variable and the column label (morning) corresponds to dependent variable.
3. The results of the linear model include a coefficient and a p-value. We display the coefficient in Fig. 27. Here, 1.282 means the complementarity of morning is 1.282 times as high as that of overnight.

It is worth mentioning that all the coefficients shown in Fig. 27 are with p-values smaller than 0.01 in their corresponding linear models. From these statistically significant coefficients, we can see that the extent of complementarity are different among periods. Generally in all cities, the complementarity during midday is higher than other periods.

3.5.2 Explain Complementarity via regression model

Based on Eq. (9), the area-level complementarity is derived from citizens' visitation to specific areas. Thus, it may be explained by the motivations of citizens' visitation. By common sense, citizens can be motivated to visit areas with

- high venue diversity within an urban area. Citizens can visit various categories of venues, such as restaurants, theaters, luxury stores.
- a large number of venues within an urban area. Even if this area only has restaurants, these restaurants are with many brands and flavors. Citizens can feel free to choose per their preferences.
- high venue density within an urban area. This means even within a small area, citizens still have lots of venues to visit.



Chicago

	O	MO	MI	A	N
O	1.000	1.282	3.705	3.700	3.312
MO	0.234	1.000	2.611	2.300	1.335
MI	0.076	0.295	1.000	0.918	0.541
A	0.088	0.299	1.058	1.000	0.610
N	0.182	0.402	1.445	1.413	1.000

LA

	O	MO	MI	A	N
O	1.000	0.313	1.496	1.796	2.802
MO	0.219	1.000	2.727	2.299	1.305
MI	0.065	0.170	1.000	0.964	0.567
A	0.080	0.147	0.987	1.000	0.624
N	0.214	0.143	0.992	1.068	1.000

NYC

	O	MO	MI	A	N
O	1.000	1.067	2.074	2.545	4.676
MO	0.230	1.000	2.296	1.888	1.461
MI	0.058	0.299	1.000	0.860	0.540
A	0.086	0.295	1.032	1.000	0.750
N	0.160	0.233	0.661	0.765	1.000

Figure 27: The coefficients of complementarity among different periods. All the coefficients are with p-values smaller than 0.01 in their corresponding linear models. The row and column labels "O", "MO", "MI", "A", "N" means overnight, morning, midday, afternoon, night.

We will examine the effect of these motivations on our calculated area-level complementarity. Before the examination, we would like to formally define the diversity by the rationale of entropy. Recall that in Section 3.2.2, for an urban area I, we define z_{Ik} as the ratio of number of venues in k-th category to total number of venues. We have a total of 10 categories. In this way, we define the venue diversity of urban area I as

$$-\sum_{k=1}^{10} z_{Ik} \log(z_{Ik}). \tag{12}$$

Essentially, the effect of aforementioned motivations can be examined via a linear regression model. The dependent variable is the complementarity from start area to end area. As mentioned in Section 3.4, a city can be considered as a graph, where each area is a node. Then, complementarity is derived by Eq. (9) in the conceptual level of a directed edge. Following this rationale, the independent variables of the linear model should also be derived in the conceptual level of a directed edge. However, venue diversity of an area, as one of the citizens' motivations, is a node-level variable, because each area is considered as a node. Potentially, citizens' visitation from the start to the end area can be jointly motivated by the characteristics of these two areas. Thus, we define a variable "joint diversity", denoted as ψ, which is the weighted sum of the venue diversity of start and end area. In this way, we derive the diversity in the conceptual level of a directed edge. Formally, to calculate the joint diversity from area I to J, it is

$$\psi_{I,J} = (1 - w_{end}) \times (\text{diversity of } I) + w_{end} \times (\text{diversity of } J) \tag{13}$$

where w_{end} is the weight of the end area and $w_{start} = 1 - w_{end}$ is the weight of the start area.

We use the same approach to define "joint number of venues" (denoted as ϕ), and "joint density" (denoted as ρ). Particularly, to firstly calculate the venue density of each area, we use the area size dataset from United States Census Bureau[3], and then the density of a specific area is the number of venues divided by the area size of that area. Then we formulate ϕ, ρ as

$$\phi_{I,J} = (1 - w_{end}) \times (\text{number of venues of } I) + w_{end} \times (\text{number of venues of } J) \tag{14}$$

$$\rho_{I,J} = (1 - w_{end}) \times (\text{density of } I) + w_{end} \times (\text{density of } J) \tag{15}$$

We are also interested in the effect of inter-area distance on complementarity. That is, we define the distance between two areas' centroids as d_{cen}, which also becomes one independent variable in our linear regression model. Additionally, the complementarity is calculated under a specific time period in a specific city. To express this effect, we define another variable "city and period", denoted as *cp*. The purpose of *cp* is to distinguish complementarity in different cities and periods. It has a total of 15 values; for example, the value "Chicago midday" means the complementarity is for Chicago during midday period.

Our purpose is to use a linear model to summarize the effect of all independent variables. However, the value of w_{end} in Eq. (13)(14)(15) is yet to be decided, since it influences independent variables: ψ, ϕ, ρ. Our idea to decide it is to put all independent variables into a linear model in order to conduct a grid search of candidate values of w_{end}. The formula for the linear model for the grid search is

$$c \sim \psi + \phi + \rho + d_{cen} + \psi : \phi + \psi : \rho + \psi : d_{cen} + cp. \tag{16}$$

[3] https://www.census.gov/

After the grid search (elaborated in Appendix C), w_{end}=0.65 is chosen since it generates the best linear model.

Now that we have chosen w_{end}, the values of $\psi, \phi, \rho, d_{cen}$ are decided. We would like to focus on them as independent variables for a linear model. At the same time, we would not like to focus on *cp* as one of the independent variables. This is because it is already expected that the complementarity value highly depends on a specific city.

Following such purposes, our approach is to build a mix-effect model [6], since it can explicitly use $\psi, \phi, \rho, d_{cen}$ as independent variables while the effect of *cp* is only implicitly presented. The basic idea of a mix-effect model is very similar to a linear model. The main difference is that we can select some variables as random effects to be only implicitly reflected in the model summary table. More specifically, our formula to build the mix-effect model is

$$c \sim \psi + \phi + \rho + d_{cen} + \psi : \phi + \psi : \rho + \psi : d_{cen} + (1|cp). \tag{17}$$

In this formula, $(1|cp)$ means that the variable *cp* is selected as the random effect to implicitly influence the intercept of the model. Then, the mix-effect model summary is shown in Table 9, where *cp* is not explicitly listed as one independent variable. Instead, its effect is implicitly reflected in the coefficient and the standard error of the intercept; in other words, the intercept is decided given a specific *cp* value. This way, we can explicit show the effect of the independent variables we would like to focus on, as listed in Table 9.

From Table 9, we can find that the p-values for all independent variables (including interaction terms) are smaller than 0.01, which means that all independent variables are statistically significant w.r.t. the dependent variable - complementarity. Therefore, complementarity can be explained by diversity, number of venues, density of urban areas, the distance among urban areas, and the interactions among these independent variables.

Another fining is that the diversity value influences the coefficients of other factors. From Table 9, let's firstly analyze the interaction between diversity and density. Assume that the diversity has a relatively high value 2.5. Then the coefficient for density is -0.072 + 0.036×2.5 = 0.018, which means the complementarity increases as the density of the start and end areas increase. The reason can be that the high diversity in the start area leads to lots of existing citizens. As the diversity of the end area is also high, these citizens are motivated to visit the

Table 9: A mixed-effects model where the dependent variable is complementarity in the postal area level.

variable meaning	**variable**	**coefficient**
intercept	intercept	142.800***
		(10.550)
diversity	ψ	-53.700***
		(3.596)
number of venues	ϕ	-1.233***
		(0.0273)
diversity: number of venues	$\psi : \phi$	0.542***
		(0.012)
density	ρ	-0.072***
		(0.008)
diversity: density	$\psi : \rho$	0.036***
		(0.004)
distance	d_{cen}	3.775***
		(0.322)
diversity: distance	$\psi : d_{cen}$	-2.248***
		(0.149)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

end area, causing the number of movements to exceed the expectation. However, when the diversity has a relatively low value 1.5, then the coefficient for density is -0.072 + 0.036×1.5 = -0.018. This means the complementarity decreases as the density of the start and end areas increase. The reason can be that the low diversity in the start area leads to only a small number of existing citizens. As the diversity of the end area is also low, these citizens are not motivated to visit the end area, causing the number of movements to be lower than the expectation. We can find similar effects from diversity to number of venues and distance. This reminds city planners to check the diversity of urban areas first before taking measures on other factors for better complementarity.

We would also like to emphasize that the aforementioned independent variables are jointly derived by the statistics in the start area and end area. This indicates that to improve the complementarity, city planners may need to take measures on both the start area and the end area.

3.5.3 Embedding patterns among periods

The aforementioned results show that complementarity values vary among periods, and can be explained by some edge-level independent variables. We are also interested in some informative node-level characteristics. That is, similar to Section 3.2.2, we will fetch the embeddings in the node (urban area) level. In the current section, these embeddings will correspond to the dependent variable - complementarity. Additionally, as complementarity is period-specific, the urban area embeddings will also be period-specific. In this way, we can further compare the differences of area embeddings among periods.

To obtain the area embeddings, we may need some straightforward area-level independent variables to start with. Recall that we have indirectly used the area-level features in Eq. (13)(14)(15) to build the mix-effect model shown in Table 9. But we may need more variables in finer granularity; particularly, the information in each venue category per area can be useful. Additionally, our mix-effect model in Table 9 is basically only a descriptive linear regression model. It may not be capable of modeling the complementarity accurately. Thus, we need another approach to model the complementarity. Instead of a linear model, we are interested in a non-linear way to potentially model the complementarity more accurately.

Design the graph neural network

Neural network is a type of candidate model since it is more capable of capturing non-linear relationship between dependent and independent variables. Additionally, we can customize the dimensions of the embeddings to potentially improve the accuracy of the model. Then, we will design a specific neural network where:

- in the input layer, we need some independent variables in the node (area) level, since they are relatively straightforward to start with.
- in the intermediate layers, there are embeddings, which are the exact variables we aim at fetching.
- in the output layers, the dependent variable is the complementarity in the edge level.

Since all our discussed variables are related to the sense of "graph", we use the concept of graph neural network (GNN) [79] to design our model, which is illustrated in Fig 28. The input is in the node level. Particularly for each node (area), we use the number of venues in each venue category as the input feature vector. As we have a total of 10 categories, the dimension of the input vector is 10. Then each input vector goes through 4 hidden layers; in this way, each node (area) is represented as an embedding. However, complementarity is the dependent variable. That is, we need to transform the node-level embeddings to the edge-level complementarity. Our approach is to concatenate the embedding of two nodes, which becomes the embedding of their corresponding edge. Then we let this edge-level embedding go through a fully connected layer, in order to transform this embedding to a single number, i.e. predicted complementarity. The mean-square-error (MSE) loss is calculated based on this predicted complementarity and the ground truth of complementarity. We train each model for a total of 40,000 epochs. In each epoch, all nodes and edges are within the same batch. In this way, we can make sure that in order to calculate a predicted edge-level value (complementarity), the two nodes forming this edge are available in the same batch; additionally, in order to calculate the MSE loss, the ground truth corresponding to this predicted edge-level complementarity is also available in the same batch.

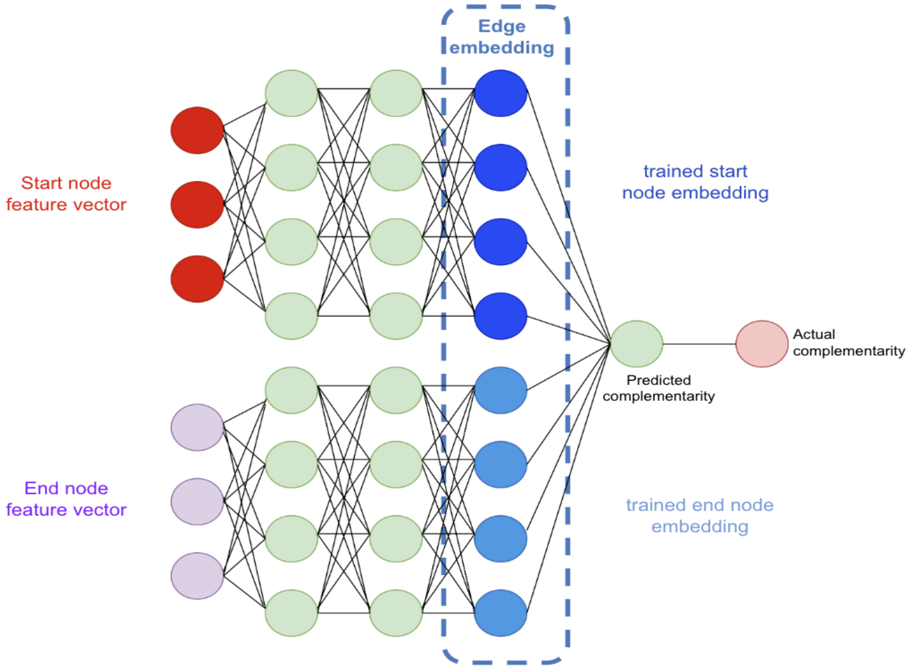

Edge embedding
Start node feature vector
trained start node embedding
Actual complementarity
Predicted complementarity
End node feature vector
trained end node embedding

Figure 28: Graph neural network structure.

It is worth mentioning that though the `node2vec` approach mentioned in Section 3.2.2 is well-known for obtaining the embeddings for nodes in a graph, it is not applicable in the current situation where the complementarity is modeled as an edge. The key reason is that `node2vec` can only handle non-negative edge values while the complementarity values can be negative. Therefore, we use the GNN to model the graph where the edge represents complementarity.

Select some edges for training

Straightforwardly, we may input all edges (complementarity values) for training. However, this may lead to a biased model. For example, as shown in the distribution of complementarity for Chicago midday period (Fig. 29), the majority of complementarity values are relatively close to 0. If we input all edges to train the GNN of Chicago midday period, this GNN will be biased to model almost all complementarity values to be close to 0. Our idea to

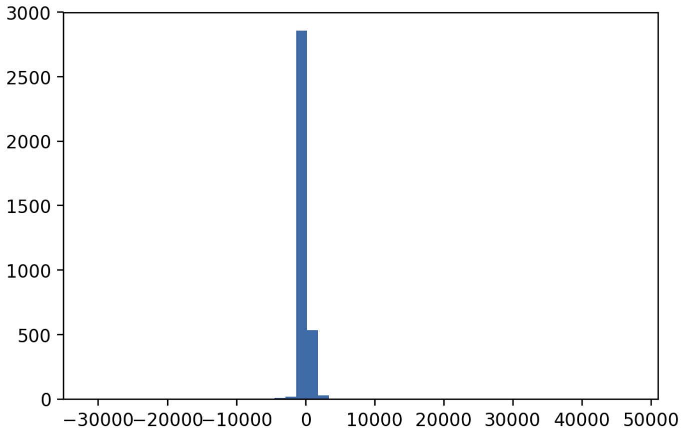


Figure 29: The distribution of complementarity values for Chicago midday period.

resolve this is, out of all edges, to select two sets of edges: (1) edges with weak complementarity (denoted as $\mathbb{E}_w^{MI}$ for Chicago midday period); (2) edges with strong complementarity including negative values (denoted as $\mathbb{E}_s^{MI}$ for Chicago midday period). Additionally, the size of these two sets should be relatively balanced, i.e., they should have similar sizes. Then we only input edges of these sets into the model for training.

The one-sample t-test is an intuitive way to identify strong or weak values, which is helpful for selecting edges for these two sets. Recall that by Eq. (10), a single complementarity value $c_{I,J}$ is calculated by 20 values $c_{I,J}^q$ from 20 simulations. The one-sample t-test can be used to compare whether the average of 20 values for an edge (the corresponding single complementarity value) is significantly different from 0. Ideally, when the output p-value of the t-test is smaller than 0.1, it means the average of the complementarity is significantly different from 0, and this edge may belong to $\mathbb{E}_s^{MI}$. However, only using the t-test cannot precisely identify edges for $\mathbb{E}_s^{MI}$. The reason is that, t-test is too strict with the difference between 0 and the average of 20 complementarity values. For example, we assume a single complementarity value is generated by 20 simulated values, i.e., an array of ones {1, 1, ..., 1, 1}. By conducting t-test on this array, the output p-value is 0.000, which means the average of the array is significantly different from 0. In this way, the t-test decides to assign this edge into the set $\mathbb{E}_s^{MI}$ with strong complementarity, but from Fig. 29, we know that the complementarity value 1 should not be recognized as strong.

Apart from t-test, we will conduct extra assistance. The key ideas of our extra assistance are: define an intermediate metric for the strength of complementarity, and set a threshold. Since the input of the t-test involves the complementarity $c^q_{I,J}$ derived from the q-th simulation, our assistance will also process every $c^q_{I,J}$ for $q \in [1, 20]$. Recall that one interpretation of complementarity is that, per Eq. (9), complementarity $c^q_{I,J}$ is the extent of over-performing or under-performing. That is, if $c^q_{I,J} \geq 0$, then $c^q_{I,J}$ is the extent of out-performing; otherwise, $|c^q_{I,J}|$ is the extent of under-performing. Also, the value $c^q_{I,J}$ (or $|c^q_{I,J}|$) itself may not be sufficient for us to sense the strength of over-performing (or under-performing). But the context of that specific edge is helpful. To explain this, some cases are provided in Table 10:

- In the 1st case, the extent of over-performing is $c^q_{I,J}$=50. The basic reason to *determine* the trend of over-performing is that the actual movement number $a_{I,J}$ (60) is *larger* than its counterpart $s^q_{I,J}$; this is the context. Then by 50/60=0.833, we can find that the extent of over-performing (50) occupies a large portion (ratio 0.833) of its basic reason (60). This provides a large and strong sense of over-performing. Thus, this edge may be suitable to belong to the set with strong complementarity $\mathbb{E}^{MI}_s$.
- The 2nd case is different, though the extent of over-performing is also $c^q_{I,J}$=50. Here, the basic reason to *determine* the trend of over-performing is that the actual movement number $a_{I,J}$ (10000) is *larger* than its counterpart $s^q_{I,J}$; this is the context. Then by 50/10000=0.005, we can find that the extent of over-performing (50) occupies only a small portion (ratio 0.005) of its basic reason (10000). This provides a small and weak sense of the over-performing. Thus, this edge may be suitable to belong to the set with weak complementarity $\mathbb{E}^{MI}_w$.
- The 3rd case is also different, where the complementarity is $c^q_{I,J}$=-50, i.e., the extent of under-performing is $|c^q_{I,J}|$=50. Here, the basic reason to *determine* the trend of under-performing is that the simulated movement number $s_{I,J}$ (60) is *larger* than its counterpart $a_{I,J}$; this is the context. Then by 50/60=0.833, we can find that the extent of under-performing (50) occupies a large portion (ratio 0.833) of its basic reason (60). This provides a large and strong sense of under-performing. Thus, this edge may be suitable to belong to the set $\mathbb{E}^{MI}_s$.

Table 10: Example cases to explain the outlier identification approach.

case index	**complementarity** $c^q_{I,J}$	**actual** $a_{I,J}$	**simulated** $s^q_{I,J}$
1st	50	60	10
2nd	50	10000	9950
3rd	-50	10	60

In the aforementioned examples, we calculate several ratios, such as 0.833, 0.005. We formally define such ratio as Complementarity Strength Ratio (CSR), which is only used in the current section (Section 3.5.3) to select two sets of edges $\mathbb{E}^{MI}_w$ and $\mathbb{E}^{MI}_s$ for training. The rationale of CSR is, as an intermediate metric, to provide a sense of strength of complementarity compared to the its basic reason. To formulate it:

$$CSR^q_{I,J} = \begin{cases} c^q_{I,J}/a_{I,J} & \text{if } c^q_{I,J} \geq 0 \\ |c^q_{I,J}|/s^q_{I,J} & \text{otherwise} \end{cases} \tag{18}$$

Based on the above cases, we already define the intermediate metric CSR, which will be the input to the t-test. However, this is not sufficient to appropriately select edges into the set with strong complementarity $\mathbb{E}^{MI}_s$. For example, in the 1st case, the t-test can only make sure 0.833 is significantly different from 0, but cannot judge whether 0.833 is sufficiently strong. Our approach for this is to set a threshold to evaluate CSR. In fact, per Eq. (18), we find that $0 \leq CSR^q_{I,J} \leq 1$ is always true. So, a reasonable threshold value is between 0 and 1. We set 0.75 to be the threshold. That is, if the average of CSR is larger than 0.75, then we assign that edge into the set with strong complementarity $\mathbb{E}^{MI}_s$.

As follows, we summarize the process to select edges into the two sets $\mathbb{E}^{MI}_w$, $\mathbb{E}^{MI}_s$ for the midday period (MI is referred to as the midday period). But this exact process is applicable to any period if the complementarity values are from that period, which is also marked as "Selected edge sets per period" in Fig. 30.

1. For midday period, each edge (from area I to J) has an array of complementarity $\{c^1_{I,J}, c^2_{I,J}, ..., c^Q_{I,J}\}$. Based on Eq. (18), we generate another array $\{CSR^1_{I,J}, CSR^2_{I,J}, ...,$ $CSR^Q_{I,J}\}$
2. For each edge, input the array $\{CSR^1_{I,J}, CSR^2_{I,J}, ...,$ $CSR^Q_{I,J}\}$ into the one-sample t-test process. This purpose is to verify whether the average of CSR for the current edge is significantly different from 0.
   - If the output p-value from the t-test is smaller than 0.1, it means the average of CSR for the current edge is significantly different from 0. Then we check whether the average of CSR is larger than the threshold 0.75. If so, we assign this edge to the set with strong complementarity $\mathbb{E}^{MI}_s$.
   - If the output p-value from the t-test is larger than 0.1, it means the average of CSR for the current edge is not significantly different from 0. We assign this edge to the set with weak $\mathbb{E}^{MI}_w$.

After this process, for each city, we can get a total of 10 selected edge sets from all 5 periods. For overnight, we get $\mathbb{E}^O_s$, $\mathbb{E}^O_w$. For morning, we get $\mathbb{E}^{MO}_s$, $\mathbb{E}^{MO}_w$. For midday, we get $\mathbb{E}^{MI}_s$, $\mathbb{E}^{MI}_w$. For afternoon, we get $\mathbb{E}^A_s$, $\mathbb{E}^A_w$. For night, we get $\mathbb{E}^N_s$, $\mathbb{E}^N_w$.

However, to train the model for midday period, we cannot only input the sets $\mathbb{E}^{MI}_s$, $\mathbb{E}^{MI}_w$ to the model. The reason is that our objective is to compare the difference of embeddings among periods. The comparison needs to be fair. That means, the selected edges to be trained should be the same for all periods, i.e., all models. But since CSR is derived from period-specific complementarity, CSR is also period-specific, causing $\mathbb{E}^O_s$, $\mathbb{E}^{MI}_s$, $\mathbb{E}^{MO}_s$, $\mathbb{E}^A_s$, $\mathbb{E}^N_s$ to be different. For example, if the edge from I to J belongs to $\mathbb{E}^{MO}_s$ (midday), it may not belong to $\mathbb{E}^N_s$ (night). To solve this problem, our approach is to take the union of $\mathbb{E}^O_s$, $\mathbb{E}^{MI}_s$, $\mathbb{E}^{MO}_s$, $\mathbb{E}^A_s$, $\mathbb{E}^N_s$; this union set for strong complementarity is denoted as $\mathbb{E}_s$. Similarly, we take the union of $\mathbb{E}^O_w$, $\mathbb{E}^{MI}_w$, $\mathbb{E}^{MO}_w$, $\mathbb{E}^A_w$, $\mathbb{E}^N_w$; this union set for weak complementarity is denoted as $\mathbb{E}_w$. Then, to train the model for any period, $\mathbb{E}_s$ and $\mathbb{E}_w$ together are the input edge sets. This process is illustrated as the whole diagram in Fig. 30. The sizes of selected edges are shown in Table 11.

Results: embedding differences among periods

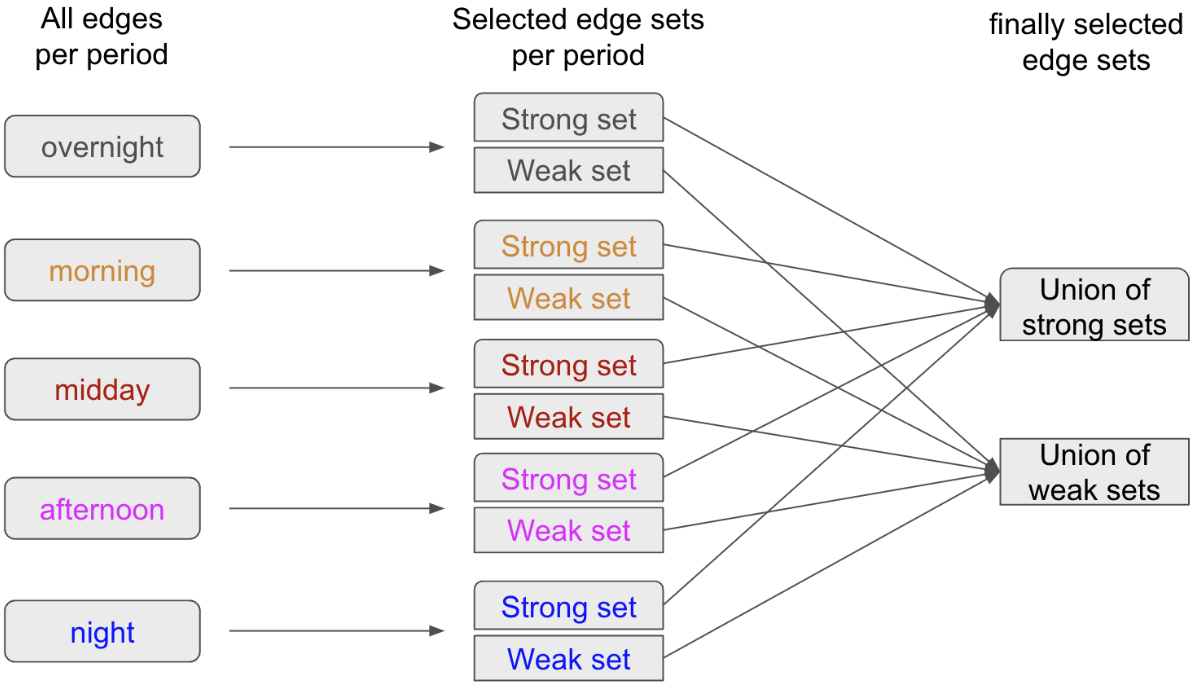


Figure 30: Process of selecting edges for training.

Table 11: Size of selected edge sets for training in each city. The selected edges consist of $\mathbb{E}_s$ and $\mathbb{E}_w$.

city	**size of $\mathbb{E}_s$**	**size of $\mathbb{E}_w$**	**number of selected edges**	**number of**
			(sum size of $\mathbb{E}_s$ and $\mathbb{E}_w$)	**all edges**
NYC	6494	5920	12414	18769
Chicago	806	663	1469	3481
LA	4274	4354	8628	11881

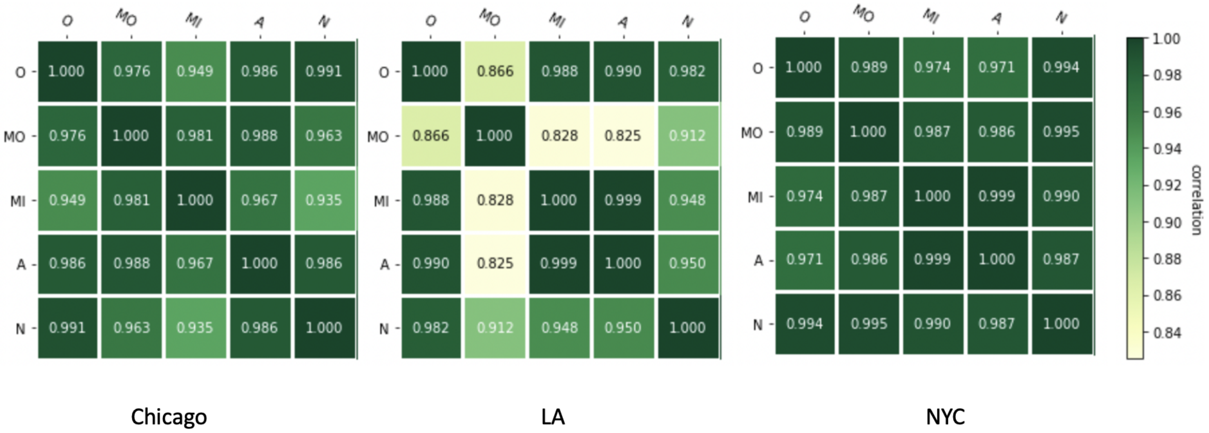


Figure 31: Correlation among area embeddings of different periods. The row and column labels “O”, “MO”, “MI”, “A”, “N” means overnight, morning, midday, afternoon, night.

As we finish the model training for each period and fetch the corresponding node embeddings, here we follow similar steps as described in Section 3.2.3 to compare the embeddings between two periods. The similarity of two areas I and J, with embedding $\mathbf{x}_I$ and $\mathbf{x}_J$ respectively, is defined as:

$$\sigma_{IJ} = \texttt{dist}(\mathbf{x}_I, \mathbf{x}_J) \tag{19}$$

where $\texttt{dist}(\mathbf{x}_I, \mathbf{x}_J)$ is the (Euclidean) distance between the embedding of area I and J. We can also interpret such similarity as inter-area relationship. Then we examine whether embeddings in two periods for the urban areas provide different views for their similarity. In particular, we denote σ^A_{IJ} and σ^N_{IJ} as the similarities between the pair of areas I and J using embeddings from afternoon and night, respectively. Then, for afternoon, we can calculate such similarity between every pair of areas; we conduct the same calculation for night period. This way, we obtain and denote σ^A as an array of all pairwise similarities for the afternoon period, and we denote σ^N as an array of all pairwise similarities for the night period. Their Pearson correlation coefficient ρ_{σ^A,σ^O} will be high if the embeddings from two periods provide similar information, and low otherwise. We apply the same process between any two periods.

Additionally, we set 32 as the embedding dimension; the reason is elaborated in Appendix D.

The results are shown in Fig. 31. One can observe that for these three cities, the correlations between any two periods are very high, all over 0.80. This means that, the inter-area relationships (to model the complementarity) are similar regardless of the time of a day. Since New York City and Chicago are more geographically compact, it is easier for citizens to move within the city for any purpose at any time. This could be the reason that generally the inter-area relationships within a day are similar. Los Angeles has slightly different patterns. Particularly, the correlations related to morning are smaller than 0.90 (i.e., 0.866, 0.828, 0.825 in lighter colors in Fig. 31), while other correlations are larger than 0.90. The reason may be that Los Angeles is geographically scattered, which introduces a conflict between long commuting distances and morning rush hour. That is, for a commuter, even if the home area is very far from the office area, she has to get over such a long distance to reach the office in the morning. This leads to very high complementarity from home areas to office areas, which can be the reason of relatively unique pattern for the morning period. But for other periods, even if a citizen should travel long distance to her final destination, she usually does not need to hurry; she may also travel a short distance to another intermediate venue before heading to the final destination.

3.6 Summary

In this chapter, we explore the extent and the pattern of excess demand in urban business entities.

Firstly, we propose `hood2vec`, which maps urban areas into the latent spaces. We demonstrate how the excess demand can potentially influence the similarity among urban areas.

Next, we propose our simulation approach to estimate the capacity demand where demands are only affected by distances and venue ratings. We consider the real-world check-in data as the total demand. By choosing the complementarity as the proxy for excess demand, we estimate the extent of complementarity (i.e., excess demand) using the difference between total demand and capacity demand.

Then, among all calculated complementarity, we provide examples of complementarity and corresponding interpretation for them from multiple perspectives. From all calculated complementarity, every source of complementarity can be provided for a specific urban entity. In order to improve the business of a specific entity, the venue owner or city planner can choose and prioritize the measures to be taken based on each source of complementarity.

We also analyze the statistical patterns of area-level complementarity. Firstly, we find that the extent of complementarity is significantly different among periods. Next, we build a mix-effect model where the complementarity is the dependent variable. We find that the area density, number of venues, inter-area distance are statistically significant independent variables, but their coefficients are influenced by the area diversity. Additionally, these independent variables are jointly derived by the statistics in the start area and end area. Such results reminds city planners that to improve the complementarity between two areas, city planners may need to consider the influence of area diversity, and then take measures on both the start area and the end area.

Then, we fetch the area embeddings via a graph neural network where the complementarity is the dependent variable. By analyzing the embeddings, we find that generally for Chicago, NYC, LA, the inter-area relationships are similar among different periods in a day, while the patterns related to LA morning period are slightly different from other periods.

4.0 Excess demand of the urban transportation

In this chapter, we will firstly estimate the excess demand of each station of a bike sharing system. We propose to use a special time interval named “excess demand pulse” as the proxy for the estimation. Next, we build a Skellam model to predict the net total demand of the bike sharing system, which is advantageous over other alternative models on predictive power and interpretability.

Proxy for capacity demand: Before elaborating our proxy for excess demand, we introduce our selection of the proxy for capacity demand first. Our selected proxy is the observed demand in each station, which directly comes from the rental and return records in the system logs. To iterate, the capacity demand is the demand volume that the operator expects. The current observed demand in the system logs should generally meet the operator’s expectation. Otherwise, it is easy for the operator to keep re-constructing the bike sharing system until the observed demand in the system logs meets her expectation. In the remaining part of this chapter, we directly use the term of the proxy “observed demand” to express “capacity demand”.

To calculate the total demand, we will sum up the capacity demand (through the proxy: observed demand) and the estimated excess demand. We will elaborate the excess demand estimation and total demand calculation in the remaining part of this chapter.

4.1 Excess demand estimation

In order to capture the pattern of excess demand and quantify it, firstly, we would like to select a statistical distribution to model the bike flows in the bike-sharing system. While several distributions have been used to model the bike arrivals (and departures) within a bike sharing, including negative binomial [16], Weibull [44, 55, 74] and Poisson [63, 35, 65, 18, 12, 33], the latter is the most common choice for this task. Gast *et al.* [30] show through a Kolmogorov-Smirnov test [69] that the trips in the Paris bike sharing system follow a Poisson distribution. In the following Section 4.2, we use a similar approach to show that the trips in our dataset fit a Poisson distribution as well.

In bike-sharing system, it is not meaningful to use complementarity as the proxy for excess demand, since a bike station is almost complementary to all types of business venues. In other words, all demands on a bike station are due to complementarity. Therefore, to iterate, the excess demand of bike sharing system is the unobserved demand we introduce in Chapter 1. More specifically, excess demand is not captured in the transaction logs in the bike-sharing system, since it appears when there is zero supply. Hence, it is very challenging to estimate it. In this section, borrowing ideas from queuing theory, we will introduce a way to estimate the excess demand. We further simulate the bike rental and return process to show the ability of the proposed approach to estimate the excess demand in a bike sharing system. Then, we apply our approach on data obtained from a real bike sharing system, Chicago's Divvy, to estimate the excess demand present in the system. Notations used in describing our approaches and models through this chapter are shown in Table 12.

At a bike station, we generally have two types of event flows occurring as illustrated in Fig 32. One flow represents the bike departure (rental) events, with the number of departures per time unit following a Poisson distribution with intensity μ [30]. This also means that the inter-departure time intervals follow exponential distribution with an average of $\frac{1}{\mu}$. The other flow represents the bike arrival (return) events, with the number of arrivals per time unit following Poisson distribution with intensity λ (and similarly the inter-arrival time intervals follow an exponential distribution with average $\frac{1}{\lambda}$). Under the assumption of the flows being independent, we can consider their union as a single flow with mixed types of events [9]. In this mixed flow, the number of events per time unit follows a Poisson distribution with intensity $(\lambda + \mu)$, while the inter-event time intervals follow an exponential distribution with average $\frac{1}{\lambda+\mu}$.

Let us assume that the number of available bikes at a station is a. Fig 33 shows a segment of the bike availability curve, where a changes from 0 to 1 after a bike arrival at t_1, and goes back to 0 after a rental at t_2. This pattern is central to our estimation of bike excess demand rate (denoted with μ_e), and we refer to this curve pattern as excess demand pulse (EDP). The EDP is our proposed proxy to quantify the excess demand in bike-sharing

Table 12: A list of notations used through Chapter 4.

Symbol	**Description**
μ	actual bike departure rate by total demand
$\hat{\mu}$	estimated bike departure rate by total demand
μ_e	actual bike departure rate by excess demand
$\hat{\mu}_e$	estimated bike departure rate by excess demand
λ	actual bike arrival rate by total demand
$\hat{\lambda}$	estimated bike arrival rate by total demand
λ_e	actual bike arrival rate by excess demand
$\hat{\lambda}_e$	estimated bike arrival rate by excess demand
a	number of available bikes
τ_f	EDP length
τ_m	the average of multiple τ_f
τ_s	the average of inter-supply intervals
t_{end}	the end time stamp of the availability curve
$t_a, t_b, t_c, t_d, t_1, t_2, t_3, t_4$	specific time stamps of the availability curve
N_μ	total bike demand volume
N_λ	total dock demand volume
Z	net total demand volume
N_{μ_o}	observed bike demand volume
N_{μ_e}	excess bike demand volume
N_{λ_o}	observed dock demand volume
N_{λ_e}	excess dock demand volume
l_{μ_e}	duration length for bike excess demand in a 30-minute interval
l_{λ_e}	duration length for dock excess demand in a 30-minute interval

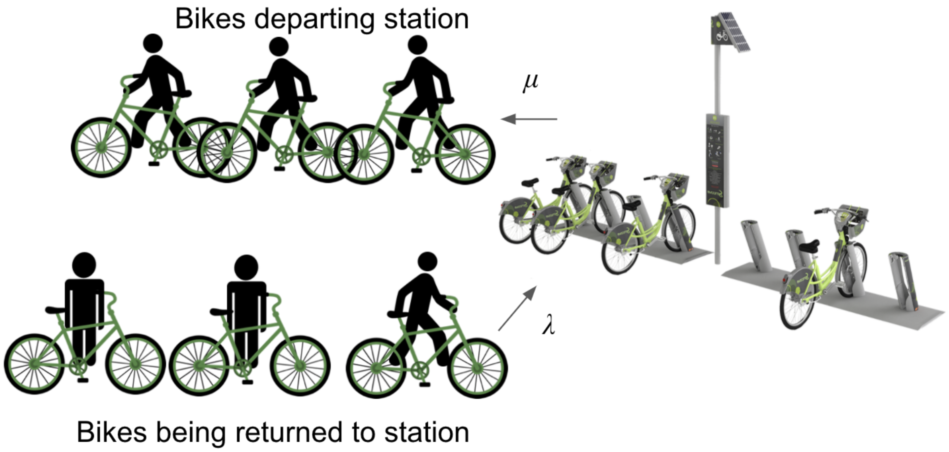


Figure 32: Bike departure and arrival event flows at a bike station.

systems. We also define $\tau_f = t_2 - t_1$ as EDP length. During the interval $(0, t_1)$, the bike availability is constantly 0, which can be interpreted by someone that there are no events (rentals or returns) happening during that time. However, this is not necessarily true. This constant 0 availability can indeed be due to no events happening during this interval, or due to failed bike rentals, that is, a customer tried to rent a bike but none was available. The pattern captured by the EDP serves as an important signal for the possible presence of excess demand and its degree. Intuitively, the presence of significant excess demand leads to situations where any supply that becomes available is consumed shortly thereafter. At the situation visualized in Fig 33 when the single bike arrives at t_1, it is quickly consumed (rented) at time t_2. In contrast, if we consider the scenario presented in Fig 34, a bike arrives at t_a but it is not consumed *quickly*. Instead, another bike arrives at t_b before a rental. Therefore, any bike demand in this case can be captured well from rental logs, and it is not excess. In other words, the pattern in Fig 34 does not provide evidence for the existence of excess demand.

Using these observations let us see how we can estimate μ_e through the bike availability curves. Fig 35 depicts a segment of the bike availability curve. Recall that the mixture of arrival and departure flows follows a Poisson distribution with intensity $(\lambda + \mu)$. That is, the

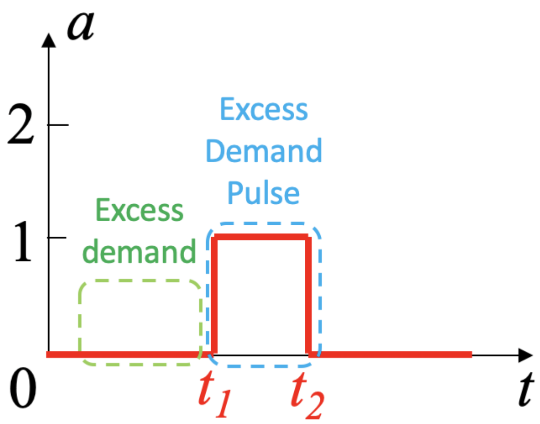


Figure 33: This bike availability curve indicates possible excess demand for $t \in (0, t_1)$.

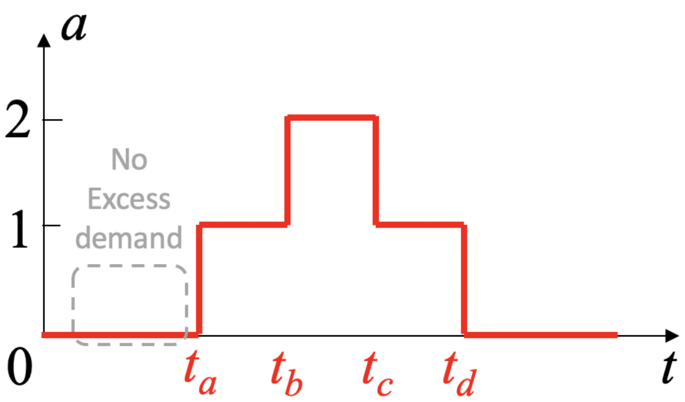


Figure 34: This bike availability curve indicates no excess demand for $t \in (0, t_a)$.

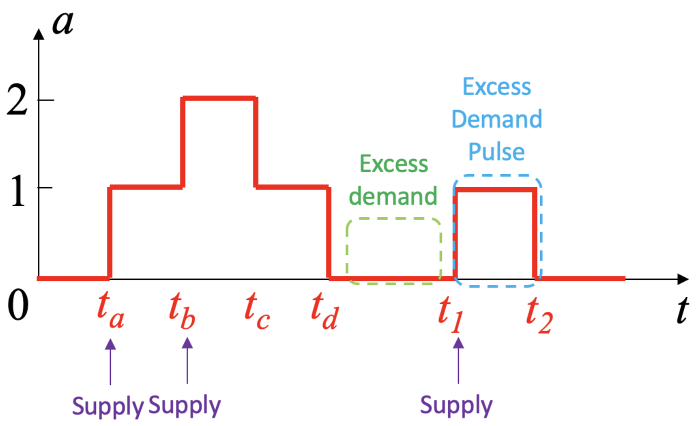


Figure 35: A segment of bike availability curve to illustrate the estimation of excess demand.

inter-event intervals of this mixture follow exponential distribution with intensity $\frac{1}{\lambda+\mu}$. If we observe an arrival event followed by departure event, such observation is caused by mixing arrival and departure flows. Thus, in such observation, the interval from the arrival to the departure event follows an exponential distribution with intensity $\frac{1}{\lambda+\mu}$. Thus, $\tau_f = t_2 - t_1$ is a sample from an exponential distribution with average $\frac{1}{\lambda+\mu}$. During a large observation period we will observe τ_F from multiple EDPs, denoting their average value as τ_m. By expectation, we should get $\tau_m \approx \frac{1}{\lambda+\mu}$. That is, the estimated intensity of the mixed flow is $\hat{\lambda} + \hat{\mu} = \frac{1}{\tau_m}$.

We can also calculate the estimated arrival rate $\hat{\lambda}$ from the data. In this paper, we focus on bike sharing systems with docks, so while there is a possibility for *excess supply* in a bike station - e.g., a user tries to return a bike to a full dock - this is not an issue in the presence of bike excess demand. In general, there cannot be bike excess demand and excess supply at the same station during the same time. Therefore, each bike supply (i.e., bike arrival) event is successfully reflected in the bike availability curve when there is bike excess demand present. To reiterate, the inter-supply (i.e., inter-arrival) intervals themselves follow exponential distribution with intensity $\frac{1}{\lambda}$. By obtaining all inter-arrival intervals from the data we can estimate their average denoted as τ_s. For example, in the segment in Fig 35, we have arrivals at t_a, t_b, t_1, resulting in $\tau_s = \frac{(t_b - t_a) + (t_1 - t_b)}{2}$. By expectation, we should get $\tau_s \approx \frac{1}{\lambda}$, i.e., the estimated arrival rate is $\hat{\lambda} = \frac{1}{\tau_s}$.

Combining the two results above, the excess demand rate μ_e can now be estimated as $\hat{\mu}_e = (\hat{\lambda} + \hat{\mu}) - \hat{\lambda} = \frac{1}{\tau_m} - \frac{1}{\tau_s}$. However, it is possible that $\frac{1}{\tau_m} < \frac{1}{\tau_s}$. This happens when the inter-arrival intervals are very short, i.e., departure rate is relatively low compared with arrival rate. However, such low departure demand indicates there is not really any excess bike rental demand, or in other words the total demand can be reflected by the rentals observed. Finally, combining all of the above observations, the estimated excess demand rate is given by:

$$\hat{\mu}_e = \max(\frac{1}{\tau_m} - \frac{1}{\tau_s}, 0) \tag{20}$$

Evaluation on synthetic data: Since we do not have the ground truth for the excess demand in real data (i.e., people that attempted to rent a bike but the station was empty), we rely on simulations to evaluate whether Eq (20) is able to accurately estimate μ_e. Our simulator begins with 0 available bikes at time $t = 0$ and ends at t_{end}. The simulator operates as follows:

- **Time to next event**: We sample an exponential distribution with average $\frac{1}{\lambda+\mu}$, to generate a random interval τ_r that represents the time duration until the next event (either an arrival or a departure).
- **Event type**: We next have to *decide* the type of event happening. For this we sample a number r_e from a uniform distribution between 0 and 1. If $r_e < \frac{\lambda}{\lambda+\mu}$ we label the next event as an arrival, otherwise it is a departure. We also update the count of available bikes a.
- **Excess demand**: If $a = 0$, i.e., there are no available bikes, the next event cannot be a departure. Every time (when $a = 0$) the next event is simulated as a departure, we mark it as a failed bike departure. This will allow us to simulate the ground truth for the excess demand.

We simulate 1,000 time points (i.e., $t_{end} = 1000$ hours), while we use $\mu = 3$ bikes/hour, $\lambda = 1$ bikes/hour. By setting $\mu > \lambda$, we can create several situations where the bike rental demand cannot be fulfilled hence generating excess demand. Finally, we repeat the simulation 400 times.

In each simulation we collect the following information:

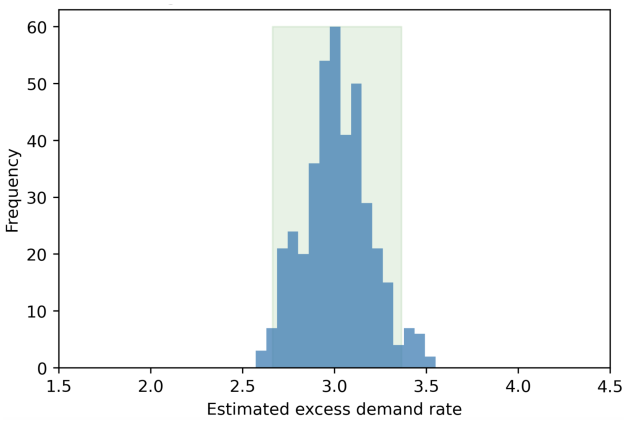


Figure 36: Histogram of estimated excess demand rate.

- The average τ_s of all the inter-arrival intervals.
- The average τ_m of all EDP lengths (i.e., $t_2 - t_1$ in Fig 35).
- We estimate the excess demand rate $\hat{\mu}_e$ using Eq (20).

In our setting, since we assume that the demand is constant at 3 bikes/hour, the excess demand is also 3 bikes/hour. Simply put, even if we do not observe any departure for a prolonged period of time in our simulation when $a = 0$, there will be a constant demand of 3 bikes/hour during these intervals. Fig 36 depicts the distribution of $\hat{\mu}_e$ from each of our simulations. As we can see the distribution is centered around 3 bikes/hour, with an average of 3.014 bikes/hour (95% CI [2.66, 3.37]). Simply put, the proposed approach is able to estimate the true excess demand in our simulations, showcasing its appropriateness for the task at hand.

4.2 Excess demand in real data

Next we are interested in applying the aforementioned approach of excess demand estimation to data from a bike sharing operator. We use data from Divvy, the bike sharing system in Chicago, and in particular we collect:

- Historical bike trip records recorded on the system[25]. A bike trip record is a tuple including the following information: <start station ID, start station name, end station ID, end station name, start time stamp, end time stamp>.
- Historical bike station status data using the Chicago Data Portal API[19]. A record of station status is a tuple of the following form: <time stamp, station ID, station name, station coordinate, number of available bikes, number of free docks, number of docks occupied by bikes>. The status of each station is recorded every 10 minutes.
- Weather data from Openweathermap[58]. Each record is a tuple including the following information: <time stamp, temperature, humidity, pressure, descriptive weather conditions>.

Distribution of Trips in Chicago's Divvy: Through our analysis above we have assumed that the trips' departures and arrivals follow a Poisson distribution. We now statistically examine the validity of this assumption. More specifically, for a given station j and a given time period t (e.g., 9-9:30am), we first focus on the number of departure trips $n_{j,t}$. By daily collecting observations for $n_{j,t}$ during a given quarter (in order to avoid seasonality), we obtain a sequence $\{n_{j,t}\}$. We calculate the average $\hat{n}$ of this sequence. We consequently repeatedly sample a Poisson distribution with mean $\hat{n}$ to generate $B = 500$ sequences of the same length as the observed one denoted as $\{r_{j,t}\}$. We then compare the distribution of the observed departures $\{n_{j,t}\}$ and the Poisson sampled ones $\{r_{j,t}\}$ using two-sample K-S test [69]. Repeating this process for every station j we obtain the average p-value $\hat{p}_j$ for the null hypothesis that the observed sequence follows a Poisson distribution. Fig 37 (left) visualizes the distribution of these p-values for all the stations in the Divvy system. As we can see they are all larger than 0.2, which means that the test cannot reject the hypothesis that the observed data follow a Poisson distribution. We repeat the same process for the arrival events and Fig 37 (right) presents the results, where we can see that again we cannot reject the null hypothesis of the arrival data following a Poisson distribution.

These results verify that we cannot reject the hypothesis that the observed bike demand and supply in the Divvy system follow a Poisson distribution. However, we also make the assumption that the excess demands follow a Poisson distribution (possibly with a different rate). Given the sparsity of the excess demand data for each station and time period, the K-S

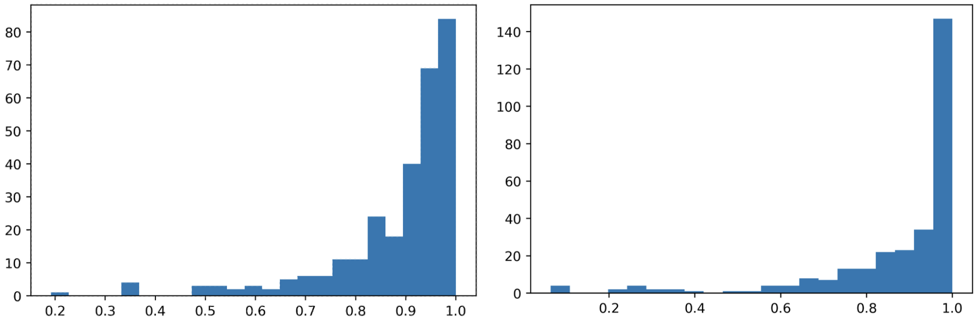


Figure 37: Average p-values from the K-S test for all stations for departures (left) and arrivals (right). The K-S test cannot reject the hypothesis that the observed data follow a Poisson distribution.

test potentially fails to reject the null hypothesis due to reduced statistical power. However, it is a very reasonable assumption that the excess demand/supply will also be following the same distribution (albeit with different parameters) as the observed demand/supply.

Estimating excess demand of bikes in Chicago's Divvy: Following the aforementioned approach of excess demand estimation, we can calculate the excess demand observed on the system. While the bike availability curves are just like the ones we simulated, there is one important difference. The excess demand rate in the real environment is not constant over time but it rather changes. For example, we expect the excess demand rate in the morning (rush hour) is higher than that in the late night. There are several factors that can lead to this temporal variation, ranging from people's schedule (e.g., during rush hours the excess demand is expected to be higher) to weather conditions that change during the day. This temporal dependency does not allow us to use all τ_f intervals in the data to estimate a single, constant, excess demand. We will need to only use limited information, localized in time, to estimate the excess demand rate during a specific time interval.

In particular, we adjust the aforementioned approach in this section as follows. Here we still use Fig 35 to describe the adjusted approach. The EDP in the interval (t_1, t_2) is able to inform us about the excess demand occurring in the immediately preceding interval (t_d, t_1). We can use Eq (20) to calculate excess demand rate in this interval. However,

$\tau_f = t_2 - t_1$ is the only EDP length that we can use to calculate τ_m given the time-varying nature. Furthermore, we need to calculate the average inter-supply interval τ_s, which again needs to be temporally localized due to its time varying nature. For the setting in Fig 35 we have arrival events at t_a, t_b and t_1. Thus, we use inter-arrival intervals, i.e., (t_a, t_b) and (t_b, t_1), to obtain $\tau_s = \frac{(t_b - t_a) + (t_1 - t_b)}{2}$. Finally, we calculate $\hat{\mu}_e$ of interval (t_d, t_1) using Eq (20).

The single EDP length aforementioned may cause the calculated excess demand rate to be extreme. For instance, if the bike was rented almost immediately after it was returned, then the excess demand rate would be calculated practically as infinite. While we could eliminate such observations - since most probably correspond to users that return the bike and re-rent it immediately just for time-limit purposes imposed by the operator - it is not clear what is the time threshold as a good standard to eliminate such observations (i.e., such extremely short EDP lengths). To avoid having to choose an arbitrary cutoff, we make use of the Bayesian average [17]. The Bayesian average is a weighted average between (i) the estimate obtained from the sample we have for the quantity of interest, and (ii) a prior belief for this estimate. The weights are the sizes of the samples respectively (for the prior it can be a sample size that is considered *stable*). As with any Bayesian analysis, the prior can be purely subjective, or uninformative etc., but it can also be calculated by data. In our case, we can focus on a period of time around the time interval of interest and estimate the excess demand for the same periods over a week. If our measurement of the interval of interest was an extreme outlier, then the prior will shrink the final estimate. For example, let us assume that we want to calculate the excess demand rate at 9:30-10:00am on a given day, which is referred to as μ_{930}. First, using Eq (20) we calculate the excess demand rates of 9:30-10:00am (interval of interest), and 9:00-9:30am, 10:00-10:30am (periods near interval of interest) of the given day. This will give us 3 observations and an observed average μ_{obs}. Then using Eq (20) we calculate the excess demand rates of 9:00-9:30am, 9:30-10:00am, and 10:00-10:30am every day since 6 days before the given day. This will essentially give us 18 observations and an estimated prior average μ_{prior}. Combining these with the Bayesian average we will get our final estimate for μ_{930} as:

$$\hat{\mu}_{930} = \frac{3 \cdot \mu_{obs} + 18 \cdot \mu_{prior}}{21} \tag{21}$$

Of course, the choice of prior can be different, but the idea is that using this approach we can smooth extreme cases in a principled way. In the Appendix E, we further discuss how we processed instances that do not follow exactly the shape of EDP discussed here but appear infrequently in the data (e.g., when multiple bikes simultaneously arrive at a station as a result of rebalancing from the operator).

Estimating excess demand of docks in Divvy: Chicago bike sharing system does not allow for self-docking [24, 23] . Thus, if a bike is returned and the dock is full, there is no way to return it, leading to excess demand for the dock. To calculate the excess demand of docks, we can still use the method used to estimate the excess demand for bikes, but we need to make the following adjustments:

- The availability curve now represents dock availability (i.e., how many racks at the station are free), rather than bike availability (i.e., how many bikes are available at the station for renting).
- 0 dock availability means that each rack at the station is occupied by a bike.
- The EDP starts with a bike departure (from a full station) and quickly ends with a bike arrival. This allows us to capture how *quickly* the rack is being utilized again, thus, capturing, the excess demand for docks (which again is time-varying).
- τ_f still denotes EDP length (based on the definitions above), while τ_m still denotes average value of τ_f.
- τ_s still denotes the average value of inter-supply intervals, but to reiterate, based on the definitions above, in this case a supply is a bike departure. So specifically, τ_s means the average value of inter-departure intervals.
- We use λ_e to denote excess demand rate of docks, which is formally defined in Eq (22):

$$\hat{\lambda}_e = \max(\frac{1}{\tau_m} - \frac{1}{\tau_s}, 0) \tag{22}$$

Excess demand in different stations: As one might expect, the excess demand rates differ among different stations. The maps in Figs 38 and 39 illustrate the sum of the excess demand for each station for bikes and docks respectively. As we can see, stations closer to the downtown area have higher excess demand rate. We further illustrate in the inset

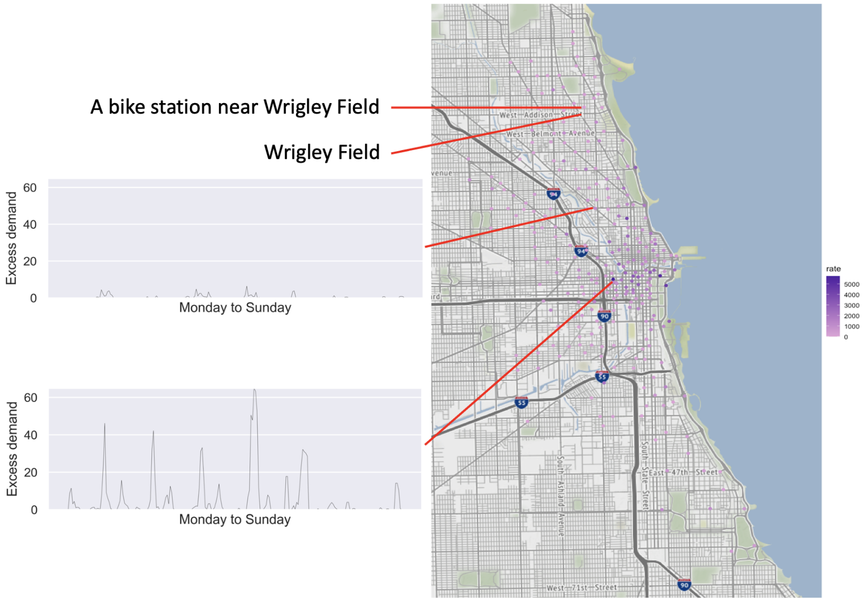


Figure 38: Cumulative bike excess demand rate for different stations. (Reprinted from [57] under a CC BY license, with permission from OpenStreetMap, original copyright 2021.)

figures the weekly patterns of the excess demand in 30-minute periods for two representative stations. As we can see these stations exhibit very different patterns in terms of levels of excess demands (both for bikes and docks). However, the relative spikes in each station appear to be similar to an extent. Furthermore, when focusing on a specific station, there seems to be a temporal shift between the excess demand for bikes and docks.

Excess demand and sporting events: In order to provide some context for the excess demand observed at the system, we examined the estimated excess demand near the Wrigley Field during game days. For example, at 1:20pm on July 8, 2018, there was a baseball game in Wrigley Field between the Cubs and the Reds [27]. There is a Divvy station only 130 meters away from Wrigley Field, which we have also marked in Fig 38. Based on our calculations, this station exhibited excess demand during particular time periods on that day. In particular, between 12:30pm and 2pm there was an average excess dock demand of more than 3 docks/30 minutes. This is possibly due to fans riding bikes to Wrigley Field, leading to non-empty docks. Furthermore, between 4pm and 5pm there was an average bike

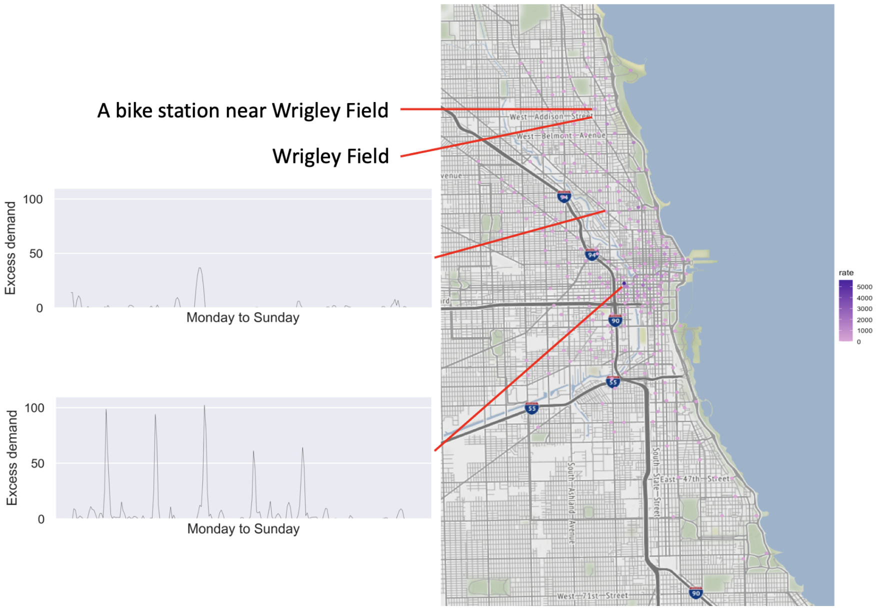


Figure 39: Cumulative dock excess demand rate for different stations. (Reprinted from [57] under a CC BY license, with permission from OpenStreetMap, original copyright 2021.)

excess demand of 2.36 bikes/30 minutes, which is possibly due to several fans making their way out of the stadium as the game was coming to an end.

4.3 Demand prediction models

The data processing described until now can facilitate a *post-hoc*, descriptive, analysis of the historical excess demand rates in a shared bike system. However, it is also important to explore the ability to perform predictions for the excess demand conditioned on various external variables. This can facilitate logistics operations, such as, rebalancing, fleet updates, etc. We define the following:

- Total bike demand volume N_μ: Number of rented bikes in a 30-minute time interval of interest. This includes both bikes actually rented and bikes attempted to be rented but there was no availability.

- Total dock demand volume N_λ: Number of returned bikes in a 30-minute time interval of interest. Again this includes both bikes actually returned, as well as, bikes attempted to be returned to a full station.
- Net total demand volume $Z = N_\mu - N_\lambda$: Difference between total bike demand volume and total dock demand volume in the same 30-minute time interval of interest.

In this section, we develop a predictive model for the net total demand volume at a station during a 30-minutes interval; i.e., build a predictive model for Z during a specific time interval. We choose Z as our dependent variable since it provides direct insights for the bike operator to decide the number of bikes to be rebalanced. Therefore, we need to estimate the bike and dock demand volumes during each 30-minute period in our data. However, it is important to note that these total demand volumes, include both the observed from the trip logs demand, as well as the excess demand that is not directly captured in these data. In particular, we perform the following steps for each 30-minute interval in our data:

- **Calculate observed demand volumes**: We obtain the number of observed departures, which is equal to the observed bike demand volume N_{μ_o} during the interval of interest, as well as, the number of observed arrivals, which is equal to the observed dock demand volume N_{λ_o} for the same interval.
- **Calculate the excess demand rate**: As per the discussion in the previous section, we also identify EDPs from bike and dock availability to calculate bike and dock excess demand rates μ_e and λ_e respectively.
- **Convert rate to volume**: If a time duration with the existence of excess demand (i.e., a duration with 0 availability) is located inside our 30-minute interval of interest, we denote the length of that duration for bike, dock excess demand as l_{μ_e}, l_{λ_e}, respectively. Then, we convert bike and dock excess demand rate to bike (N_{μ_e}) and dock (N_{λ_e}) excess demand volume by multiplying with l_{μ_e}, l_{λ_e}:

$$\begin{aligned} N_{\mu_e} &= \mu_e \times l_{\mu_e} \\ N_{\lambda_e} &= \lambda_e \times l_{\lambda_e} \end{aligned} \tag{23}$$

Using the above, we finally calculate N_μ, N_λ, Z as:

$$\begin{aligned} N_\mu &= N_{\mu_o} + N_{\mu_e} \\ N_\lambda &= N_{\lambda_o} + N_{\lambda_e} \\ Z &= N_\mu - N_\lambda \end{aligned} \tag{24}$$

Following the above process, we are able to obtain the net total demand volumes in the Divvy system for each 30-minute interval during the 2018 year.

To build our prediction model for the net total demand volume, we consider a set of variables that are expected to be correlated with the demand for bikes and docks. More specifically, we use the independent variables listed in Table 13.

Each data record used to build our model describes a 30-minute interval of observations. Given that the weather data are only available on the top of the hour, we interpolate them for the half hour interval. Having identified the covariates to use in our model, we start by exploring two generalized linear models, namely, Poisson regression and Skellam regression. With the first approach, we model the total demand volumes for the bike and dock demand independently, while with the second approach we model directly their difference, i.e., the net total demand Z. We also explore and evaluate the predictive performance of a feed forward neural network and XGBoost on the same set of features.

4.3.1 Poisson regression

To estimate Z, an intuitive approach would be to predict the total bike departures N_μ and bike arrivals N_λ, and then calculate $Z = N_\mu - N_\lambda$. Bike departures and arrivals have been widely modeled as Poisson flows [30, 63, 35, 65, 18, 12, 33], so a Poisson regression is an intuitive candidate model. A Poisson regression essentially models the expected value of the dependent variable through a linear combination of a set of independent variables $\mathbf{X}$ as:

$$\lambda_Y = e^{\alpha + (\mathbf{b} \cdot \mathbf{X})} \tag{25}$$

The parameters α and $\mathbf{b}$ are obtained through maximum likelihood estimation. We can also estimate the distribution for the dependent variable Y as:

$$p(Y = k|\mathbf{X}, \mathbf{b}, \alpha) = \frac{e^{k \cdot (\alpha + (\mathbf{b} \cdot \mathbf{X}))}}{k!} \cdot e^{-e^{\alpha + (\mathbf{b} \cdot \mathbf{X})}} \tag{26}$$

In our case, we have two processes that we need to model, namely the bike demand and the dock demand. Therefore, we learn two separate regression models using the covariates described above. For the rest of the paper, we will refer to this model as the "Two-Poisson regression" model.

4.3.2 Skellam regression

The Two-Poisson regression model assumes that the two processes - rentals and returns - are independent and hence, we can model them separately. However, this is not necessarily the case (The correlation between total bike demand volume N_μ and total dock demand volume N_λ of a station can be up to 0.885), and in these situations the estimations will be biased [43, 59]. However, we can directly model variable Z through a Skellam distribution since it represents the difference between two Poisson distributions [68]. In fact, if $(X, Y) \sim BP(\lambda_1, \lambda_2, \lambda_3)$, where λ_3 captures the covariance between X and Y, then their difference $Z = X - Y$ follows the Skellam distribution:

$$P(z) = e^{-(\lambda_1 + \lambda_2)} \cdot \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} \cdot I_z(2\sqrt{\lambda_1 \lambda_2}) \tag{27}$$

where $I_z(x)$ is the modified Bessel function. What we can observe is that the distribution does not depend on the covariance (λ_3) of the two Poisson distributions [68].

Therefore we can model the net total demand Z through a Skellam regression. In particular:

$$\begin{aligned} Z &\sim Skellam(N_\mu, N_\lambda) \\ \ln(N_\mu) &= \mathbf{b}_1 \cdot \mathbf{X} \\ \ln(N_\lambda) &= \mathbf{b}_2 \cdot \mathbf{X} \end{aligned} \tag{28}$$

where $\mathbf{X}$ denotes independent variables. $\mathbf{b}_1$ and $\mathbf{b}_2$ denote the coefficients to be learnt. We fit the model using Maximum Likelihood Estimation. Implementation source code can be found at `https://github.com/xinliupitt/skellam_regression`.

Table 13: Independent variable list. The first three variables are numerical, and the remaining are categorical.

Name	**Description**
temperature	temperature (unit: Kelvins)
cloud percentage	percentage of clouds in the sky
wind speed	wind speed (unit: meter/sec)
day of a week	day index of a week: Mon - Sun
interval index	30-minute interval index of a day (e.g., 6:00 - 6:30, 6:30 - 7:00)
holiday indicator	binary indicator of whether the record falls in weekend or federal holidays (1) or not (0)
cloud indicator	binary indicator of weather being “cloud” (1) or not (0)
rain indicator	binary indicator of weather being “rain” (1) or not (0)
mist indicator	binary indicator of weather being “mist” (1) or not (0)
snow indicator	binary indicator of weather being “snow” (1) or not (0)
thunderstorm indicator	binary indicator of weather “thunderstorm” (1) or not (0)

4.4 Results

In this section we will present our evaluation results for predicting the net total demand. We will evaluate the predictive performance across two dimensions:

- Peak - vs - non-peak hour predictions
- Training based on observed - vs - total demand

Specifically, for the latter, we are interested in quantifying the predictive gains achieved by considering the excess bike and dock demand, and not only using recorded bike rentals and returns.

4.4.1 Peak and non-peak hours

Typically “peak-hours” for a transportation system include weekdays morning (7am-9:30am) and evening commute (4pm-6:30pm). However, for a bike sharing system there is also seasonality, especially during the summer months [81]. Our data also support this seasonality. In particular, the net total demand during peak hours in the summer months is approximately 6 times higher as compared to that during non-peak hours of the year. For this reason, our results for peak hours below will be focused on the summer months. Different peak hours also have different patterns across seasons. Given the imbalance between the records for the peak hours per season and non-peak hours (peak hours per season cover a little less than 15% of the observations), a single model would be *overwhelmed* by the latter and will not able to identify the peak hour patterns in different seasons. Hence, we build separate models for different time periods. In particular, we learn a single model for non-peak hours, while we build two separate peak hour models (one for the morning and one for the evening peak hours). Predicting the net total demand for (particularly) the peak-hour periods is very important for the bike share system operator for various management operations, such as conduct an effective rebalancing. For learning each model, we split the data from all 300 stations and use 80% of the them to train the model, 10% as the validation set to optimize the regularization shrinking parameter, and the remaining 10% for out-of-sample evaluation. All models use L1 regularization, while we use the mean squared error

Table 14: MSE of different time periods under various ML models.

Model	Excess (7-9:30)	All records (7-9:30)	Excess (16-18:30)	All records (16-18:30)	Excess (non-peak)	All records (non-peak)
Skellam	**36.2**	**6.4**	**36.4**	**10.3**	42.6	**2.7**
Two-Poisson	37.6	6.7	37.2	10.6	45.3	2.8
Neural	40.1	6.8	39.6	10.8	43.1	2.8
XGBoost	36.3	9.3	43.1	16.2	**40.2**	3.4
Constant	44.6	8.8	68.2	16.7	67.0	3.1

(MSE) as our loss. The Skellam model training process follows the regression training setup in Appendix F.

Baseline models: We compare our proposed modeling (Skellam regression) with the following four baselines: (i) two independent Poisson models (Section "4.3.1"), (ii) a feed forward neural network, (iii) XGBoost, (iv) constant prediction. They are referred to as "Two-Poisson", "Neural", "XGBoost", "Constant", respectively in Table 14. For the models except constant prediction, we apply L1 regularization and use the validation set to optimize the shrinking parameter. In particular, the "Two-Poisson" model follows the regression training setup in Appendix F.. For the neural network we use 5 hidden layers, 32 units per layer and a batch size of 32 for training. For XGBoost, we set the number of estimators to 10,000. For the constant prediction, we use the average net total demand existing in the training set as our prediction for each out-of-sample record.

Table 14 presents the MSE on the test set for two peak hours periods (in the columns marked with "7-9:30", "16-18:30") as well as the non-peak hours (in the columns marked with "non-peak"). For each period, we present the MSE over all the records in the test set (in Table 14 columns marked with "All Records"). In the test set, there are some records with non-zero excess demand; that is, when calculating the ground truth Z of those records, either N_{μ_e} or N_{λ_e} is non-zero. To understand better any gains existing in predictions, we also specifically present the MSE of those records (in Table 14 this corresponds to the columns marked with "Excess"). As aforementioned these instances are very important for the bike sharing system operator, since these are the situations where operations such as rebalancing are crucial. Note that the records with non-zero excess demand occupies 10% of the dataset for peak hours, and 2% for non-peak hours. For two peak-hour periods, as we can observe, the Skellam regression exhibits the lowest error among all the models examined. The benefits are even larger, in situations where the excess demand is non-zero. For non-peak hours, as we can see, Skellam exhibits only slight benefits over the two-Poisson model and the neural network. This could be attributed to the fact that during non-peak hours, there is an overall low demand for the bike sharing system, and hence, the two-Poisson and neural network models can capture this signal. Finally, XGBoost seems to perform slightly better than Skellam regression for records with non-zero excess demand. However, these records only occupy 2% of the dataset for non-peak hours (these could represent situations where there are special events - e.g., summer street fairs - that boost demand during non-peak hours).

Apart from its performance in terms of MSE, Skellam regression has two additional advantages over the alternative models considered. First, the Skellam regression as a generalized linear model is interpretable. This is particularly important from an operator's perspective, since it can lead to actionable insights. For example, in a model built for a station during non-peak hours, for the independent variable "temperature" we obtain two coefficients: $\mathbf{b}_{1,temp} = 6.57$ for bike demand and $\mathbf{b}_{2,temp} = 6.90$ for dock demand (Eq (28)) . These coefficients indicate that higher temperature is correlated with more people renting bikes for biking, i.e., higher bike demand. Since these riders need to return the bikes, the dock demand is also positively correlated with the temperature. Secondly, and most importantly, the Skellam regression model allows us to get a better estimation of the uncertainty of our prediction. In particular, we do not only get a single point estimate for the expected value of the net total demand, but rather its whole probability distribution. For example, let us assume that our predictions are $\hat{N}_\mu = 12.67$ and $\hat{N}_\lambda = 10.29$. This means that the net total demand is $\hat{Z} = 2.38$. Recall, that $\hat{N}_\mu$ and $\hat{N}_\lambda$ are the two parameters of the Skellam distribution, and hence, we can plot the probability mass function for Z as presented in Fig 40. This distribution allows us to answer questions, such as *"what is the probability that there will be excess demand during a specific time period?"*. Questions are important for the

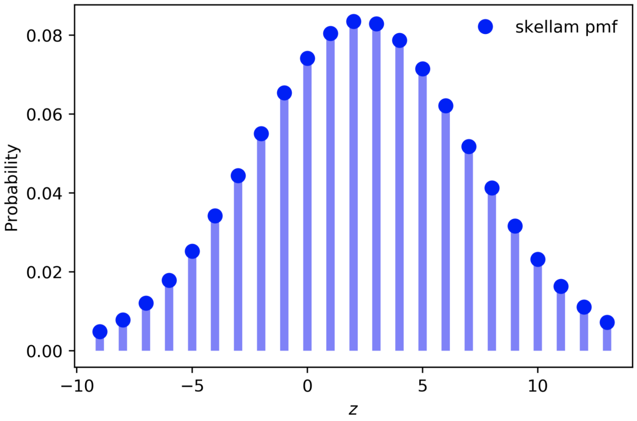


Figure 40: Skellam probability distribution with parameters $\hat{N}_\mu = 12.67$, $\hat{N}_\lambda = 10.29$. $\hat{Z} = 2.38$.

system operators, providing them with a more holistic view of the system.

4.4.2 Total and observed demand in training

For the results we presented above, we use the total demand N_μ and N_λ to calculate the dependent variable $Z = N_\mu - N_\lambda$. One of the motivations for our study is the fact that excess demand is not directly available in the trip/dock availability logs obtained from the bike system operator. Therefore, a lot of existing literature simply uses the observed demand for building predictive models. For these models, 0 trips from a station during a period is an indicator of 0 demand, even though as we have seen this may very well be an instance of actually high (excess) demand. However, what if even by simply using the observed demand to train our models, we can still get a good prediction for the net total demand. To examine this we build our model using only the observed demand when we train the model. We then evaluate the predictions on the test set and the results are presented in Table 15.

As we can see, when training our models using the total demand ("Observed+Excess" in Table 15), the predictions have obvious performance gain (as expected). These gains are of course higher when making predictions for periods with excess demand, as one might have expected as well.

Table 15: MSE of different time periods under Skellam model.

Model	Excess (7-9:30)	All records (7-9:30)	Excess (16-18:30)	All records (16-18:30)	Excess (non-peak)	All records (non-peak)
Observed +Excess	36.2	6.4	36.4	10.3	42.6	2.7
Observed	47.5	10.0	52.2	11.9	45.8	2.9

4.5 Discussion and Conclusions

In this chapter, firstly, we identify that the proxy to estimate the capacity demand of the bike-sharing system is the observed demand, which is directly recorded in bike trip logs. Then we propose our approach/proxy to estimate "excess demand" in bike sharing systems (e.g., how many customers attempted to rent a bike from an empty station). This type of demand is not directly recorded in bike trip logs. Key to our approach/proxy for estimating excess demand is identifying temporal segments in the bike availability data, that include changes in the availability from zero to non-zero. Through simulations, we verify the ability of our approach to estimate the excess demand present in the system. Consequently we apply our approach on data obtained from Chicago's Divvy bike sharing system to estimate the excess demand present in Divvy system. To predict the net total demand (which includes the observed and excess demand), we learn a Skellam regression model through maximum likelihood estimation, which shows advantages over other alternative models, both in terms of predictive performance and interpretability. Moreover, our Skellam regression model, as a generalized linear model, allows us to get a better estimation of the uncertainty of our prediction, since we essentially obtain the whole probability distribution of our dependent variable.

5.0 Limitations and Future works

In this chapter, we will describe limitations of our works on several aspects. In each aspect, we will discuss corresponding future directions.

5.1 Proxy for excess demand

We estimate the excess demand based on the proxy we select. This means, our estimated excess demand is limited to the patterns under the selected proxy. Particularly, in urban business, our selected proxy is the complementarity, which may not well capture the patterns in time dimension. This may cause our estimated excess demand lose some time-dimensional information. In urban transportation, our selected proxy is EDP, which only relies on the data from bike stations. It may lose some information on the customer's side. Such information can be important to improve our excess demand estimation.

A future direction is to add other proxies for excess demand. In urban business, we can add the proxy - data of customer waiting and spent time at a specific venue. This may help us capture the time-dimensional information for excess demand. For urban transportation, we may add the proxy - how customer operates the bike system's mobile app. For example, a customer of the bike sharing system may use the corresponding mobile application to explore the bike availability of stations near her location. This search itself is a signal of bike demand, and in the case where there are no available bikes nearby we can consider this to be part of the excess demand, which adds necessary information to improve our excess demand estimation.

5.2 Proxy of capacity demand

In urban business, our selected proxy of capacity demand is the simulation which relies on 2 factors: distance and rating. Our limitation is that in the simulation, we assume people prioritize the distance choice over the rating.

A future direction is to slightly change the proxy of capacity demand based on the trade-off between distance and rating. For example, we may assume people prioritize the rating choice over the distance. We may also assume people only care about distance but not rating. Under different assumptions, we estimate the capacity demand and then excess demand. In this way, we can study the patterns of excess demand under slightly different proxies of capacity demand. Such pattern differences may help us further improve the excess demand estimation.

5.3 Data time range

For urban business, since the FourSquare movement data is available only from April 2017 to March 2019, our estimated excess demand is limited to the context of this time range.

A future direction is to study the excess demand in different year ranges. For example, if the movement data can be available during COVID-19 pandemic, we can estimate the excess demand during pandemic, and compare the excess demand patterns before and after pandemic. We may also see how the process of pandemic gradually changes the patterns of excess demand.

5.4 Types of bike systems

When estimating the excess demand in bike sharing system, we assume this is a docked system, i.e., bikes should be returned to designed stations. This limits our excess demand estimation to the docked system.

A future direction is to extend our approach to a dockless system, where a customer can return a bike anywhere in a city. In such setting, before applying our approach, we may need to divide the whole city into lots of predefined areas, each of which imitates the “station” in the docked system. However, it is challenging to elaborate this dividing process and the granularity of the predefined areas, which should be carefully and further studied.

6.0 Conclusions

In this paper, we firstly find that complementarity can be a proxy for excess demand by examining the fast food restaurants near highway exits. We also find that the demand distribution is influenced by distances and venue ratings, which are important factors to investigate when we move on to explore demand patterns in more general urban areas.

Next, by choosing the complementarity as the proxy, we estimate the excess demand for the urban business. Particularly, we propose our approach, which is incorporates real-world and simulated data, to estimate the complementarity for urban business entities. For each urban business venue, we estimate every source of its complementarity. For urban areas, we reveal the complementarity patterns among different periods in a day, and find that the complementarity can be explained by the venue diversity, venue density, number of venues and inter-venue distance of urban areas. We fetch the embeddings of urban areas via a graph neural network and reveal the inter-area relationship in the latent space. Using these results, venue owners can improve their business strategy to satisfy more excess demand and increase their revenue.

For urban transportation, we estimate and predict the excess demand in bike sharing systems. The proxy of excess demand is a temporal segment in the bike availability data, that include changes in the availability from zero to non-zero. Assisted by this proxy, we propose our approach to estimate the excess demand based on queuing theory. We verify through simulations its ability to estimate the excess demand present in the system. Then we learn a Skellam regression model to predict the net total demand, which shows advantages over other alternative models, both in terms of predictive performance, as well as, interpretability. Using these results, the bike sharing operator can strategically rebalance bikes to satisfy more excess demand, which provides convenience to citizens and improves the city's transportation condition.

Appendix A Could ArcGIS work for the complementarity topic?

For the topic of complementarity, there are two major tasks to accomplish: (1) simulate citizen movements, based on which the complementarity will be calculated; (2) analyze the pattern of complementarity. ArcGIS, as a well-known tool in GIS study, has the toolboxes for simulations and GIS information analysis. We are interested in whether these two tasks can be completed by ArgGIS.

ArcGIS does not have a straightforward function to calculate the complementarity. From the book [36], in Chapter 7 for example, ArcGIS seems to have this tool to visualize the interactions among venues. However, such interaction is not related to the concept of complementarity, which should be based on venue visitations. More specifically for that tool,

- It only discusses the distance patterns among venues in a specific urban area.
- The analysis by that tool has too many assumptions, such as multiple candidate locations to open restaurants, a distance threshold for an interaction to exist. However, the complementarity analysis does not have such assumptions.

In fact, any analytical tool in ArcGIS [40, 66] does not analyze visitations; in other words, they cannot be used to obtain the complementarity in our paper.

One may think we could directly ArcGIS's simulation tool to complete the simulation. Then we use the simulation results to calculate the complementarity via our own codes. However, ArcGIS's simulation tool does not meet our requirements. Firstly, it cannot use external venue data (e.g., our Foursquare dataset) as the simulation context. Also, it does not support customizing details of agents' motivations. For example, agents will select candidate venues within a "width". Agents also tend to visits highly rated venues. However, ArcGIS does not allow us to customize the width and rating-based probabilities.

Based on these attempts and findings, ArcGIS is not a suitable tool for the research topic on complementarity.

Appendix B Simulation implementation based on the geohash

Recall that in the first stage of our simulation idea, for the agent's each start venue v_1 (each movement), we need to select a set of candidate end venues within the ring area $d \pm \Delta r$. This means, we need to brute all venues in the dataset to see if they are within ring area $d \pm \Delta r$. Assume there are a total of N_v venues in the dataset. As we need to simulate $(N_{B \to B} + N_{NB \to B})$ movements, the computational complexity is $N_v(N_{B \to B} + N_{NB \to B})$, which is very high.

To reduce the computational complexity of simulations, we apply the concept of geohash and its related tools in our practical simulation approach. A geohash is a collection of venues which are geographically close. In this way, for an agent's each start venue v_1 in the simulated movement, we only need to brute force the distances from v_1 to a total of N_{geo} geohashes and then identify geohashes which are located in our ring area $d \pm \Delta r$. In fact, the total number of geohashes N_{geo} is smaller than the total number of venues N_v (shown in Table 16), which is the key reason to reduce the computational complexity to $N_{geo}(N_{B \to B} + N_{NB \to B})$. Additionally, the `proximitypyhash` API[1] is applied in our simulation, which further accelerates the computational process related to the geohash.

[1] https://pypi.org/project/proximitypyhash/

Table 16: Number of N_v and N_{geo} for each city.

	New York City	Los Angeles	Chicago
N_v	10884	20372	8681
N_{geo}	6212	6140	4795

Appendix C Weight of independent variables to explain complementarity

Our objective is to find which candidate value of w_{ends} outputs the best linear model for complementarity (i.e., dependent variable) based on formula Eq. (16). Our candidate values are 0.60, 0.65, ..., 0.90, 0.95, where all candidate values being larger than 0.5. The reason is that complementarity means citizens are motivated to go from the start area to the end area, which indicates the features of the end area is very attractive. Since w_{ends} is the weight of features of the end area, setting $w_{ends} > 0.5$ can let the weight of the end area always larger than the start area.

The steps of selecting w_{end} are elaborated as follows:

I. Choose one candidate value of w_{ends}, such as 0.60. Calculate all independent variables to generate a dataframe given the currently chosen candidate w_{end}.

1. Randomly split the whole dataframe into 80% training set and 20% validation set.
   a. Use the training set to build a linear model based on Eq. (16).
   b. Input the validation set to the built linear model to generate the predicted complementarity. Compare the predictions and ground truth of complementarity, which outputs an mean-square-error (MSE) value.
2. Repeat Step 1 and its nested steps 100 times, which generates 100 MSE values. Record the average of these MSE values.

II: Repeat Step **I** (by traversing all candidate w_{end} values) and its nested steps. Then we obtain a specific average MSE value under each candidate w_{end} value.

III: Select the w_{end} value corresponding to the lowest average MSE value.

By executing the above steps, we finally select w_{end}=0.65.

Appendix D Dimensionality selection in neural network

For the graph neural network, we carefully choose the dimension of the embedding, i.e., the number of elements in an embedding vector. We tried 32 and 128 as the embedding dimension, and finally decided to use 32 for the following reasons.

The first reason is that lower dimension can make the distance calculation between embeddings/vectors become more robust. Under a high dimension, the largest and smallest distances have so little difference [8, 4], which makes all distances become relatively meaningless.

The second reason is that under 32 dimensional embeddings, the model converges relatively better. The training loss of 32 and 128 dimensional embeddings are shown in Fig. 41, 42, respectively. As observed, the loss of Fig. 41 has less fluctuation, which indicates a better convergence of 32 dimensional embedding than that of 128.

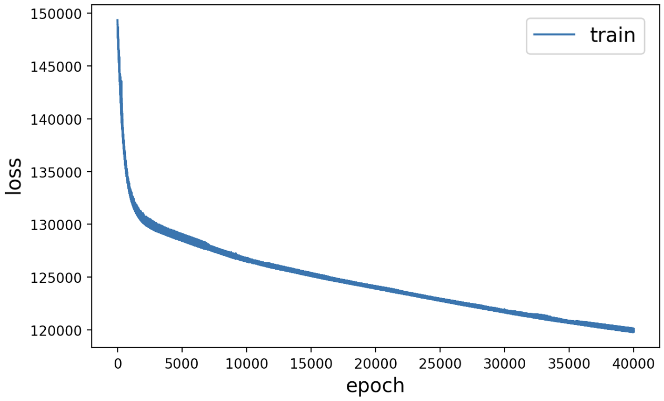


Figure 41: Loss of Chicago midday with embedding dimentionality 32.

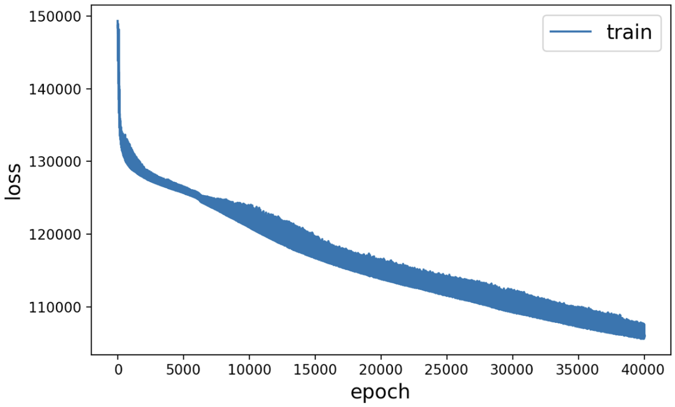


Figure 42: Loss of Chicago midday with embedding dimentionality 128.

Appendix E Special cases of excess demand in bike-sharing

While the excess demand can be estimated as described in the main text for the vast majority of the instances, rebalancing from the operator can break down the calculations and a slightly different approach is needed.

For example, let us consider the situation in Fig 43. As we can see in this case, the availability at t_0 changes from 0 to k ($k \leq 4$ as observed in the real data of Divvy bike sharing system), potentially due to the bike sharing operator relocating/rebalancing k bikes to this station (of course, other reason are possible, such as, a group trip, but the treatment of the situation is the same regardless of the reason for causing it). Next the k bikes are consecutively consumed at times t_1, t_2, $...t_k$ respectively. This means that the demand rate is high, which leads to the supply of all the k bikes being consumed quickly before any new supply of bikes arrives. The EDP in this case is the curve between $[t_0, t_k]$, which to reiterate it is different from the *typical* EDP discussed in the main text. As we discussed in Section "Excess demand estimation", in this case the excess demand rate μ_e equals to the departure rate μ. The estimated departure rate $\hat{\mu}$ can be calculated by inverting the average of intervals between rentals; i.e., $[t_0,\, t_1]$, $[t_1,\, t_2]$, $[t_2,\, t_3]$, $\ldots$, $[t_{k-1},\, t_k]$. Then the average value of these intervals is $\dfrac{t_k - t_0}{k}$. Finally, the excess demand rate is estimated by inverting $\dfrac{t_k - t_0}{k}$, i.e., $\hat{\mu}_e = \dfrac{k}{t_k - t_0}$.

The scenario shown by the bike availability curve in Fig 44 is a generalized case of Fig 43. The difference is that starting at t_0, the bikes are consecutively rented (at t_1, t_2, $...t_e$ respectively) up to the point when a supply arrives at t_g, where e is the total number of consumed bikes before t_g. In Fig 44, the EDP is the curve during $[t_0, t_e]$, which is terminated by the supply arrival at t_g. The reason follows our explanation for Fig 3 in Section "Excess demand estimation". In particular, the supply arrival at t_g indicates that the bikes are not consumed *quickly* anymore and the ensuing rental would be recorded at the rental (observed) logs. Therefore, we calculate the excess demand using the departure records during $[t_0, t_e]$, when bikes are consumed *quickly*. Following the calculation method aforementioned the

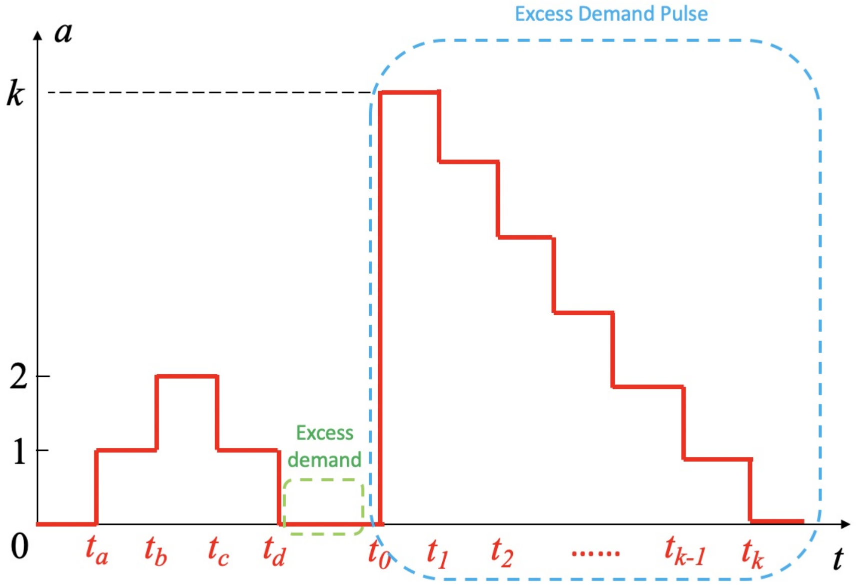


Figure 43: A segment of bike availability curve with a bulk of (potentially rebalanced bikes) arriving at t_0, all consumed by consecutive rentals.

excess demand in this case is $\dfrac{e}{t_e - t_0}$.

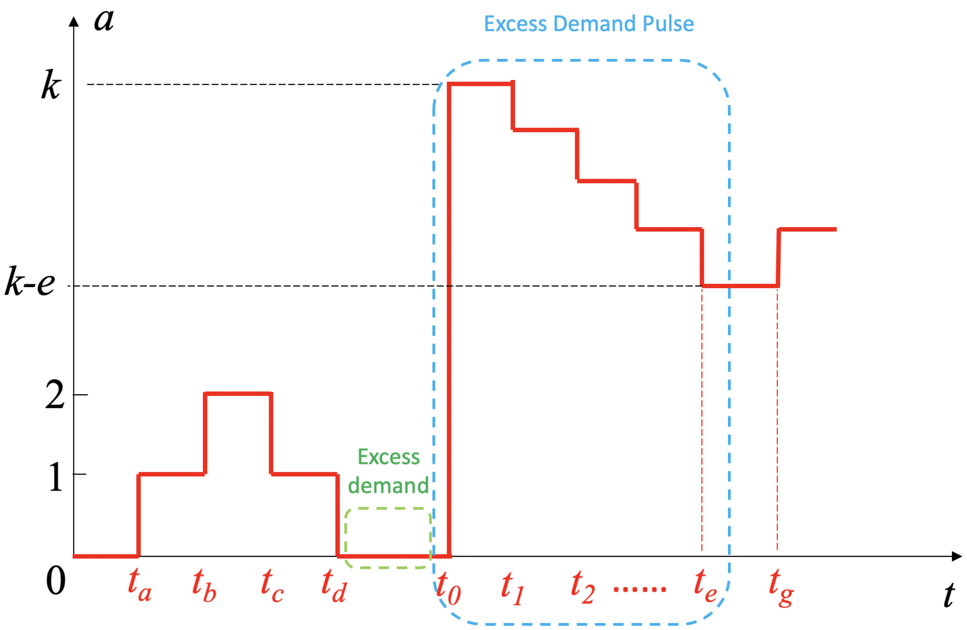


Figure 44: A segment of bike availability curve to describe the generalized case of excess demand with a bulk of bikes arriving at the dock at time t_0.

Appendix F Regression training setup

For training the Skellam and the Two-Poisson regression models, we need the dependent variables to be integers, since the probability distribution for these models is discrete. However, from Eq (20)(22)(23), the values of our estimated excess demand volumes N_{μ_e}, N_{λ_e} can be non-integers, since they are obtained through the estimation of the excess demand rate. In order to be able to train the models we use sampling. In particular, the estimated excess demand volumes are the expected values of the Poisson processes for departures and arrivals. Therefore, we can sample two Poisson distributions with intensity N_{μ_e}, N_{λ_e} respectively, and obtain specific integer instances to update/replace N_{μ_e} and N_{λ_e}. This sampling process also incorporates some of the uncertainty around the excess demand volumes, captured by the whole probability distribution. Finally, we obtain integer instances for N_{μ} and N_{λ} as dependent variables of the Two-Poisson model, Z as dependent variable of the Skellam model.

Appendix G Publication list

Papers contributing to this dissertation:

- Liu, Xin, and Konstantinos Pelechrinis. “Complementarity estimation for urban business.” In preparation.
- Liu, Xin, and Konstantinos Pelechrinis. “Excess demand prediction for bike sharing systems.” Plos one 16.6 (2021): e0252894.
- Liu, Xin, Konstantinos Pelechrinis, and Alexandros Labrinidis. “hood2vec: Identifying similar urban areas using mobility networks.” Future Cities Challenge Session in Netmob 2019.
- Liu, Xin, and Konstantinos Pelechrinis. “A Data-Driven Examination of Hotelling’s Linear City Model.” Proceedings of the 10th ACM Conference on Web Science. 2019.

Other papers during my PhD study:

- Liu, Xin, Mai Abdelhakim, Prashant Krishnamurthy, and David Tipper. ”Identifying malicious nodes in multihop IoT networks using dual link technologies and unsupervised learning.” Open Journal of Internet Of Things (OJIOT) 4, no. 1 (2018): 109-125.
- Liu, Xin, Mai Abdelhakim, Prashant Krishnamurthy, and David Tipper. ”Identifying malicious nodes in multihop IoT networks using diversity and unsupervised learning.” In 2018 IEEE International Conference on Communications (ICC), pp. 1-6. IEEE, 2018.
- Abdelhakim, Mai, Xin Liu, and Prashant Krishnamurthy. “Diversity for detecting routing attacks in multihop networks.” 2018 International conference on computing, networking and communications (ICNC). IEEE, 2018.

Bibliography

[1] Robert Aboolian, Oded Berman, and Dmitry Krass. Competitive facility location and design problem. *European Journal of Operational Research*, 182(1):40–62, 2007.

[2] Robert Aboolian, Oded Berman, and Dmitry Krass. Competitive facility location model with concave demand. *European Journal of Operational Research*, 181(2):598–619, 2007.

[3] Robert Aboolian, Oded Berman, and Dmitry Krass. Efficient solution approaches for a discrete multi-facility competitive interaction model. *Annals of Operations Research*, 167(1):297–306, 2009.

[4] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[5] Burcu Aydin, Kemal Guler, Enis Kayis, and Mehmet O Sayal. Estimation of unobserved demand, September 18 2014. US Patent App. 14/354,100.

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[7] Oded Berman and Dmitry Krass. The generalized maximal covering location problem. *Computers & Operations Research*, 29(6):563–581, 2002.

[8] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[9] U Narayan Bhat. *An introduction to queueing theory: modeling and analysis in applications*. Birkhäuser, 2015.

[10] Burcin Bozkaya, Seda Yanik, and Selim Balcisoy. A gis-based optimization framework for competitive multi-facility location-routing problem. *Networks and Spatial Economics*, 10(3):297–320, 2010.

[11] Alessia Calafiore, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, and Alex Singleton. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems*, 85:101539, 2021.

[12] Giuseppe C Calafiore, Carlo Novara, Francesco Portigliotti, and Alessandro Rizzo. A flow optimization approach for the rebalancing of mobility on demand systems. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5684–5689. IEEE, 2017.

[13] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, 2013.

[14] Enhui Chen, Zhirui Ye, Chao Wang, and Mingtao Xu. Subway passenger flow prediction for special events using smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1109–1120, 2019.

[15] Longbiao Chen, Daqing Zhang, Leye Wang, Dingqi Yang, Xiaojuan Ma, Shijian Li, Zhaohui Wu, Gang Pan, Thi-Mai-Trang Nguyen, and Jérémie Jakubowicz. Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 841–852, 2016.

[16] Lu Cheng, Zhifu Mi, D'Maris Coffman, Jing Meng, Dining Liu, and Dongfeng Chang. The role of bike sharing in promoting transport resilience. *Networks and Spatial Economics*, pages 1–19, 2021.

[17] Mung Chiang. *Networked Life: 20 Questions and Answers.* Cambridge University Press, 2012.

[18] Federico Chiariotti, Chiara Pielli, Andrea Zanella, and Michele Zorzi. A dynamic approach to rebalancing bike-sharing systems. *Sensors*, 18(2):512, 2018.

[19] Chicago. City of chicago data portal. `https://data.cityofchicago.org/`, 2021.

[20] Jacob Cosman and Nathan Schiff. Monopolistic competition in the restaurant industry. Technical report, mimeo, John Hopkins University, 2019.

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

[22] Justin Cranshaw and Tae Yano. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In *NIPS Workshop of Computational Social Science and the Wisdom of the Crowds*, 2010.

[23] Divvy. Divvy for everyone member agreement. `https://www.divvybikes.com/d4ememberagreement`, 2020.

[24] Divvy. How do i know if my bike is docked properly? `https://help.divvybikes.com/hc/en-us/articles/360033484451-How-do-I-know-if-my-bike-is-docked-properly-`, 2020.

[25] Divvy. Divvy: Chicago's bike share program. `https://www.divvybikes.com/`, 2021.

[26] 4SQ Eng. What neighborhood is the "east village" of san francisco?

[27] ESPN. Reds vs. cubs - summary - july, 8, 2018. `https://www.espn.com/mlb/game?gameId=380708116`, 2018.

[28] Cheng Feng, Jane Hillston, and Daniël Reijsbergen. Moment-based probabilistic prediction of bike availability for bike-sharing systems. In *International Conference on Quantitative Evaluation of Systems*, pages 139–155. Springer, 2016.

[29] Jon Edward Froehlich, Joachim Neumann, and Nuria Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[30] Nicolas Gast, Guillaume Massonnet, Daniël Reijsbergen, and Mirco Tribastone. Probabilistic forecasts of bike-sharing systems for journey planning. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 703–712, 2015.

[31] Raymond Gerte, Karthik C Konduri, Nalini Ravishanker, Amit Mondal, and Naveen Eluru. Understanding the relationships between demand for shared ride modes: case study using open data from new york city. *Transportation research record*, 2673(12):30–39, 2019.

[32] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[33] Qiao-Chu He, Tiantian Nie, Yun Yang, and Zuo-Jun Max Shen. Beyond rebalancing: Crowd-sourcing and geo-fencing for shared-mobility systems. *Available at SSRN 3293022*, 2019.

[34] Pierre Hulot, Daniel Aloise, and Sanjay Dominik Jena. Towards station-level demand prediction for effective rebalancing in bike-sharing systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 378–386, 2018.

[35] Ramon Iglesias, Federico Rossi, Rick Zhang, and Marco Pavone. A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems. *The International Journal of Robotics Research*, 38(2-3):357–374, 2019.

[36] John R Jensen and Ryan R Jensen. *Introductory geographic information systems.* Pearson Higher Ed, 2012.

[37] Pablo Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, 74(3):035101, 2006.

[38] Pablo Jensen. Analyzing the localization of retail stores with complex systems tools. In *International Symposium on Intelligent Data Analysis*, pages 10–20. Springer, 2009.

[39] Feifan Jia, Haiying Li, Xi Jiang, and Xinyue Xu. Deep learning-based hybrid model for short-term subway passenger flow prediction using automatic fare collection data. *IET Intelligent Transport Systems*, 13(11):1708–1716, 2019.

[40] Kevin Johnston, Jay M Ver Hoef, Konstantin Krivoruchko, and Neil Lucas. *Using ArcGIS geostatistical analyst*, volume 380. Esri Redlands, 2001.

[41] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[42] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. Geo-spotting: mining online location-based services for optimal retail

store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM, 2013.

[43] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[44] Zhaoyu Kou and Hua Cai. Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Physica A: Statistical Mechanics and its Applications*, 515:785–797, 2019.

[45] Alexander Kubis and Maria Hartmann. Analysis of location of large-area shopping centres. a probabilistic gravity model for the halle–leipzig area. *Jahrbuch für Regionalwissenschaft*, 27(1):43–57, 2007.

[46] Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[47] Anran Li and Kalyan Talluri. Estimating demand with unobserved no-purchases on revenue-managed data. *Available at SSRN 3525773*, 2020.

[48] Yang Li, Xudong Wang, Shuo Sun, Xiaolei Ma, and Guangquan Lu. Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. *Transportation Research Part C: Emerging Technologies*, 77:306–328, 2017.

[49] Yexin Li, Yu Zheng, and Qiang Yang. Dynamic bike reposition: A spatio-temporal reinforcement learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1724–1733, 2018.

[50] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2015.

[51] Frank Limehouse and Robert E McCormick. Impacts of central business district location: A hedonic analysis of legal service establishments. *US Census Bureau Center for Economic Studies Working Paper No. CES 11-21*, 2011.

[52] Junming Liu, Leilei Sun, Weiwei Chen, and Hui Xiong. Rebalancing bike sharing systems: A multi-source data smart optimization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014, 2016.

[53] Junming Liu, Leilei Sun, Qiao Li, Jingci Ming, Yanchi Liu, and Hui Xiong. Functional zone based hierarchical demand prediction for bike system expansion. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 957–966, 2017.

[54] M Donald MacLaren. The art of computer programming. volume 2: Seminumerical algorithms (donald e. knuth). *SIAM Review*, 12(2):306–308, 1970.

[55] Ashkan Negahban. Simulation-based estimation of the real demand in bike-sharing systems in the presence of censoring. *European Journal of Operational Research*, 277(1):317–332, 2019.

[56] Eoin O'Mahony and David B Shmoys. Data analysis and optimization for (citi) bike sharing. In *Twenty-ninth AAAI conference on artificial intelligence*. Citeseer, 2015.

[57] OpenStreetMap. Openstreetmap copyright and license. `https://www.openstreetmap.org/copyright`, 2021.

[58] OpenWeatherMap. Current weather and forecast - openweathermap. `https://openweathermap.org/`, 2021.

[59] Konstantinos Pelechrinis and Wayne Winston. A skellam regression model for quantifying positional value in soccer. *Journal of Quantitative Analysis in Sport (in print)*, 2021.

[60] Sergio Porta, Vito Latora, Fahui Wang, Salvador Rueda, Emanuele Strano, Salvatore Scellato, Alessio Cardillo, Eugenio Belli, Francisco Cardenas, Berta Cormenzana, et al. Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488, 2012.

[61] Sergio Porta, Emanuele Strano, Valentino Iacoviello, Roberto Messora, Vito Latora, Alessio Cardillo, Fahui Wang, and Salvatore Scellato. Street centrality and densities of retail and services in bologna, italy. *Environment and Planning B: Planning and design*, 36(3):450–465, 2009.

[62] Stuart S Rosenthal and William C Strange. Evidence on the nature and sources of agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2119–2171. Elsevier, 2004.

[63] Hamid R Sayarshad and Joseph YJ Chow. Non-myopic relocation of idle mobility-on-demand vehicles as a dynamic location-allocation-queueing problem. *Transportation Research Part E: Logistics and Transportation Review*, 106:60–77, 2017.

[64] Arieh Schlote, Bei Chen, and Robert Shorten. On closed-loop bicycle availability prediction. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1449–1455, 2014.

[65] Jasper Schuijbroek, Robert C Hampshire, and W-J Van Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257(3):992–1004, 2017.

[66] Lauren M Scott and Mark V Janikas. Spatial statistics in arcgis. In *Handbook of applied spatial analysis*, pages 27–41. Springer, 2010.

[67] Michael Sheret. Note on methodology: The coefficient of variation. *Comparative Education Review*, 28(3):467–476, 1984.

[68] John G Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A*, 109:296, 1946.

[69] N V Smirnov. Approximate distribution laws for random variables, constructed from empirical data. *Uspekhi Mat. Nauk*, 10:179–206, 1944.

[70] Shivaram Subramanian and Pavithra Harsha. Demand modeling in the presence of unobserved lost sales. *Management Science*, 67(6):3803–3833, 2021.

[71] Glen L Urban. A mathematical modeling approach to product line decisions. *Journal of marketing research*, 6(1):40–47, 1969.

[72] Liqin Wang, Yongfeng Dong, Yizheng Wang, and Peng Wang. Non-symmetric spatial-temporal network for bus origin–destination demand prediction. *Transportation Research Record*, page 03611981211039844, 2021.

[73] Rui Xue, Daniel Jian Sun, and Shukai Chen. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. *Discrete Dynamics in Nature and Society*, 2015, 2015.

[74] Qiang Yan, Kun Gao, Lijun Sun, and Minhua Shao. Spatio-temporal usage patterns of dockless bike-sharing service linking to a metro station: A case study in shanghai, china. *Sustainability*, 12(3):851, 2020.

[75] Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen, and Thomas Moscibroda. Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th annual international conference on mobile systems, applications, and services*, pages 165–178, 2016.

[76] Ji Won Yoon, Fabio Pinelli, and Francesco Calabrese. Cityride: a predictive bike sharing journey advisor. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 306–311. IEEE, 2012.

[77] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

[78] Yang Yue, Yan Zhuang, Anthony GO Yeh, Jin-Yun Xie, Cheng-Lin Ma, and Qing-Quan Li. Measurements of poi-based mixed use and their relationships with neighbourhood vibrancy. *International Journal of Geographical Information Science*, 31(4):658–675, 2017.

[79] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 793–803, 2019.

[80] Chunjie Zhou, Pengfei Dai, Fusheng Wang, and Zhenxing Zhang. Predicting the passenger demand on bus services for mobile users. *Pervasive and Mobile Computing*, 25:48–66, 2016.

[81] Zoba. Winter is coming... for 90% of all micromobility markets. `https://medium.com/zoba-blog/winter-is-coming-for-90-of-all-micromobility-markets-d2085bedb2a7`, 2019.