# Modeling Urban Venue Dynamics through Spatio-Temporal Metrics and Complex Networks

**Krittika D'Silva**

Department of Computer Science and Technology
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Jesus College September 2020

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Krittika D'Silva
September 2020

# Acknowledgements

I'd like to first thank my supervisor, Cecilia Mascolo, for her support, guidance, and patience. Her mentorship over my time at Cambridge has enabled me to learn and grow as an independent researcher. I am also grateful for the friends I have made through the Computer Lab. I also thank my coauthors for their guidance and collaborations over the course of my PhD and for the support of Lise Gough, the backbone of the department.

I am grateful for the friends I have made in Cambridge. Tori and Lior who have been there for all of the ups and downs of the journey. Alice, for being my closest supporter, mentor, and friend. Yani, Bruno, and Val for bringing a piece of Canada to Cambridge. I made many life-long friends through Jesus College and will always cherish the memories and experiences we've shared. I am also grateful to the Gates Cambridge Scholarship for not only supporting my studies but also for the opportunity to join a community for life.

I am fortunate to have worked on a number of exciting projects outside of Cambridge during my PhD. I am grateful to the entire team at the United Nations Global Pulse Lab in Jakarta who made me feel welcome. I am especially thankful to Derval Usher and George Hodge. My eight weeks at NASA FDL were a whirlwind and I am thankful to the organizers for the opportunity and to my team for their friendship, collaborative spirit, and support.

As an undergraduate, I had a number of mentors who guided me through formative experiences which led me to Cambridge. Dr. Sanders and Dr. Hafner guided me through many undergraduate research projects and spurred my interest in research. I am grateful for their support and to John Cagle who has my been my mentor and steadfast supporter. I am also grateful to Nicki Dell for her guidance and support. I am thankful to have had the opportunity to work with and learn from Bill Thies at Microsoft Research. And to Luis Ceze and Karin Strauss for the opportunity to work on a number of exciting projects. I will forever be thankful to have met and learned from Gaetano Borriello who showed me the value in working to serve others.

I am grateful to Kevin for the constant support and love. I also thank my brother, Suneet, who has been one of my closest supporters every step of my journey and my grandparents for their love. Finally, I thank my parents for their sacrifices in supporting me through my education. Their encouragement and wisdom have shaped the person I am today.

# Abstract

The ubiquity of GPS-enabled devices, mobile applications, and intelligent transportation systems have enabled opportunities to model the world at an unprecedented scale. Urban environments, in particular, have benefited from new data sources that provide granular representations of activities across space and time. As cities experienced a rise in urbanization, they also faced challenges in managing vehicle levels, congestion, and public transportation systems. Modeling these fast-paced changes through rich data from sources such as taxis, bikes, and trains has enabled prediction models capable of characterizing trends and forecasting future changes. Data-driven studies of urban mobility dynamics have been instrumental in helping deliver more contextual services to cities, support urban policy, and inform business decisions. This dissertation explores how novel algorithmic architectures and techniques reveal and predict business trends and urban development patterns.

The research informing this dissertation harnesses principles from network science, modeling cities as connected networks of venues. Building upon a foundation of research in complex network theory, urban computing, and machine learning, we propose algorithms tailored for three computing tasks focused on modeling venue dynamics, characteristics, and trends. First, we predict the demand for newly opened businesses using insights from movement patterns across different regions of the city. Through this analysis we demonstrate how temporally similar areas can be successfully used as inputs to predict the visitation patterns of new venues. Next, we forecast the likelihood of business failure through a supervised learning model. We analyze the value of varying features in predicting business failure and explore their impact across new and established venues and across different cities worldwide. Finally, we present a deep learning architecture which integrates both spatial and topological features to predict the future demand for a venue. These works highlight the power of complex network measures to quantify the structure of a city and inform prediction models.

This dissertation leverages vast amounts of data from spatio-temporal networks to model venue dynamics. The research puts forward evidence to support a data-driven study of geographic systems applied to fundamental questions in urban studies, retail development, and social science.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

From financial transactions and medical sensors to transportation systems and satellite imagery, in this day and age large volumes of data are continuously generated across numerous domains. An infographic assembled by Raconteur showed that daily an average of 500 million tweets are created, 294 billion emails are sent, and 95 million Instagram photos and videos are shared [136]. At the current pace, 1.5 quintillion bytes of data are produced each day [111]. This data and the information it encodes offer opportunities to characterize and understand systems. Machine learning algorithms have emerged as models that are able to detect patterns in large datasets. Their insights surpass traditional metrics as machine learning models can build sophisticated representations of the world and inform decisions not only in the field of technology but also in policy, media, and retail, among others [62, 110].

Over the past decade, mobile applications, including social media outlets and ride-sharing platforms, have emerged as valuable sources of data for urban environments. The ubiquity of GPS-enabled devices, in conjunction with sensor-powered mobile applications, has facilitated the study and understanding of the social and geographical movements of individuals at an unprecedented scale. Location-based social networks (LBSNs) are popular systems that allow their users to connect and interact online in relation to real-world places such as businesses and landmarks [117]. As millions of users interact on LBSNs they generate sequences of digital traces with high spatio-temporal granularity. The movements of these users elucidate numerous insights such as the connectivity of different areas of a city, their characteristics at varying times of day, and their growth patterns over time [188, 189].

A broad range of stakeholders can benefit from these types of insights, including government agencies, non-profit organizations, and private entities. Prior academic work has shown that businesses in particular benefit from insights that enable them to predict future demand trends and, in turn, future profits [39, 171]. Moreover, there are a plethora of other use cases for these datasets within the context of businesses. Financial transaction data, such as information on deliveries, payments, and invoices, are used to build pricing models [171]. Social media data, including tweets, posts, and check-ins, are used to predict future consumer behavior [105]. Crowd-sourced reviews from a variety of applications such as Yelp, TripAdvisor, and Google Maps, are used to inform trends in customer perceptions [70].

Historically, businesses collected data via surveys and their individual transaction information [80]. These datasets, however, are limited as they provide a singular view of a venue without contextual data from bordering businesses and neighborhoods; these approaches are also often subjective in methodology and narrow in scope [182]. This was similarly the case in urban studies, which in this context refers to a field exploring the development and growth of cities. Historically, urban studies heavily relied on the use of conventional population surveys and individual interviews [101]. These methods are generally expensive, time-consuming, and provide a static view of the system [182]. In contrast, more recent data from sensors and social media provide an objective, large-scale, and often real-time perspective [69]. Unlike theoretical and survey-based approaches, empirical data-driven computational methodologies also have the potential to quantify our understanding of complex systems.

The past decade has seen a surge in research around and applications of machine learning. These algorithms refit and improve their representations of data over training intervals; therefore, a key requirement is vast quantities of data. As large datasets have become increasingly available and access to cloud computing has become increasingly accessible, machine learning has risen to prominence [81]. These models have proven superior in many contexts as they do not require hard-coded strict rules as expected by some previous models [170].

In parallel, network theory has proven to be a versatile tool to model the complexity of systems. Network theory is the study of graphs as a representation of relationships between discrete objects [13]. Networks are found naturally in community structures, transportation routes, and biological organisms [29, 169]. As such, network science has been applied broadly beyond graph theory because of its ability to capture complex interconnected relationships. There is tremendous potential to integrate network approaches with machine learning techniques to generate innovative models. However, research in this space is still very much in its nascency and further work building methodologically novel architectures and models is imperative to identify new patterns, associations, and insights in data. These models must be contextual and consider differences in attributes across domains. This is the focus of the research in this dissertation. It presents machine learning methodologies for urban environments. Specifically, it focuses on spatio-temporal models built using techniques from network science.

This dissertation advances the thesis that the study of urban areas through location-based networks can enhance our understanding of human mobility and enable us to model complex venue dynamics. Using mobility and business popularity data, we build unique and contextual prediction models of the popularity of new venues, the likelihood of failure of existing venues, and the future demand of existing businesses. This work builds upon prior research in this space and demonstrates the value in spatio-temporal machine learning models of venue activity. Insights from this work aim to inform new policy, methodologies, and further research efforts. In the following sections we present the thesis (Section 1.1) and contributions (Section 1.2) of this dissertation.

## 1.1 Thesis and Substantiation

As discussed above, the recent proliferation of urban datasets, especially those related to mobility, offers unique opportunities to model and understand the complex movements of individuals at the level of a city and across different cities worldwide. Within this context, this dissertation seeks to advance the field of urban computing using innovative machine learning architectures and techniques to offer methodologies as well as insights into human mobility patterns and their impact on business dynamics. It incorporates insights from network science by modeling cities and businesses as individual entities within a connected system. The thesis of this dissertation is that *the spatio-temporal study of urban areas through location-based analytics and complex network theory can advance our understanding of human mobility and enable us to model venue dynamics.*

In this dissertation we substantiate this statement with three core research themes. First, we explore how identifying clusters of temporally similar areas in a city can forecast the weekly popularity dynamics of a new business. Next, we demonstrate how mobility-derived features can predict the failure of an existing business. Lastly, we introduce a novel deep learning framework to model the future popularity of existing businesses. In the next section, we detail the contributions of this dissertation and discuss how they are linked to the research themes described here.

## 1.2 Contributions and Chapter Outline

This dissertation makes several novel contributions to the field of urban computing with machine learning. The contributions are detailed in the following chapters:

## Chapter 3: New Venue Demand Prediction

The research in this chapter harnesses mobility data from Foursquare, a LBSN, to build proxies for urban activities and understand how new venues become popular with time. Estimating revenue and business demand of a newly opened venue is paramount as these early stages often involve critical decisions such as first rounds of staffing and resource allocation. Traditionally, this estimation has been performed through coarse-grained measures such as observing numbers in local venues or businesses at a similar location. We propose a novel prediction framework able to use characteristic temporal signatures of places together with k-nearest neighbor metrics to capture similarities among urban regions. This methodology forecasts the weekly popularity dynamics of a new venue. We also show the model is capable of forecasting the popularity of a new venue one month following its opening by using locality and temporal similarity as features. We demonstrate that temporally similar areas of the city can be used as inputs for predictions of the visit patterns of new venues, with an improvement of 41% compared to a random selection of wards. Our models have the potential to impact the design of location-based technologies and decisions made by new business owners.

## Chapter 4: Modeling Business Closure

We build a supervised prediction model to investigate whether mobility-derived features can foretell the failure of retail businesses, over a six-month horizon, across 10 cities spanning the globe. As in the previous chapter, we use data from Foursquare to model business dynamics. We hypothesize that the survival of such a retail outlet is correlated with not only venue-specific characteristics but also broader neighborhood-level effects. Through careful statistical analysis of mobility data we uncover a set of discriminative features which include a neighborhood's static characteristics, venue-specific customer visit dynamics, and a neighborhood's mobility dynamics. Our contributions are not only in building the model but also in identifying and analyzing a range of features to demonstrate the predictability of survival of Food and Beverage businesses. We demonstrate that classifiers trained on our features can predict survival with high accuracy. We achieve AUCs of 0.85 and 0.90 for Singapore and New York respectively with corresponding precision/recalls at $\approx 80\%$. These results represent a 10-15% improvement in accuracy over past work in this space. We also show that the impact of such features varies across new and established venues and different cities. Besides achieving a significant improvement over past work on business vitality prediction, our work demonstrates the vital role that mobility dynamics play in the economic evolution of a city.

### Chapter 5: Studying Business Growth

We propose a novel deep learning framework that aims to model the popularity and growth of urban venues. We create a novel architecture with a Graph Convolutional Network (GCN) that integrates both spatial and topological features into a temporal model which predicts the demand of a venue at the subsequent time-step. Our experiments demonstrate that our model can learn spatio-temporal trends of venue demand and consistently outperform baseline models. Our experiments conducted using Foursquare data demonstrate that our model can learn spatio-temporal trends of venue demand and consistently outperform baseline models. Relative to state-of-the-art deep learning models, our model reduces the Root Mean Square Error by $\approx 28\%$ in London and $\approx 13\%$ in Paris. Our approach demonstrates the power of complex network measures and GCNs in building prediction models for urban environments. The model has numerous applications within the retail sector to better estimate venue demand and growth. More broadly, the methodology and results can support policymakers, business owners, and urban planners to develop models to characterize and predict changes in urban settings.

## 1.3 Publication List

Included below are the academic publications which make up the research conducted to substantiate this thesis. These consist of my efforts and technical contributions but would not have been possible without the support and guidance of my collaborators. Some of the publications are directly related to this dissertation. In particular, Chapter 3 draws from [45] and [46], Chapter 4 is based on [44], and Chapter 5 builds on work currently under review.

**Works related to this dissertation**

- [45] If I build it, will they come? Predicting new venue visitation patterns through mobility data. Krittika D'Silva, Anastasios Noulas, Mirco Musolesi, Cecilia Mascolo, Max Sklar. *In Proceedings of the ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL).* Redondo Beach, CA, USA. November 2017.

- [46] Predicting the temporal activity patterns of new venues. Krittika D'Silva, Anastasios Noulas, Mirco Musolesi Cecilia Mascolo, and Max Sklar. *In EPJ Data Science.* May 2018.

- [44] The Role of Urban Mobility in Retail Business Survival. Krittika D'Silva, Kasthuri Jayarajah, Anastasios Noulas, Cecilia Mascolo, Archan Misra. *In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT).* Volume 2, Issue 3. Ubicomp 2018. Singapore. October 2018.

- Modeling Urban Dynamics with Multi-Modal Graph Convolutional Networks. Krittika D'Silva, Jordan Cambe, Cecilia Mascolo, and Max Sklar. Currently under review.

**Other publications:**

- A Generative Machine Learning Framework For Synthesizing Symptomatic ECG Astronaut Health Data. Eleni Antoniadou, David Belo, Krittika D'Silva, Brian Wang, Brian Russell, Frank Soboczenski, Annie Martin, Graham Mackintosh, and Tianna Shaw. *In NASA Human Research Program Investigators' Workshop.* Galveston, Texas, USA. January 2020.

# Chapter 2

# An Overview of Urban Modeling

This dissertation builds machine learning prediction models of venue dynamics using data from location-based networks. The previous chapter provided an overview of these research themes. This chapter explores prior literature and the current state-of-the-art modeling human mobility and place dynamics. We also discuss the role of machine learning and applications of network science.

The structure of this chapter is as follows. In Section 2.1, we outline research in human mobility demonstrating that trends in populations often follow a universal distribution. In Section 2.2, we present past work using machine learning for prediction tasks in urban environments. These include models applied to transportation systems, public services, and business dynamics, among others. In Section 2.3 we provide an overview of network science and network metrics. We summarize prior research using network structures in machine learning prediction tasks for urban environments. Finally, we summarize the chapter in Section 2.4 and discuss the contributions of this dissertation in its research community.

## 2.1 Modeling Human Mobility

Cities are complex systems that are constantly changing over time. With public transportation and commuters following routine mobility patterns and visitors and tourists quasi-stochastically traveling between different neighborhoods, urban environments have tremendous potential to be modeled and analyzed [123]. Characterizing human mobility in urban environments has numerous applications. It is especially valuable when considering the sustained trends towards urbanization in recent history. Rapid urbanization has had a dramatic impact on cities worldwide [83].

A recent report from the United Nations showed that the percentage of the world's population living in urban areas increased from 30% in 1950 to approximately 55% in

2018 [115]. These trends are expected to continue in the near term and have implications for resource allocation, urban planning, and economic growth. They introduce both new challenges and opportunities. On one hand, rapid urbanization can result in higher rates of air pollution, energy consumption, and traffic congestion [102]. Conversely, urbanization can also result in increased job opportunities and access to educational, health, and cultural resources [185]. The ability to quickly model and interpret these changes can inform urban governance and mitigate potential issues.

Amongst the first works on large-scale analyses of human mobility was in 2006 when Brockmann et al. [23] tracked American dollar bills as a proxy for human movements around the United States. The researchers created an online game called *Where's George* that allowed users to input the serial number of selected dollar bills into their database with details on where that bill was found. Through this game, Brockmann et al. gathered millions of trajectories of bills around the United States and subsequently modeled movement patterns nationally. The authors showed that the distributions of an individual's traveling distance decay following the power law. Specifically, the probability of traveling a certain distance for a certain time interval followed $P(r) \propto r^{-(1+\beta)}$ where $\beta = 0.59$.

In 2008, Gonzalez et al. [67] examined the trajectories of 100,000 mobile phone users over a six-month period. Their analysis of the data found that human movements contained a high degree of temporal and spatial regularity. The regularity was so apparent that the authors, after correcting for differences in travel distances, found that individual travel patterns could be modeling with a single spatial probability distribution. Similarly, Noulas et al. [117] looked at trajectories of mobility using Foursquare data from cities around the world. They conducted this study with over 10 million transitions from 34 cities. Their work described a rank-distance model that estimated the probability of transitioning from one place to another is inversely proportional to the power of their rank, defined as the number of intervening opportunities between those two places. Song et al. [165] showed that the movements of individuals have a periodicity as users tend to frequently revisit locations and often are at venues of significance such as their work or home. Works such as these highlight the predictability of certain types of human movements.

Other research has explored the regularity of urban mobility as users follow their daily routine [123, 152]. Pappalardo et al. [124] used mobile phone and GPS data to identify trends in the travel distances of individuals. Their work found two distinct mobility profiles, returners and explorers. Returns limit their mobility to a few regularly frequented locations. Conversely, the mobility of explorers consisted of a broader range of locations which could be characterized as having their recurrent and overall mobility being very dissimilar. Their work created a gravity model which distinguished the two separate classes of mobility. Pappalardo et al. showed that the two classes also had a strong impact on social interactions. Specifically, their work found that individuals tended to engage in social interactions with others of the same profile.

Cho et al. [33] also conducted research using cell data to model large scale movements. They proposed a model of human mobility, based on check-ins, using separate spatial and temporal Gaussian components. The work looked at spatial, temporal, and social trends and found that social relationships can help explain between 10% to 30% of movements while periodic behavior explained between 50% to 70%. The authors built a model to predict future movements of individuals which incorporated both of these insights. Their model more accurately predicted mobility patterns over their baselines by a factor of two.

Giannotti et al. [64] modeled mobility using GPS receivers placed on-board tens of thousands of cars. From the trajectory patterns of these cars, the authors extracted mobility trends and insights. This work found subgroups of vehicles characterized by similar movements that informed their *mobility atlas*, a catalog of the mobility behaviors of an area.

The studies described above conducted large-scale analyses of human mobility. These studies produced many insights, including showing that trends in human mobility often follow a universal distribution. They demonstrated that the predictability of human travel can be used to build models of urban mobility and identify characteristics of different cities and neighborhoods within cities. Harnessing data to model urban trends forms the backbone of this dissertation.

## 2.2 Urban Spatio-Temporal Machine Learning

There are countless applications of new datasets to urban studies. Data sources such as social media, transport, and financial records each portray cities from a different perspective. These data are especially important within the context of *smart cities* where urban areas are supported by computer-aided systems and sensors [181]. The data collected in *smart cities* are used in a broad range of applications, including optimizing the efficiency of operations and supporting urban planning. Insights from the data can also help support the development of services such as transportation systems, trash collection, and ambulance deployment.

While insights from these data are crucial to answer many important questions, there are limitations associated with the data. Firstly, while specific demographics vary based on application, users of LBSNs tend to be biased towards individuals who are younger and technologically inclined [72]. As such, LBSN data generated by older individuals or those in more rural or poor areas can often be more limited and under-represented. In these cases, it can be helpful to incorporate many different types of data into the model including those that are from both passive and active sensing modalities. In addition to demographics, the type of data input into LBSNs can also be biased. For example, users may prefer to checkin to certain types of venues over others or prefer not to post about their movements based on privacy or security concerns. These are important considerations to make when extrapolating from the data and inferring trends. Lastly, an inherent limitation of these data is that they only represent the online experiences of individuals which can differ from their offline experience. These limitations in the data highlight the importance of understanding each dataset and the contextual limitations associated with it.

The research in this dissertation uses datasets from urban environments to create machine learning methodologies. We next explore prior research in spatio-temporal urban modeling using machine learning in conjunction with contextual insights from cities for a broad range of applications.

**Transportation Systems:** As cities have seen a rise in urbanization they have faced challenges that include managing vehicle levels, congestion, and public transportation systems. Modeling these changes with data from sources such as taxis, bikes, and trains has enabled prediction models capable of characterizing trends and forecasting future changes [3].

Li et al. [103] examined the trajectories of users of a bike-sharing system to predict changes in the distribution of bikes in the city. The authors spatially clustered bike stations and then used a Gradient Boosting Regression Tree to predict the total number of bike check-outs in the city and the rent proportion for each cluster of bikes. Their work used correlations in the temporal profile of stations together with variations in the weather patterns of the city as features of the model. It more accurately predicted check-out percentages across bike stations over baseline approaches with up to a 0.23 reduction in the Error Rate. Chiang et al. [32] applied a similar approach to taxi data. They used the locations of taxi bookings in Singapore to estimate the future spatio-temporal distribution of bookings. Their model divided the city into grid cells and subsequently used a Poisson process to model the daily rate of bookings in each cell. The authors used a Gaussian Mixture Model to predict the demand for taxis within each individual cell. Chiang et al. showed that this model can be used to detect deviations in trends based on events such as concerts or large festivals in a given area.

Dai et al. [38] used data of the inflow and outflow of individuals to a metro station in Zhengzhou City, China to predict future passenger demand patterns. Their framework combined outputs from k-nearest neighbors and Adaboost models into a Random Forest Classifier. This approach harnessed both spatial and temporal clustering to analyze trends. Specifically, they aggregated origin-destination pairs based on their time interval and location and subsequently generated inflow and outflow characteristics for each station. These features were fed into their machine learning model to predict future demand. Models such as these have the potential to support metro authorities' decision processes from day to day and aid in emergency management situations.

Integrating anomalous data or real-time events is often crucial to accurately model the uncertainty of real-world systems. However, early detection of anomalies in transportation data is a research question in itself. Zhang et al. [183] created a similarity-based algorithm that incorporated an anomaly score for each region and time of the city using the road network, taxi, and bike datasets. The anomaly scores were fed into a Support Vector Machine to detect anomaly patterns across the cities. Kong et al. [96] developed an anomaly detection algorithm for traffic flow which correlated anomalies of similar temporal and spatial segments and fed them into an ensemble machine learning model. Their model input data from road networks, taxi mobility, and metro usage to detect regional irregularities across a city. These works exemplify the value of using data from many types of transportation systems in informing a prediction.

**Public Services:** In addition to transportation systems, machine learning has proven to be valuable when applied to public services. Use-cases include modeling waste collection, crime prevention, and environmental changes, among others. We next discuss applications machine learning to support public services in cities worldwide.

Zimmerman et al. [192] used machine learning to monitor air quality. Their model took in data from low-cost sensors that track the concentrations of different air pollutants such as $CO, CO_2$, and $NO_2$. The authors created a machine learning model using random forest techniques to build calibration models capable of improving the performance of low-cost air quality sensors. Chou et al. [35] monitored and predicted the water quality levels in numerous reservoirs across Taiwan. Their work used data collected over ten years from 20 different water reservoirs. The data included attributes of the water such as the pH level and phosphorus concentration. The authors worked to use these data to predict the Carlson's Trophic State Index of the water, a common metric of water quality. Their architecture consisted of an ensemble neural network model based on a tiering methodology. These works demonstrate the value in sensor data in conjunction with machine learning to predict trusted metrics such as that of water or air quality.

Puissant et al. [134] used satellite imagery and object-based image detection to build a random forest model capable of predicting the number of urban trees. Their work mapped and monitored changes in urban green spaces which is valuable because of their impact on air, climate, and water quality. The authors' model build machine-generated features for different segments of an image and then computed a metric to quantify the level of urbanization of a plot of land. Vermeiren et al. [172] similarly used satellite imagery to model the city of

Kampala, the capital of Uganda. They modeled the growth of the city and predicted future growth patterns and the potential impact of that growth on the quality of life of citizens.

Johnson et al. [86] used a gradient boosting tree to model waste collection across the city of New York. Their model trained on historical weather and waste tonnage data to predict future waste generation. It accurately ($R2 > 0.88$) forecast weekly waste generation data for all of the 232 areas of New York City across different types of waste taking into consideration seasonal variations, weather events, and holidays. Models such as these can help inform the development of and changes to many types of services in a city.

**Venue Dynamics:** Modeling urban venue characteristics and dynamics is an active area of research [93, 137]. Check-ins, where users register a visit to a specific venue, are rich sources of data to model the dynamics of businesses and venues more broadly [7].

Lichman et al. [105] predicted the spatial distributions of check-in and geolocation data from two popular LBSNs, Twitter and Gowalla. Their kernel density estimation method captured and predicted individual and population spatial patterns applied across the state of California. Scellato et al. [150] built a prediction framework of the number of future visits to a venue as well as the duration of stay of those visits. The model determined the significant locations of users based on their movements and then applied a nonlinear time series analysis to predict their future movements. Baumann et al. [19] combined attributes from 18 prediction algorithms to predict the likelihood of a transition occurring between two places in a city. The authors used a heuristic method based on majority voting which created not only a highly accurate model but also one which included confidence metrics of the outputted predictions. Kwon et al. [100] used reviews and tips from Foursquare and Yelp to predict *user churn*, the percentage of users who stop using the service. The authors found distinct engagement patterns of users who leave the service which they used to inform a deep learning prediction task to distinguish churning users from those who stay.

Lu et al. [109] developed a framework to characterize the lifecycle of Points of Interest (POIs) by incorporating urban geographic features and human mobility dynamics. Using this data, they formulated a classification task to predict the life status of POIs, determining whether the are *booming*, *decaying*, or *stable*. Yang et al. [178] proposed LBSN2Vec, a hypergraph embedding approach for LBSN data. Their model improved the performance of friendship and location prediction tasks through automatic feature learning (i.e. learned node embeddings). Their hypergraph included user-user edges, such as friendships, and user-time-POI-semantic hyperedges, such as check-ins.

These works exemplify how traditional machine learning algorithms can be applied to urban environments to better predict trends and characteristics. Data representing the dynamics of a city can be helpful in not only predicting movements within a city but also in helping to understand or characterize a neighborhood or area within the larger city. A popular tool to model the interconnectedness of cities it that of network science which we discuss next. The research in this dissertation uses network science as a framework for modeling venue dynamics by characterizing the relationships between different areas of a city and their different levels of interconnectedness.

## 2.3 Network Theory

Over the past decade, networks have become an adaptable tool to represent systems with interacting components, from brain pathways in living organisms to online social networks [29, 169]. Network science is valuable in the context of urban environments because it captures their complexity and numerous interconnected areas. It also enables abstractions from the inherent complexity of these systems while providing a robust framework to model interactions between the different constituents. In computer science, network theory is defined as the study of graphs in which the graphs are composed of sets of many connected nodes that vary in characteristics and which interact in different ways [13]. Networks are found naturally in nature, including social networks, information networks, and biological networks. These networks can either be directed, where edges link two vertices symmetrically, or undirected, where edges link two vertices asymmetrically. The edges between nodes can also be weighted in which a number (i.e. a weight) is assigned to each edge [53]. There are numerous types of networks, a few of which are briefly described below:

- **Spatial network:** A spatial network is one in which the nodes are embedded in a metric space and the edges are associated with a distance, such as Euclidean distance [15]. A pair of nodes in the network is therefore connected if the distance between the pair of nodes is smaller than a set radius. Numerous examples of spatial networks emerge in the real world, including public transportation journeys, air travel routes, and mobile phone networks [40, 67]. All of these examples involve systems in which the underlying space is relevant.

- **Social networks:** A social network is one in which the vertices are representations of social interactions between the nodes [97]. A pair of nodes is connected if the two nodes have an interaction within a given time frame. Examples of social networks abound, including organizational arrangements in companies, criminal networks, and community structures [144, 87]. For each, the social connection between the different nodes in the networks is a crucial attribute.

- **Temporal networks:** Temporal networks are unique in that the links between nodes are only active at certain time points [75]. The attributes that define edges between the nodes can be any characteristic of the network. Examples of temporal networks include neural connectivity of the brain, disease proliferation, and communication networks [92, 48]. Temporal networks are frequently used to model the dynamics of systems.
- **Multi-Layer networks:** A multi-layer network is one in which there are numerous different types of relations between nodes in the networks [17, 95, 53]. Each layer contains a set of intra-layer connections of a specific attribute and the network as a whole also contains inter-layer connections as well [41, 166]. For example, given a network of cities in the world, one layer could include air travel routes between the cities while another layer could encode postal delivery systems between the cities. This multi-modal approach can often create a more sophisticated representation of a system.

The network science approach described above provides an opportunity to utilize network metrics such as centrality, connectivity, and betweenness to quantify attributes of the system [15, 14]. Measures such as clustering, neighbor degree, and assortativity help characterize trends within the dataset. As one example, modularity can detect community structure that exists within the network. This dissertation uses network theory with large urban datasets to harness network features and properties to glean novel observations about venues and cities worldwide. These specific properties are described in more detail next.

### 2.3.1 Network Properties

Networks have numerous attributes that can be calculated to analyze the characteristics of nodes or subsets of nodes in the network. We next give an overview of traditional network properties which are throughout this dissertation.

- **Size:** The size of a network refers to the number of nodes in the network.
- **Density:** The density of a network is defined as the number of existing edges divided by the total number of possible edges. This metric characterizes how tightly connected the nodes in a network are to each other.
- **Degree:** The degree of a node is defined as the number of edges connected to it. Naturally, the degree of a network is the average degree of all nodes within the network. In a directed network, this metric can be extended to calculate in or out degree by measuring the total number of incoming or outgoing edges. In weighted networks, the

edge weights can be incorporated and normalized to create a weighted degree metric. At the level of an individual node, this indicates the importance of a node.

- **Distance:** Network distance is defined as the length of a path from a given node to another node. For a weighted graph, this can incorporate the weight of the edge between a pair of nodes. This metric is used as a basis for many other metrics defined below.

- **Shortest path length:** The shortest path length for a pair of nodes is the shortest distance between those two nodes. The average shortest path length for a network is defined as the average shortest path between all pairs of nodes. This metric is an indication of how closely connected nodes in the network are to each other.

- **Clustering coefficient:** The clustering coefficient is measured as the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of links. It provides an indication of the tendency of a given network to form triangles, that is to gather locally into fully connected groups. It varies between 0 and 1 with higher values implying a higher number of triangles in the network (see [116] for more details). The clustering coefficient for a network reflects how tightly connected or overlapped neighborhoods of nodes are to each other.

- **Closeness centrality:** The closeness metric is a measure of how "near" a node is to all other nodes in the network. It is calculated by measuring the sum of the shortest distances between a node and all other nodes. It accounts for the tendency of categories to be close to each in terms of shortest paths [116]. It varies between 0 and 1 where a higher closeness centrality score for a node suggests higher proximity to other nodes in the network.

- **Modularity:** Modularity is a metric that indicates how well-defined communities are within the network [22]. Networks with high modularity have dense connections between nodes in modules but sparse connections between nodes in different modules. Modularity values fall within the range -1 and 1, with greater positive values indicating greater presence of community structure.

### 2.3.2 Machine Learning in Urban Network Science

The rise of ubiquitous computing in the past decade has generated many works that use networks to model the intricacies of urban environments, from individuals moving between venues to neighborhood connectivity changing over time [28, 47]. We next explore past research using machine learning to model cities with network science.

**Traffic Prediction:** Traffic prediction is a popular application for complex networks. Dong et al. [43] used network properties to model congestion, showing how temporal profiles of freeways can be used to better predict traffic flow rate. Their model also examined how factors, such as spatial trends from neighboring areas, contributed to variations in traffic flows. Sole-Ribalta et al. [164] used networks to identify traffic hotspots on road networks. Their methodology identified traffic junctions that are susceptible to high levels of congestion if mobility demand were to increase in those areas. Liu et al. [108] proposed a method to represent the trajectory of moving vehicles on road networks by using topic modeling to reveal their spatial interaction patterns. These works help to not only understand the flow of traffic on the road but also to quantify the impact of real-world events on the surrounding area.

**Venue Dynamics:** Complex networks have also been applied to model the dynamics and characteristics of individual venues. Prior work has looked to predict the ideal location for a business as well as whether or not it will succeed [110, 62]. Numerous works have also explored whether a transition or a link will occur between a pair of locations [33, 174]. Others have explored which spatial features are most likely to attract individuals to a given venue [119]. Fortuna et al. [55] used a Foursquare network of venues and users to predict the existence of a check-in between a user-venue pair in a city. Chen et al. [30] modeled the evolution of different neighborhoods in a city using the urban transport network. Noulas et al. [119] used Foursquare transitions within the city of London and worked to predict whether a transition from a given venue to another was likely to occur at the next time step. Their work looked at a number of network properties, including common neighbors, degree centrality, and neighbor overlap, as features for the link prediction task. Their results showed that a gravity model, which incorporates geographic distance, interaction quantity, and mobility dynamics, best predicts the likelihood of a new link appearing.

**Urban Spatial Structure:** There is a large body of literature using network analysis to characterize and model the structure of urban environments. Zhong et al. [188] described how the spatial structure of a city changes over time relative to its connectedness. They utilized the network properties of a city to identify the structure of city hub, centers, and borders by constructing a weighted directed graph from these travel records. They discussed the concept of *polycentrality*, suggesting that Singapore is becoming progressively more polycentric over time as new subcenters and communities emerge as the city grows. Zhong et al. [189] similarly worked to detect the urban spatial structure of functional centers by using transportation data of individuals. Their model used a centrality as well as attractiveness index to detect functional centers and the impact those centers have on transportation and the movement of individuals in the city.

Capponi et al. [26] examined the attractiveness of local businesses to their neighborhoods based on the time of day and day of the week. The authors modeled the demand for venues using complex network metrics such as closeness centrality to build features for a supervised learning prediction task. Their machine learning model substantially outperformed traditional urban metrics in analyzing the popularity of a business. Work by Ratti et al. [141] used 12 billion phone calls from across Great Britain as proxies of human interactions. Their model used spectral modularity optimization, a network metric that partitions nodes, to unveil spatial structures and visualize different regions of the country. Their work also quantified the impacts of partitioning different regions Great Britain by the potential disruption it would have on the human network. Williams et al. [177] refined typical spatial network properties, such as centrality and reachability, from a temporal perspective. Their work is relevant because, as in many real-life systems, interactions between nodes are generally non-instantaneous and involve a component of time. The authors applied their metrics to a number of mobility datasets, from students interacting with each other as they move around a college campus to flight patterns as planes travel around the world.

**Multilayer Models:** Multilayer networks have been applied to a number of real-world systems from massive online multiplayer games to disease propagation [24, 168, 5, 76]. In particular, multilayer networks have been used to model urban characteristics and dynamics as they provide the ability to analyze cities a from multidimensional perspective [78, 5, 60]. Work conducted by Aleta et al. [6] pursued the former, building a model using the bus, metro, and tram data from nine cities of varying sizes. The authors found a number of interesting universal properties related to the underlying structure of the cities. Their multilayer transport network showed the noticeable impact that transport changes, such as new lines or disruptions to existing lines can have on the mobility of individuals in a city. Hristova et al. [77] quantified the social and geographic brokerage of Foursquare venues and used that knowledge to characterize the social diversity of a location. The authors built a multilayer network using Foursquare check-ins and Twitter connections to build both a social as well as spatial graph. This enabled them to model venues that serve a *bridging* role while others that serve a *bonding* role.

**Graph Neural Networks:** Graph Neural Networks (GNNs) are a class of deep learning algorithms that extend neural networks to process data represented in a graphical structure [149, 148, 184]. The goal of a GNN is to learn a state embedding which encodes information of the neighborhood of each node. Numerous algorithms have been proposed to create graph embeddings all of which aim to represent similarity in the embedding space that approximates similarity in the original network. Generally, these models train in an unsupervised manner using only the graph structure. DeepWalk is popular algorithms which uses random walks to inform the loss function [131]. The DeepWalk algorithm first performs random walks on the nodes in the graph to generate node sequences. It then runs the skip-gram algorithm to learn the embedding of each node based on the generated node sequences. This approach mirrors the word2vec algorithm used in natural language processing. GraphSage was proposed as a more generalizable and inductive approach to building graph embeddings. [71]. The GraphSage algorithm aims to incorporate the aggregation of each node in its neighborhood into the representation. It has a three-step approach through a mean aggregator, LSTM aggregator, and pooling aggregator. GNNs have increasingly been applied to a wide range of domains, including text reasoning, disease classification, and image segmentation [104, 16, 143]. Although very recent, there is a growing body of work using GNNs for urban applications.

Yu et al. [180] used a GNN to model time-series data of traffic flow in Beijing and California. The authors presented a deep learning architecture with two gated sequential convolution layers and one spatial graph convolution layer in between. This approach enabled them to incorporate the nonlinearity and complexity of traffic flows by aggregating spatial and temporal features across both their datasets. Zhao et al. [186] similarly forecast traffic patterns through GNNs. Their two-part model first aggregated data from different time windows to gather spatial features and then fed those features into a multi-graph model that found non-Euclidean correlations among different spatial areas. Chai et al. [27] built a multi-graph convolutional neural network model to predict available bike numbers at stations across a city. Their model enabled a fine-grained analysis at the level of individual stations by constructing multiple graphs of the bike-sharing system: a distance graph, an interaction graph, and a correlation graph. Their algorithm then merged the graphs together into one fused representation which incorporated many learned features of the individual stations.

Previous work has demonstrated the power of network representations to model urban dynamics and properties. Our work in Chapter 5 presents a novel approach to modeling time-series data with GNNs with an architecture that combines long short-term memory (LSTM) neural networks with GNNs to integrate both spatial and topological features into a temporal model for venue demand prediction. Exploration through new frameworks and deep learning algorithms is imperative as they reveal new insights and properties of real-world systems. One of the novel contributions of this thesis is the development of innovative spatio-temporal networks models that use machine learning architectures to characterize and predict future venue dynamics. This is discussed in further detail in the following section where the contributions and future outlook of this dissertation are presented.

## 2.4 Present Dissertation and Future Outlook

This chapter reviewed current research modeling human mobility and place dynamics. This is a fast moving area of research with continued advances in machine learning architectures, network models, and computational capacities. In this chapter, we began with an overview of human mobility examining a variety of ways in which mobility patterns were modeled prior to the advent of machine learning and *big data*. We subsequently presented applications of research that used machine learning to model urban environments. We then gave an overview of network science and network metrics and reviewed prior research using networks in machine learning prediction tasks for cities.

This dissertation builds upon prior research and takes steps towards building a deeper understanding of and prediction tools for venue dynamics in urban environments. In Chapter 3, we harness the temporal popularity patterns of venue categories to predict the weekly temporal profile of a newly established venue. Our approach is based on two key principles. First, we exploit the fact that despite their differences, neighborhoods in a city can exhibit a high degree of temporal synchronicity, even if they are geographically located far from one another. Second, we make use of the principle of locality that suggests that the temporal profile of venues in the same neighborhood are highly correlated as individual movements in a city tend to be constrained by distance. We combine the properties of synchronicity and locality into a prediction framework that uses the characteristic temporal profile of neighborhoods in a city in conjunction with k-nearest neighbor metrics to capture similarities among urban regions. It used a Gaussian Process model to forecast the weekly popularity dynamics of new venues using spatial and temporal features. This work focused on one city, London, and specifically on all venues in the city that were newly opened. In Chapter 4, we expanded the venues considered to include retail venues from 10 cities worldwide. As the analysis included both new and established retail venues, the models had a more robust set of data to learn from. First, we identify and analyze a range of features and perform a comprehensive study to demonstrate the predictability of survival or failure of a business in the subsequent 6-month period. We next the generalizability of our approach, modeling multiple cities worldwide as well as differences between newly founded and established venues. We build supervised learning models to predict the likelihood of closure of retail venues. The models created a set of discriminative features that impacted the likelihood of survival of retail businesses. Finally in Chapter 5, we expanded the scope of the analysis to include all categories of venues. This provided a rich dataset to train a deep learning model to predict the popularity and growth of urban venues. We created a novel architecture using two Graph Convolutional Network (GCNs) We employ two GCNs that build spatial and topological characteristics respectively. Spatial features represent the location of venues relative to others and topological features represent the connectivity of a venue to other venues. These are then fed into a long short-term memory (LSTM) network which predicts the demand of a venue at the subsequent time-step.

One of the biggest challenges in modeling is that models can often be overly-simplistic, considering only a small subset of data types and attributes. This dissertation presents novel methodologies for existing data taking into account complex attributes of cities and their dynamics. The intersection of machine learning and network science has numerous valuable applications, which we highlight through the work in the subsequent chapters.

# Chapter 3

# Predicting the Temporal Activity Patterns of New Venues

Following an introduction in Chapter 2 to research on human mobility, machine learning, and complex networks, this chapter introduces a model which uses machine learning to predict the future demand of new venues. Specifically, this chapter demonstrates the potential in using Gaussian Processes in conjunction with temporal similarity metrics to build urban forecasting models.

The main contribution of this chapter is a prediction framework able to use characteristic temporal signatures of places together with k-nearest neighbor metrics to forecast weekly popularity dynamics of a new venue establishment in a city. We show that temporally similar areas of a city can be successfully used as inputs of predictions of the visit patterns of new venues, with an improvement of 41% compared to a random selection of wards as a training set for the prediction task. We apply these concepts of temporally similar areas and locality to the real-time predictions related to new venues and show that these features can effectively be used to predict the future trends of a venue. Our findings have the potential to impact the design of location-based technologies and decisions made by new business owners.

## 3.1 Introduction

Cities are complex systems that constantly change over time. From city dwellers that commute to work on a weekday morning to visitors who arrive in town for business or leisure, the urban landscape is transforming at a fast pace. The way in which city neighborhoods become popular over time has been a fundamental area of study in traditional urban studies literature as it is critical to city governance [89, 18]. The rise of mobile technologies and collective sensing in the last decade has contributed to the generation of large datasets that describe activity dynamics in cities and has created new opportunities for research in the area; for example, several works have proposed the use of cellular data to understand collective mobility dynamics and inform planning decisions [142, 25, 140, 20, 85]. Beyond cellular data, the increasing popularity of services like Twitter and Foursquare has yielded new inputs for capturing the *heartbeat* of a city [57, 159, 187]. Further, trends in credit card transactions have also been used to study human patterns across space and time [160].

On the venue level, temporal dynamics and the spatial configuration of urban activities has helped decide where to geographically place new retail facilities [84, 91] as well as to power mobile applications such as local search [154] by exploiting place temporal dynamics. Nevertheless, little work has looked at predicting what happens after a new venue opens in a city neighborhood; *will it become popular?* and moreover, *at which times of the week should the owner of a retail facility expect high volumes of customer traffic?* This information is important during the early stages of a new business when staffing levels must be decided, supplies bought, and opening times established.

In this chapter, starting from the premise that mobility in a city is driven by local urban activities, we provide an analytical framework that captures the popularity dynamics of urban neighborhoods. We then exploit these temporal patterns across areas to predict the popularity dynamics of newly established venues. The primary data input we use for our study is a longitudinal dataset from location-based service Foursquare describing mobility in terms of user *check-ins* at public venues in the city of London. Our approach can be summarized as follows:

- **Temporal characterization of urban activities across regions:** First, we show how the temporal profile of an area in terms of the number of mobile users that visit over the course of a week varies significantly from neighborhood to neighborhood. These temporal profiles are shaped by the daily and weekly circadian rhythms of moving populations as well as their choice of specific urban activities at key times. Further, we demonstrate how the popularity dynamics of venue categories give rise to the temporal patterns of the urban areas that contain them, highlighting how urban activities and population levels at a neighborhood are inherently interconnected temporal processes.

- **Predicting the popularity dynamics of newly established venues:** Next, we harness the temporal popularity patterns of venue categories to predict the weekly temporal profile of a newly established venue. Our approach is based on two key principles. First, we exploit the fact that despite their differences, neighborhoods in a city can exhibit a high degree of temporal *synchronicity*, even if they are geographically located far from one another. For example, the city of London includes neighborhoods such as SoHo and Camden Town, both of which attract a young late-night crowd. These shared traits could mean these neighborhoods are likely to become popular at similar times. Second, we make use of the principle of *locality* that suggests that the temporal profile of venues in the same neighborhood are highly correlated as individual movements in a city tend to be constrained by distance. We combine the properties of *synchronicity* and *locality* in a novel k-nearest neighbor model to determine similarities between neighborhoods in cities and use this understanding to predict the characteristic demand curve of a new venue. We use classic Gaussian Processes along with a k-nearest neighbor approach to build a model which predicts a new venue's popularity temporal profile. Our results perform significantly better than our random baseline, decreasing the normalized root mean square error by 41%.

- **Real time demand prediction of new venues:** Finally, given a new venue, our goal becomes to predict how the demand changes as that venue matures over time. With each progressive month of growth for a new venue, it is to be expected that some venues flourish and grow in popularity over time while others may be less successful and see their demand decline over subsequent months. Although sparsity can be an issue when working at a fine spatial granularity, making the formulation of a regression problem challenging, we show that it is possible to improve the prediction of expected levels of visits for the next time step by using an approach based on neighborhood *synchronicity* and *locality*. We train a Gaussian Process model on the month to month trends of check-ins to other venues and wards and use these as inputs to forecast the popularity of a new venue. In doing so, we incorporate recent changes in urban mobility which could arise, for instance, due to the presence of new events nearby or other anomalies such as transport disruptions.

This approach enables a fine-grained dynamic estimation of activity for new venues. Obtaining analytics in this context can help can business owners predict demand in dynamics for their business and therefore plan better the provision of services to their customers.

The remainder of the chapter is organized as follows. In Section 3.2, we motivate our work and introduce the related work in the area. Section 3.3 gives an overview of our approach and Section 3.4 introduces a formalization of our framework. Section 3.5 reports on our temporal analysis of venues and neighborhoods in London. In Section 3.6, we describe a method for predicting the characteristic temporal profiles of a set of new venues using a batch-learning approach, whereas in Section 3.7 we present an analysis of the real-time extension of this approach. We conclude the chapter with Section 3.8 discussing our results.

## 3.2 Related Work

The ubiquity of GPS sensors on mobile devices as well as the introduction of the mobile web have been game changers with respect to the scales and types of data available about human mobility. The rise of services such as Foursquare and more generally applications that rely on geo-tagging technologies (e.g. Twitter, Instagram, Flickr) combined with the accessibility to the corresponding APIs have offered novel views to collective mobility activity in cities [118, 122]. This has led to more granular representations of urban activities across space and time [57, 187, 159], and has been used to characterize cities in terms of their urban growth patterns [37] and cultural boundaries in terms of their culinary patterns [158]. Additionally, such location-based technologies have significantly improved the quality of experience for mobile users as they navigate the city. Foursquare's data science team exploited the weekly temporal visitation patterns of venues to power its local search engine [154]. Google Maps recently incorporated the feature *popular times* that appears when search results about places are shown to users [68], while Facebook launched its in-house place search service [51]. The commercialization potential of such services has naturally expanded beyond the realm of location-based technologies and has contributed to coining the term *location intelligence* when referring to business intelligence relying on geospatial data. In this direction, several works have appeared on retail optimization in cities, including the identification of the best location to open a new shop [84, 91] or the ranking of areas according to their real estate value [59].

Before datasets from location-based services became available, those collected from cellular networks paved the way for understanding the collective dynamics of urban activities. Ratti et al. in [140] present one of the first works that demonstrates how urban landscapes transform in real time as populations move around the city. Beyond dynamic visualizations, the authors in [142, 25] characterize in statistical terms urban activity variations and provide interpretation on the observed patterns in terms of the underlying urban activities, such as transport and residential land uses, which drive population volumes regionally. Becker et al. in [20] use cellular data to characterize mobility trends across different metropolitan areas, while the authors in [85] propose using cellular data as an alternative to travel surveys so that more accurate spatio-temporal representations of mobility flows are obtained. Urban transport data has often been another source for capturing city dynamics [146, 162].

While in many of the works mentioned above, information on the temporal visitation patterns of users to locations has been used as input, none has looked at predicting the temporal signatures of visits to venues per se. This is where the primary novelty of the present chapter lies. Considering the temporal patterns of user visits at newly established venues as our main prediction task we are hoping to offer new insights on location-based analytics for business owners that would empower them to make more informed choices on staffing, provision of goods, resources in the early days of their new business. We also envision the use of similar approaches to inform several tasks that are applicable in the urban domain. These may include the spatial deployment of taxi fleets and pooling services [12, 147] or the allocation of police or ambulance resources [138, 156].

## 3.3 Our Approach at a Glance

In this section, we describe how our approach to the prediction of visitation patterns to new venues harnesses temporal similarities of urban neighborhoods. Our analysis revolves around the concept of a characteristic weekly temporal profile, a time series representing typical changes in demand of a given entity over the course of a week. We explore characteristic profiles of venues, which fluctuate based on user mobility patterns, and neighborhoods that are composed of venue profiles.

We begin with the premise that venue categories have different characteristic weekly temporal profiles. These profiles represent variations in demand based on a user's propensity to visit that category at a given hour of the week. For example, the category of Travel & Transport is likely to correlate with changes in rush hour traffic while Food could instead be dependent on typical meal times. Different neighborhoods in a city have characteristic weekly temporal profiles which are made up of contributions from different categories. We posit that neighborhoods which have similar temporal profiles or similar contributions of venues to their temporal profile could be predictors for each other. We apply this idea towards the analysis of new venues: *given a new venue in a given neighborhood in a city, can we use the the demand profile of venues in temporally similar wards as predictors for our new venue of interest?* For new venues opening up in a city, often no prior information is known about the expected popularity or demand dynamics. The ability to approximate and better understand these metrics can be crucial for the success of a new business owner.

This characteristic temporal curve provides a static representation of the typical changes in demand of a venue over the course of a week. We build upon our analysis of new venues by dynamically predicting how the demand of a venue will change. Starting from one week after a venue has opened, we show that we can use data from venues with similar characteristic curves to more accurately predict the changes in demand at the next time step.

## 3.4 Notation and Definitions

### 3.4.1 Dataset

LBSNs have recently experienced a surge in popularity, attracting millions of users around the world. The widespread adoption of these services in addition to location-sensing mobile devices has created a wealth of data about the mobility of humans in cities. Foursquare, a popular location-centric media platform, enables users to check into different locations and share that information with their friend group. As of August 2015, Foursquare had more than 50 million active users and more than 10 billion check-ins [176].

In this chapter, we use a longitudinal dataset describing urban mobility and activity patterns in Greater London that spans three years and millions of check-ins. For each venue, we have the following information: geographic coordinates, specific and general category, creation date, total number of check-ins, and number of unique visitors. The specific and general categories fall within Foursquare's API of hierarchical categories. A full list of the categories can be found by querying the Foursquare API. General categories are overarching groups to one of which each specific category is assigned. Examples of general categories could include *Food* or *Travel & Transport* while examples of specific categories could be *Chinese Restaurants* or *Italian Restaurants*, which both aptly fall under the category of *Food*. In addition to data about venues, the dataset also contains *transitions* within London. A transition is defined as a pair of check-ins by an anonymous user to two different venues within the span of three hours. A transition is identified by a start time, end time, source venue, and destination venue. Our dataset includes 18,018 venues and 4,000,040 transitions for Greater London. The dataset comprises of check-ins from December 2010 to December 2013. This dataset was obtained through a collaboration with Foursquare.

### 3.4.2 Formalization

In this section, we introduce a formalization of our model. Electoral wards are the main building blocks of administrative geography in the United Kingdom; Greater London consists of 649 electoral wards and these spatial units uniquely identify London boroughs [54]. We use wards $w \in \mathbf{W}$ as a means of subdividing Greater London. We also consider venues $v \in \mathbf{V}$. A venue has a precise geographic location in a ward. A venue $v$ is represented with a tuple $v = < loc, g, s >$ where $loc$ is the geographic location of the venue, $g$ its general category and $s$ is its specific category.

We define a *time interval* $t$ as the interval $[t\Delta, (t+1)\Delta]$ of duration $\Delta$. For example the time interval $t = 0$ indicates the interval $[0, \Delta]$, the time interval $t = 1$ indicates the time interval $[\Delta, 2\Delta]$ and so on. In our work, each time interval represents distinct hours and do not overlap.

**Definition 1: Temporal Profile of a Ward.** Similarly we define the temporal profile of a ward $w$ in an interval $[0, T]$ as the following sequence (i.e, time series):

$$C^w[0,T] = \{c_t^w\} \quad \text{with} \quad t = 0, 1, \ldots T-1 \tag{3.1}$$

where $c_t^w$ is the *total* number of check-ins in the ward $w$ during the time interval $t$.

**Definition 2: Temporal Profile of a Venue.** We define the *temporal profile of a venue v* in an interval $[0, T]$ as the following sequence (i.e, time series):

$$C^v[0,T] = \{c_t^v\} \quad \text{with} \quad t = 0, 1, \ldots T-1 \tag{3.2}$$

where $c_t^v$ is the *total* number of check-ins to venue $v$ during the time interval $t$.

**Definition 3: Aggregate Temporal Profile of Venues of a Generic (Specific) Category in a Ward.** We then define $\mathbf{V}_{g,w}$ as the set of the venues of *generic* category $g$ in a ward $w$. Similarly, we define $\mathbf{V}_{s,w}$ as the set of the venues of *specific* category $s$ in a ward $w$.

Therefore, the *aggregate temporal profile* of venues of *generic* category $g$ in a ward $w$ in a time interval $[0, T]$ is defined as the following sequence (i.e, time series):

$$C^{V_{g,w}}[0,T] = \{c_t^{g,w}\} \quad \text{with} \quad t = 0, 1, \ldots T-1 \tag{3.3}$$

where $c_t^{g,w}$ is the *total* number of check-ins to venues of general category $g$ in the ward $w$ during the time interval $t$. The temporal profile of venues of a specific category in a ward can be defined similarly.

## 3.5 Temporal Patterns of Mobile User Activity

Having formally defined the concept of temporal profiles, in this section, we discuss temporal trends of wards within Greater London and demonstrate how the composition of those wards plays a crucial role in creating a characteristic profile for that ward. We begin with an examination of all wards at one particular point in the day, highlighting that different categories dominate different wards at any given point. We then analyze more closely the characteristic temporal profile of two wards in London and discuss how their different category types contribute to their different overall profiles. We then quantify the similarity between the overall temporal profile of the 15 most popular wards in London and discuss how similarity in temporal visitation patterns could inform predictions for the temporal profile of a new venue.

### 3.5.1 Regional Temporal Activity Patterns

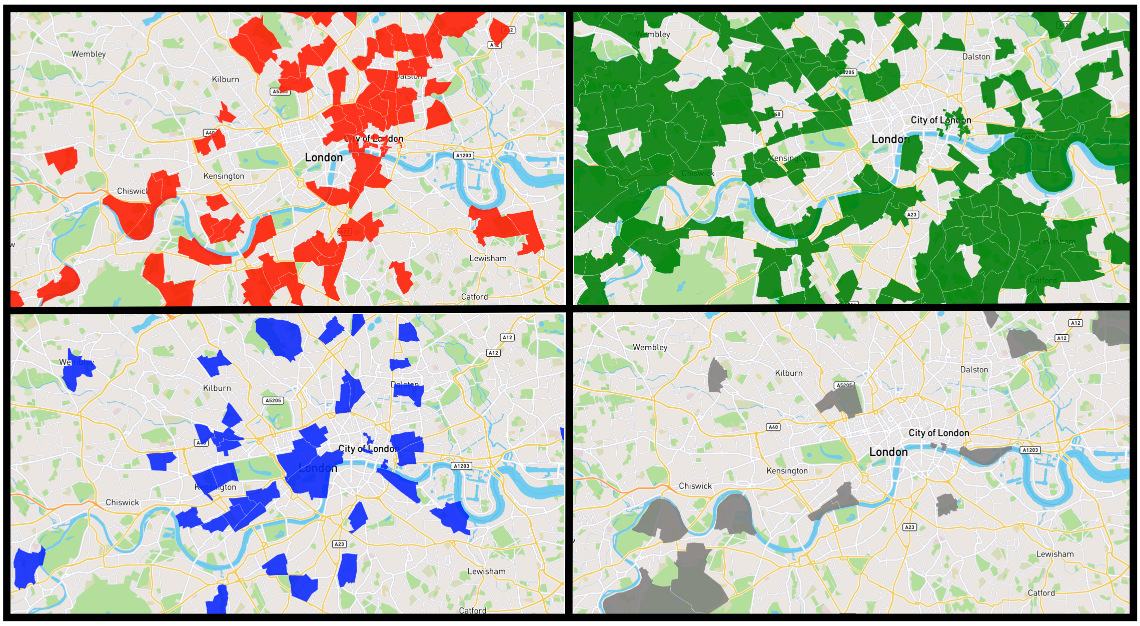



Fig. 3.1 The most popular category in each ward at 17:00. *Nightlife Spots* are represented in red, *Travel & Transport* in green, *Food* in blue, and *Outdoors & Recreation* in grey.

Looking broadly at all wards within the city of London, we choose one hour of the day to highlight the idea that the popularity patterns of different neighborhoods can be dominated by different categories. Figure 3.1 shows the most popular category in each ward in London in the time interval $t = 17$ where $\Delta = 1$ (i.e., between 5 to 6pm). For certain wards, this time of the day could be dominated by transport traffic as individuals commute towards home. For others, the most significant contributor could be nightlife, as individuals head to the pub for an evening drink. Similarities in the contribution of different categories to the overall temporal trend of a ward could be an indication that those wards attract individuals with similar demographics or have similar characteristics. An analysis of these similarities can be harnessed to better model, characterize, and profile different wards in a city.

Looking more closely at category types, Figure 3.2 presents the characteristic temporal profile of three categories: *Nightlife Spots*, *Colleges & Universities*, and *Gyms* or *Fitness Centers*. Each profile is a direct function of a users's propensity to visit at a given hour of the day and day of the week (T = 168). The profiles of different venue categories in a ward establish the overall profile of that venue. A close examination of different wards within the city of London and of the categories which make up those wards present a number of interesting insights on how those vary in terms of their temporal visitation patterns.

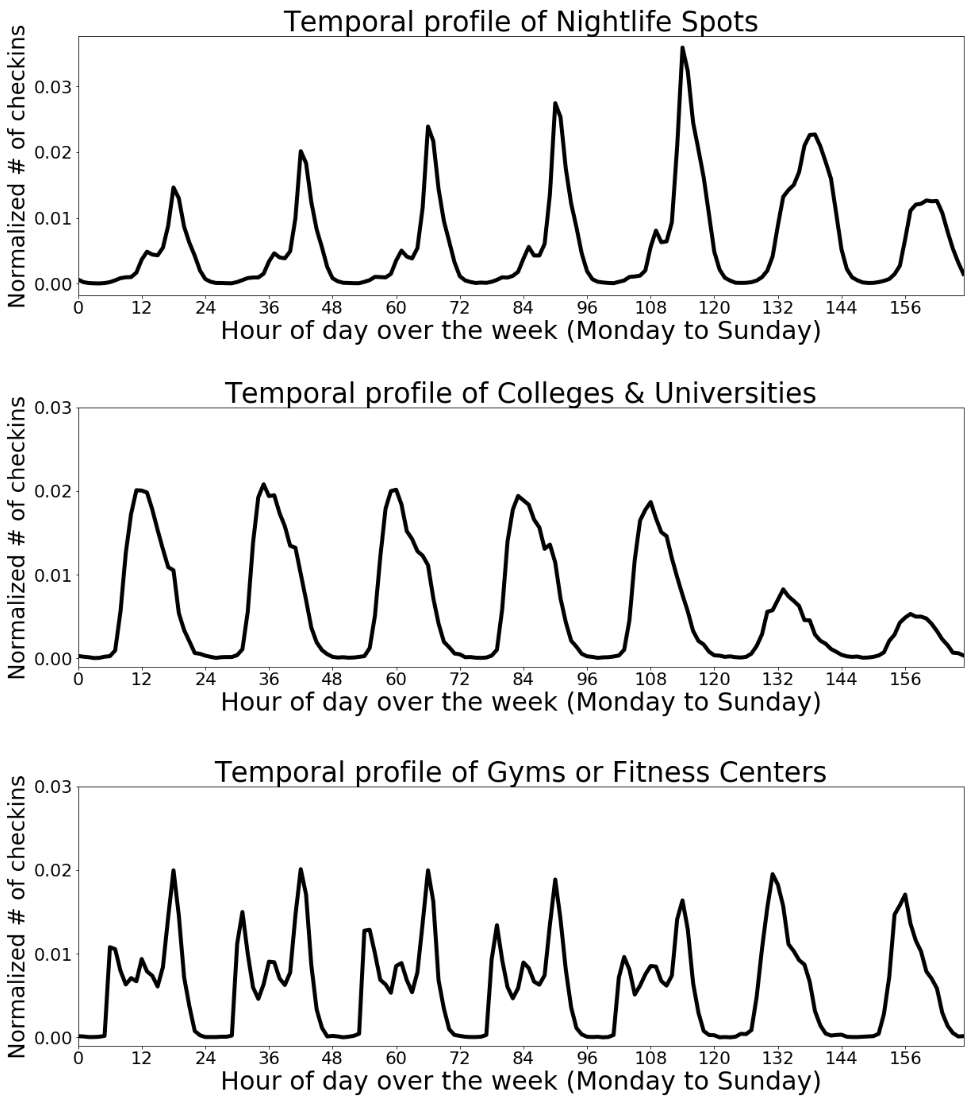



Fig. 3.2 Normalized temporal profile of different categories of venues.

In order to illustrate this let us consider two wards of interest: St. Pancras & Somers Town, which contains a major transportation hub, offices, and academic institutions, and Camden Town with Primrose Hill, which contains a variety of venues and tourist attractions. Figure 3.3 shows the average number of check-ins in each ward for each hour of the day over the course of one week, aggregating across a number of weeks. This signal creates a characteristic temporal profile which acts as a temporal signature for the ward. The overall signal, shown in black, is different for these two wards. The number of check-ins at Camden Town steadily increases over the course of the day while the number check-ins at St. Pancras has two large peaks, one in the morning and another in the evening. Examining the three main categories (*Food*, *Travel & Transport*, and *Nightlife Spots*) that characterize these two wards

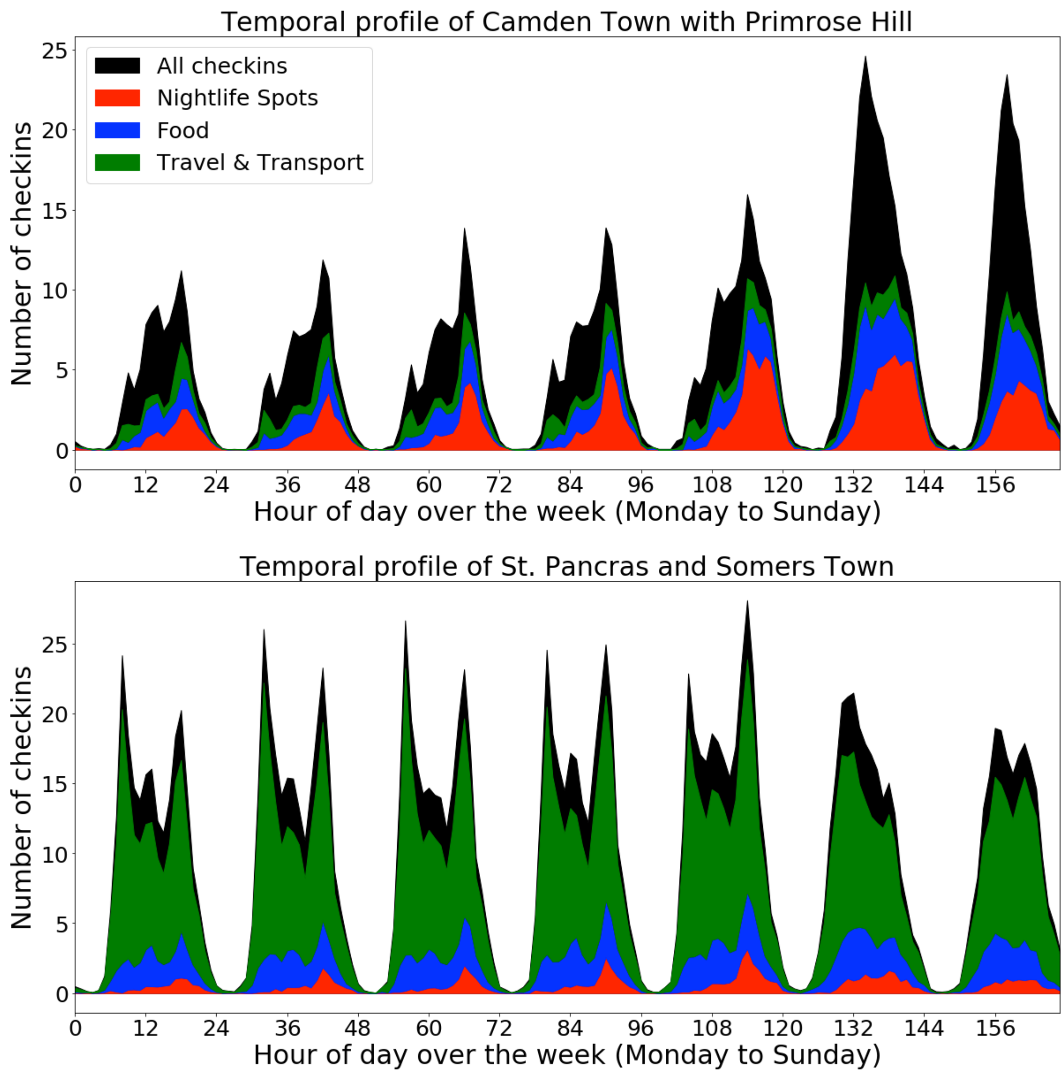



Fig. 3.3 Daily temporal profiles and category breakdown of St. Pancras & Somers Town and Camden Town with Primrose Hill, two contrasting wards in London.

can help to better understand this observation. Camden Town has significant contributions from *Nightlife* venues which gradually increases over the course of a day. Conversely, St. Pancras is dominated by *Travel & Transport*, causing the overall temporal profile of the ward to peak at rush hour. These trends suggest that Camden Town is likely a more youth dominated area while St. Pancras is a hub for commuters or travelers, as they actually are [10].

### 3.5.2 Utilizing Similarities in Visitation Patterns

Similar observations can be generalized to the rest of the wards in London. Different regions feature different degrees of similarity, an insight which we exploit in Section 3.6 to predict the characteristic temporal curves of new venues. We quantify the similarity between two temporal profiles using the Jensen-Shannon divergence (JSD) [106]. We use the JSD instead of the Kullback-Leibler divergence (KLD) since the former is a symmetric similarity measure between two functions whereas the latter is not. Our analysis showed that the JSD improved our percent accuracy by 7 to 10% over the KLD. The JSD between two wards $w_i$ and $w_j$ is calculated as follows:

$$JSD(C^{w_i}, C^{w_j}) = H\left(\frac{C^{w_i} + C^{w_j}}{2}\right) - \frac{H(C^{w_i}) + H(C^{w_j})}{2} \tag{3.4}$$

where $H$ is the Shannon entropy. The JSD provides an *information-theoretic* metric that quantifies how two profiles, which that can be seen as distributions over time, are similar. A low value of the JSD between the temporal profile of two wards represents a high similarity.

In Figure 3.4 we present the Jensen-Shannon divergence between the temporal profiles of the 15 most popular wards in London with regards to their total number of check-ins. There is an evident range in similarity between the various wards. For instance, Hyde Park is very similar to St. Pancras and Somers Town as both are central travel hubs that handle the large commuter flows to local corporate offices and government buildings. Other characteristic examples are wards such as St. James's and West End that attract a large tourist population visiting the attractions in the respective areas. These results suggest that similarities in temporal profiles can be useful indicators of similarities in the characteristics of two wards. We aim to use this insight for our overarching goal of predicting the characteristic temporal profile of a new venue. This similarity can be used for prediction.

In the next section, we explore how the temporal profile of a new venue becomes more and more stationary with each consecutive week and then demonstrate how similar wards can be used for the prediction task.

### 3.5.3 Temporal Visitation Patterns of New Venues

We now focus on the temporal characteristics of new venues. These venues represent an interesting case study as upon their launch, unlike existing venues or geographic areas, there is no historic information on their expected popularity patterns over time. We introduce next basic properties of the new venues data that we use during evaluation. We demonstrate how their temporal profiles converges to a stationary state over time, a process that will let us

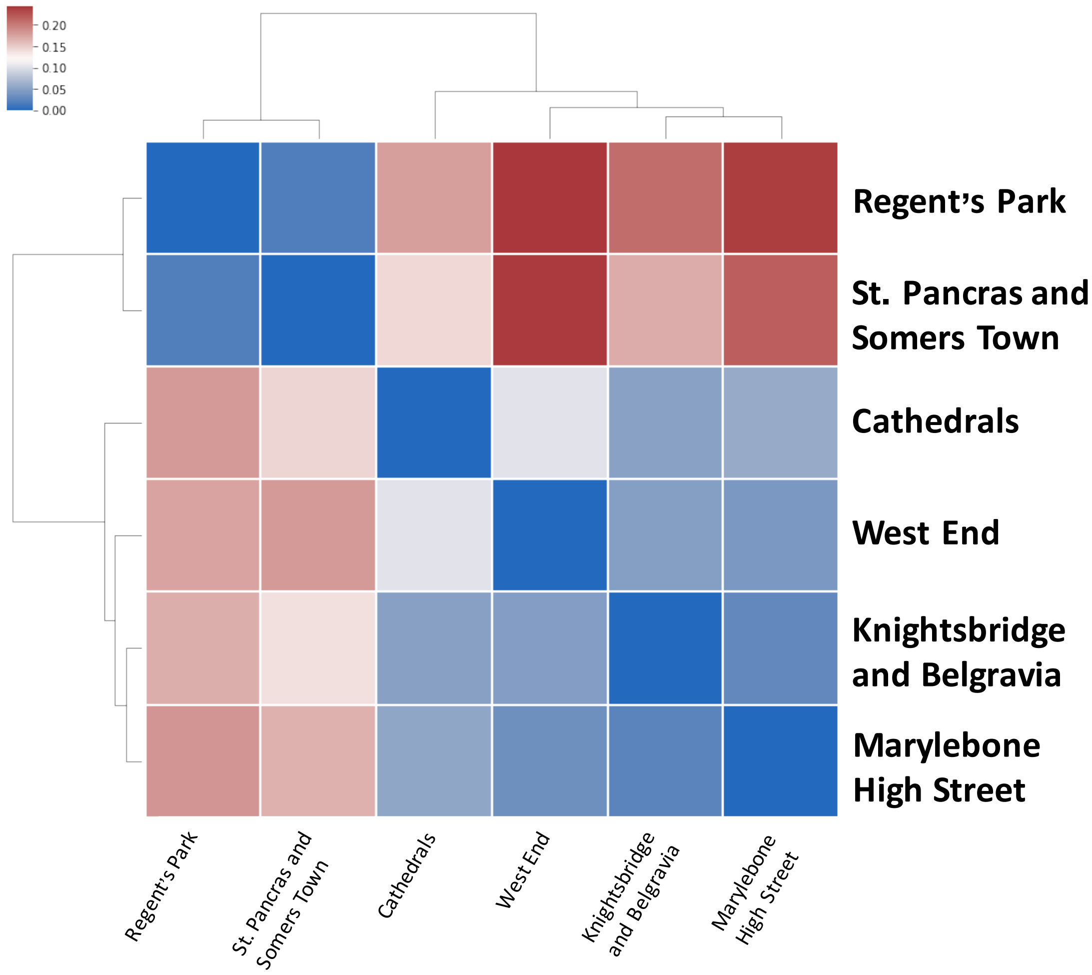



Fig. 3.4 J-S divergence of the characteristic weekly temporal profile of the 15 most popular wards. Smaller values signify a smaller divergence and thus more similarity.

define the prediction task presented in Section 3.6.

**Identification of new venues.** The Foursquare dataset includes a list of all public venues in the city of London. Critically for the present work the *creation time* of every venue is available in the database. The creation time refers to the date the venue was crowdsourced by Foursquare users. Prior research on Foursquare data has shown that venues added after June 2011 were highly likely (probability above 0.8) to actually be new venues opening in an area rather than existing venues being added to the system for the first time [37]. We look at all new venues that were added to Foursquare after June 2011 that had at least 100 check-ins. This results in a list of 305 venues which is used for the following analysis. Within this list of newly opened venues, 32% of the venues fall within the general category of *Food*,

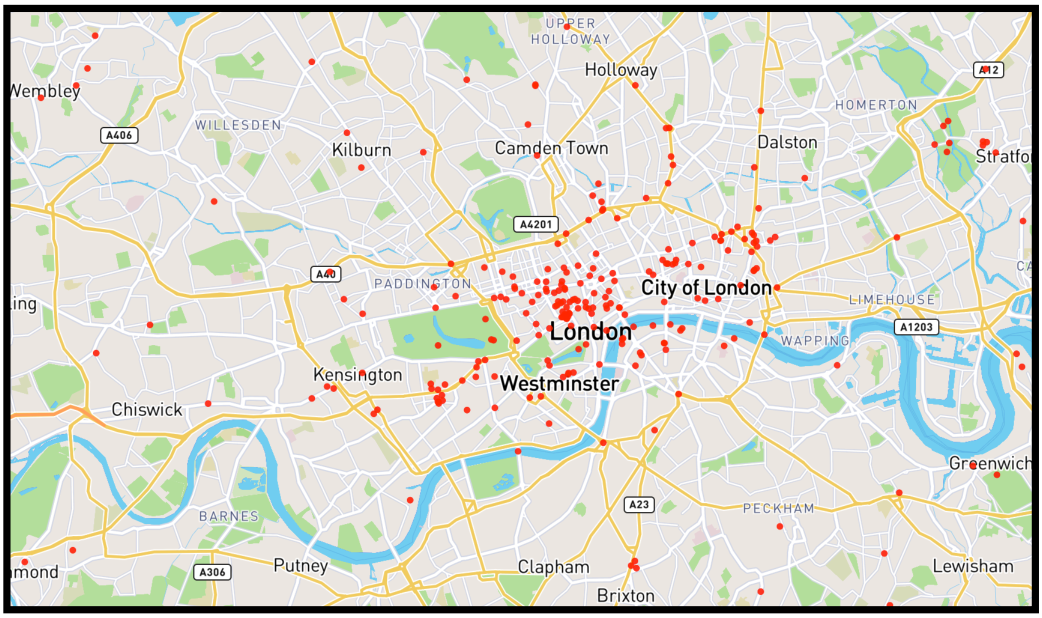



Fig. 3.5 Coordinates of the set of 305 new venues in London considered in the study.

32% under *Travel & Transport*, and 8% under *Nightlife Spots*. These venues were scattered around the city, with their focal point being in center of the city as depicted in Figure 3.5.

**Defining a venue's stationary temporal profile.** For each of the new venues in our set, we first examine the total number of checkins at each time step for each week after the venue opened (ie. the weekly temporal profile). To avoid sparsity issues working at this level of granularity, we create a cumulative temporal profile per week, summing the total number of checkins at each time step with each consecutive week. Although the raw values of cumulative sums have little significance, the trend over the course of the week represents the characteristic curve of the venue and indicates the weekly demand trend. We normalize the curve for each week by dividing by the sum of all checkins for the venue, up to the time of observation. With each consecutive week, we see this curve showing a higher degree of stationarity. We measure the stationarity of this temporal profile over time by calculating the variance of the temporal curve at time $t$ relative to time $t-1$.

Our data suggests that the temporal profile of a new venue becomes stationary when the value of the variance relative to the prior week is $\sigma^2 < 2.6 \times 10^{-5}$. On average, this occurs 5 weeks after a venue has opened. Note that we build the profile of a venue considering a week's temporal span. This captures the most essential temporal patterns of activity at a venue, which includes diurnal variations, but also differences between weekends and weekdays.

## 3.6 Predicting the Temporal Signature of New Venues

Having built an understanding of similarities in the temporal profile of categories and wards, we aim to apply these findings to predict the stationary temporal profile of a new venue. We adopt a k-nearest neighbors approach in which we find the $k$ most temporally similar wards to the ward in which the new venue is located. We look not only at the overall profile of wards as a means of comparison but also at the profile of categories within those wards. The temporal profile of those wards serve as predictors for the new venue and are used to train a Gaussian Process model [139].

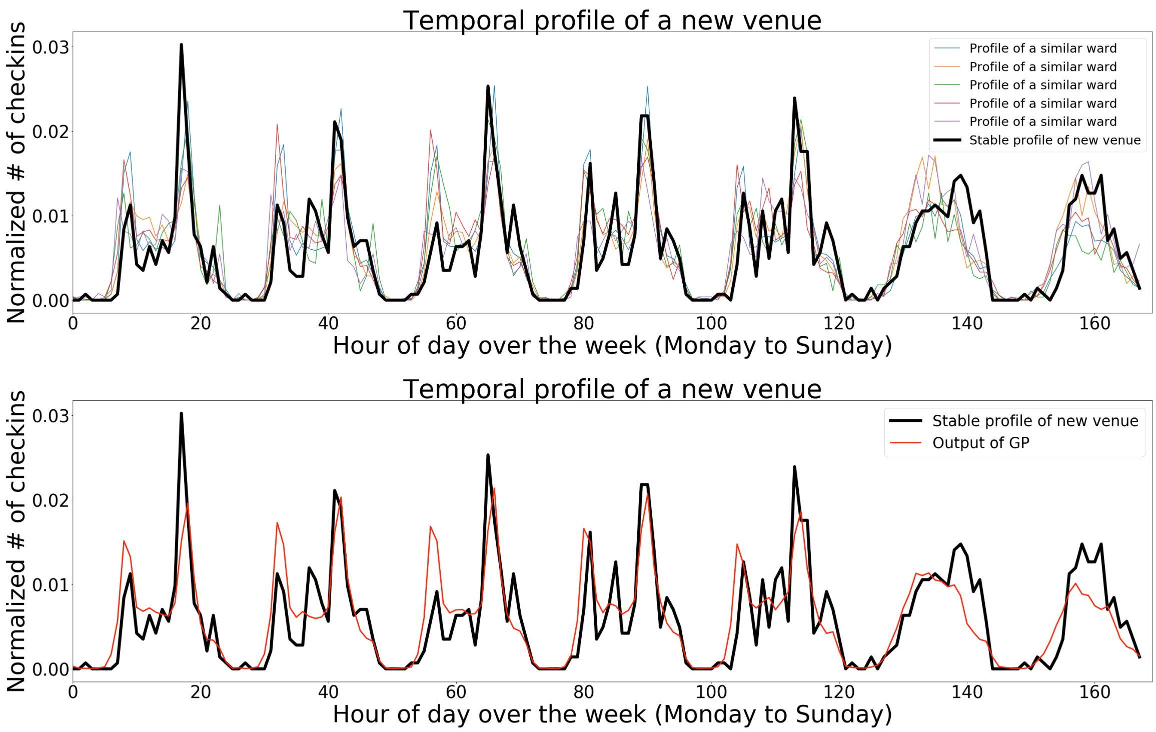



Fig. 3.6 Top panel: the normalized stable temporal profile of the new venue with the profile of similar wards. Bottom panel: the output of the GP trained on the similar ward profiles; this serves as a prediction of the profile of the new venue.

### 3.6.1 Discovering Area-Wide Similarities in Popularity Dynamics

We have seen that similarities in the temporal profiles of wards can be useful indicators of similarities in the characteristics of two wards (Section 3.5.2). We use this idea for our model in which we begin with the basis that two venues of the same category in two different wards are likely to have similar temporal patterns if the overall temporal patterns of their wards are similar. This idea is illustrated in Figure 3.6 which shows the stable temporal profile of a new venue $v_i$ and the temporal profile of the five most similar wards. Further, Figure 3.7 shows the normalized root mean squared error (NRMSE) between the stable profile of a new venue and each of the five similar wards as well as between the profile output from the GP. The figure demonstrates that GP predictors provides a better prediction with respect to simply using the profiles of similar wards.

For a given new venue $v_i$, our methodology to predict its temporal profile is as follows. For clarity, we will describe an example in which we assume $v_i$ is an Italian restaurant called *The Meaning of Life* in ward 42.

1. Determine the general category, specific category, and ward of that venue. For our example, the general category is *Food*, the specific category is *Italian restaurant*, and the ward is 42.

2. Determine the temporal profile of the ward for the general category of interest. In this example, we would determine the overall temporal profile of *Food* venues in ward 42. Formally, we determine $C^{V_{g,w}}[0,T]$ where $T = 168$.

3. Determine the $N$ most similar wards. For all other wards in the city, compare their general category's temporal profile to that of our ward of interest and determine the $N$ most similar wards where similarity is defined as $JSD(C^v, C^w) \quad v \neq w$. This is referred to as the set of *temporally similar wards*. For our example, this would entail finding the $N$ wards whose Food temporal profile is most similar to that of ward 42.

4. Calculate the specific temporal profile for each ward in the set of *temporally similar wards*. For our example, this would mean we would calculate the temporal profile of Italian restaurants for each of the $N$ similar wards.

5. Create a representative curve. These $N$ temporal curves serve as the basis of our prediction of the profile of our new venue $v_i$. To create a representative curve from those $N$ profiles, we use each of the profiles as inputs to a Gaussian Process (GP) because of its ability to recognize latent periodic trends. The output from the GP becomes our temporal prediction.

### 3.6.2 Gaussian Processes Model

Our k-nearest temporal neighbors algorithm finds temporal profiles that are likely to be similar to the venue of interest. We harness Gaussian Processes (GPs) to build a regression model to capture the periodic trends in those profiles. GPs were chosen as the model for this

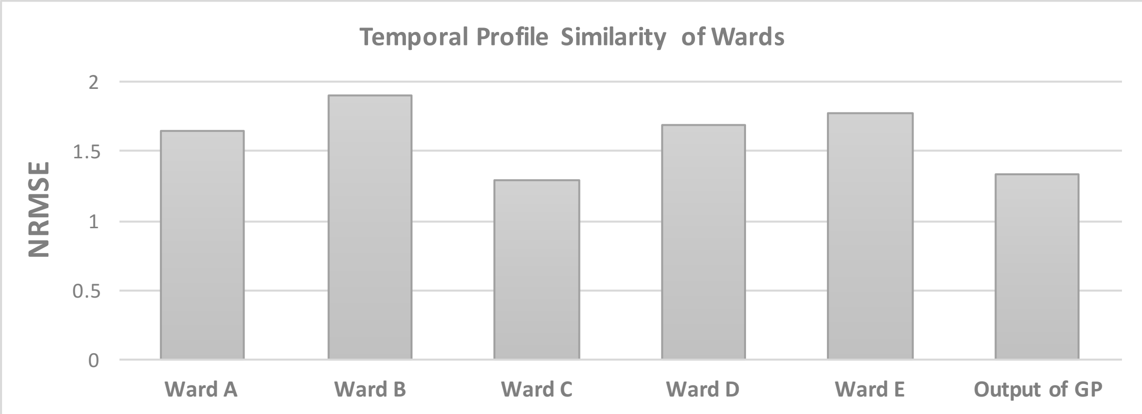



Fig. 3.7 The NRMSE between the stable temporal profile of the new venue and the temporal profile of five similar wards. “Output of GP" is the NRMSE between the output of the trained GP model and the temporal profile of the new venue.

task as they found to be more accurate than other supervised learning models such as decision trees for certain time-series data. GPs are typically employed in time-series tasks because of their flexibility at capturing the complex structures without succumbing to overfitting. GP regression is a Bayesian non-parametric which models a distribution over an infinite set of random variables. A GP model is described by its prior mean and covariance functions. For this work, as is standard, we set the prior mean to zero [139].

**Model Overview:** For the analysis of this we use a product of two Radial Basis function kernels as the base kernel functions which define the covariance matrix of the distribution. We use two kernels to describe two types of periodicity in our data, over the course of a week as well as over the course of a day.

Given the periodicity over the course of a day and a week, we posit that Gaussian Processes are able to recognize latent periodic trends in the data. The inputs to our Gaussian Process are the temporal profiles of the similar wards. We then have the GP predict a temporal pattern for an interval of $[0, T]$ where $T = 168$ for the hourly week's profile. We then compare this prediction to stable temporal profile of the venue of interest.

**Implementation:** The code was written in Python and is available open-source on Github[1]. The analysis of different wards across London was conducted using geopandas, scipy, and numpy libraries [88, 173, 121]. Visualisations of the characteristics of different wards were created using matplotlib and the Django GEOS API [79, 42]. The Gaussian Process model was created using the scikitlearn Gaussian Process library [130].

[1] https://github.com/kritts/urban_modelling_gp

### 3.6.3 Evaluation

In this section, we present an evaluation of our algorithm introducing first a set of baselines as comparators and metrics for comparing the experimental results.

**Baselines.** To evaluate our proposed solution, we compare our results with a number of baseline approaches. For each, the past temporal profiles (i.e., earlier than the venue creation time) are used as features for a GP and the outputs of the GP are the prediction of the characteristic temporal profile of the new venue. The baselines and their descriptions are listed in Table 3.1.

**Metrics.** To analyze the accuracy of our prediction, we calculate the NRMSE between the predicted temporal profile and the stable profile for each venue. We first look at the value of NRMSE as we vary the number of neighbors $N$. Our results show $N = 10$ to be the best indicator of temporal similarity of neighbors. This value was chosen for the subsequent analysis presented in this chapter.

**Results.** Using $N = 10$, we calculate the normalized root mean squared error for the output of each algorithm compared to the actual stable curve of each new venue. Table 3.1 presents a summary of these results. Temporally similar wards using the specific category of the venue proves to be the best predictor of the temporal profile of a new venue.

| Criteria | Description of Criteria | NRMSE |
|---|---|---|
| `TempGen` | Temporally similar wards, same general category | 1.614 |
| `TempSpec` | **Temporally similar wards, same specific category** | **1.575** |
| `Random` | Random wards | 2.692 |
| `SameAll` | Same ward, all categories | 2.1941 |
| `SameGen` | Same ward, same general category | 1.884 |
| `SameSpec` | Same ward, same specific category | 1.760 |
| `AllAll` | All wards, all categories | 1.937 |
| `AllGen` | All wards, same general category | 2.190 |
| `AllSpec` | All wards, same specific category | 2.028 |

Table 3.1 Comparative analysis of different similarity criteria.

## 3.7 On-line Prediction of Mobility Trends at New Venues

As demonstrated in the previous section, predicting the stable temporal curve of a venue is possible and can become more accurate when temporal information from other venues or areas is selectively transferred. Human mobility patterns, despite being characterized by a high degree of regularity, can change over time. Although the characteristic temporal curve indicates the typical temporal trends, the ability to predict demand in real-time can directly help a shop owner to dynamically approximate demand trends, even during the first few months after a venue has opened. In this section, we aim to forecast the popularity of a venue and predict the success of that venue after it has opened. On a day to day basis, we expect the profile of check-ins to vary, however, the variation in demand on a per month basis is less susceptible to noise and a more reliable indicator of the success of a new venue.

### 3.7.1 On-line Prediction Task

When working at the granularity of individual venues in an on-line manner, the sparsity of the data becomes a real concern. In order to address this issue, for each venue, we aggregate check-ins over the course of a month. Predicting a significant increase or decrease in demand for a venue for the following month is useful information for a shop owner to know as it can inform crucial business decisions and who are unlikely to have historic data for the venue. To inform our predictions we utilize locality, the past demand trends of venues in our neighborhood, and temporal similarity, the demand trends of temporally synchronous areas of the city. We next provide a roadmap of how local trends in user mobility can be exploited to forecast alterations of traffic at individual venues in future months.

For each new venue, our prediction task is to estimate the change in demand for each month following the first month of business; then, starting at the first week in which it opened, we aim to predict whether the demand of the venue at its next time step will increase, decrease, or remain stable (i.e., a 3-class prediction task). Remaining stable is defined as remaining within ten percent of the prior value. It is worth noting that the proposed methodology does not depend on the choice of this interval. With each subsequent month since a venue has opened, we have more data with which we can better understand the venue.

We inform our predictions by selectively using temporal information from other areas of the city. A number of different criteria were used when selecting these inputs, which are listed in Table 3.2. For example, when using the “history" of a venue, for each month after the venue opens, we train a GP on the demand for each month. We then predict the relative demand at each progressive month, continuously learning from the previous months (i.e., month 1 to 3 would be used in the training dataset when predicting month 4). This real-time

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| `History` | 0.6748 | 0.6697 | 0.6853 | 0.7286 | 0.7278 |
| `TempGen` | 0.7507 | 0.7820 | 0.7691 | 0.7824 | 0.7903 |
| `TempSpec` | 0.7729 | 0.7804 | 0.7829 | 0.7991 | 0.8104 |
| `Random` | 0.5102 | 0.5185 | 0.5248 | 0.5682 | 0.5993 |
| `SameAll` | 0.7149 | 0.7310 | 0.7382 | 0.7349 | 0.7352 |
| `SameGen` | 0.7403 | 0.7481 | 0.7592 | 0.7480 | 0.7791 |
| `SameSpec` | **0.7859** | **0.7915** | **0.7981** | **0.8216** | **0.8221** |
| `AllAll` | 0.6812 | 0.6892 | 0.6489 | 0.6893 | 0.6832 |
| `AllGen` | 0.6824 | 0.6853 | 0.6935 | 0.7088 | 0.7129 |
| `AllSpec` | 0.7201 | 0.7209 | 0.7403 | 0.7459 | 0.7402 |

Table 3.2 AUC values of the real-time prediction with a varying number of months of training data.

prediction methodology harnesses the Bayesian nature of GPs and their ability to predict and react to anomalies in the data (i.e., if there is a sharp peak at a given time step $t$ the GP uses that insight when predicting $t+1$). This mirrors situations in which the demand of a venue can sharply increase because of events on a given day.

### 3.7.2 Evaluation of the On-line Prediction Task

We predict the relative demand curve in real-time for one week for each of our 305 new venues. Table 3.2 presents the AUC value when predicting the relative demand curve after the training phase.

Our baseline is to train the model on the history of the new venue after the first month from its opening. We examine the use of temporally similar wards as a predictor. We use as our prediction the history of 10 wards because this value provided the optimal representation of similarity. We also examined as inputs venues in wards that have the same general and specific category as the venue of interest. Our results show that locality does have an impact on predictions as the predictions using venues in the same ward are higher than using venues in all wards throughout the city. Further, we see that temporal similarity can also be used to improve predictions; this may be because similar venues could have synchronous peaks in demand following a similar impact from real-world events.

## 3.8 Discussion

The evaluation results discussed in the previous section have shed new light on the temporal dynamics of user activity in location-based services.

At neighborhood level, we have seen that areas that are far from each other can be synchronized with regards to their temporal activities. Moreover, the temporal frequencies of such activities tend to be stationary over certain periods of time due to regularities in human mobility patterns. We exploited this information to predict the temporal popularity profiles of newly established venues in Section 3.6, essentially transferring information from the level of an urban region to that of a specific venue. This form of analytics can provide new insights to new business owners who can plan supplies and staffing in their facilities during the cold start period of a new opening. Beyond retail venues, the idea can be expanded to other types of places, such as parks or outdoor spaces. Predicting how urban spaces are used over time can improve planning, including the design of schedules for their maintenance or police them. Despite the regularity patterns observed in human mobility, variations over time will exist due to social events or unforeseen circumstances such as travel disruptions. These can result in historically unexpected increases or decreases in mobility flows towards a venue.

To examine whether such variations can be captured on a venue level we experimented with an on-line prediction task, where the goal was to predict relative changes with respect to historic patterns of a venue. This is a challenging task from the point of view of data sparsity. Even very popular venues in location-based services will have only a handful of check-ins observed in a small time window. We have demonstrated that it is possible to pick up trends in this setting using a Gaussian Process model trained on data inputs from recent mobile user activity at nearby venues.

## 3.9 Conclusion

In this chapter, we have investigated the prediction of the temporal dynamics of newly established venues using the check-in data of millions of Foursquare users. We have also introduced the concept of *temporally similar areas* in a city, areas that share patterns in the movement of people to different types of venues within those areas.

We have shown that the characteristic temporal curve of a new venue provides valuable insight for new shop owners who can use that information to better inform supply purchases, opening hours, and demand. Characteristic curves can also support the design of location-based technologies. Additionally, our models help to demonstrate how a particular venue influences the overall temporal profile of the neighborhood it is located in. This knowledge can help design more interpretable models and build urban applications that are aware of the behavioral choices made by citizens on a local level, rather than those that treat population dynamics as a blind optimization task.

The following chapter further explores the use of Foursquare data to forecast venue dynamics by modelling the likelihood of closure using spatial, network, and venue-specific metrics.

# Chapter 4

# The Role of Urban Mobility in Retail Business Survival

In the previous chapter, we modeled the demand prediction of new venues through temporal similarity metrics and Gaussian Process models. In this chapter, we further explore machine learning models to predict venue dynamics by forecasting the potential failure of retail establishments. We examine the impact of our predictions on varying types of venues and explore the predictive potential of different features.

The main contribution of this chapter is a model capable of predicting business survival with high accuracy, achieving approximately 80% precision and recall across the cities. We show that the impact of different features varies across new and established venues and across cities. Besides achieving a significant improvement over past work on business vitality prediction, this chapter demonstrates the vital role that mobility dynamics play in the economic evolution of a city.

## 4.1 Introduction

There is a strong economic and policy interest in both uncovering the causes of failure of retail businesses and in predicting their likelihood. Broadly, an establishment's failure susceptibility can be ascribed to a variety of *controllable* and *uncontrollable* factors. Controllable factors could include the quality or price of the store's product offerings, its operating hours, and its customer satisfaction. Conversely, uncontrollable factors could include unemployment rates of the city, overall economic conditions, and urban policies. Establishing what constitutes *failure* is a challenge in itself and has had a critical role in limiting the number and extent of existing studies on business survival [129, 175]. Prior works have utilized financial records where they consider *bankruptcy* as a failure. However, this approach is limiting as it does not capture cases where a proprietor decides to shut down an establishment. Despite such efforts, the inherent low frequency of financial reporting lends itself to (1) studies that focus on static macro factors leading to failure and (2) a failure in recognizing establishments that are at high risk of mortality in the near future.

The recent proliferation of urban datasets, especially related to urban mobility and social media activity, offers interesting opportunities for high-fidelity sampling of these controllable factors. For example, mobility data can reveal the *urban dynamics* of different locations (e.g., does a neighborhood attract visitors from various other neighborhoods?), whereas location-based social network (LBSN) data can elucidate *consumer interactions* at the individual venue level (e.g., how popular is the venue relative to others in its vicinity?). Researchers have explored the use of social media for business analytics–e.g., Wang et al. [175] utilized LBSN data to predict the failure of food establishments using a set of over 600 restaurants in New York City (NYC) over a 6 month period, and Karamshuk et al. [91] provided empirical strategies for using LBSN-based features to find optimal locations for new stores.

In this work, we utilize two complementary, large-scale longitudinal datasets: (1) venue *check-ins* on Foursquare, observed in ten cities across the globe, and (2) *taxi trip records*, observed across Singapore and New York City, to develop a predictive model for retail business failures. We examine the role of a number of features on retail business survival, across both a broader swathe of retail categories and specifically for food & beverage (F&B) establishments. As F&B is known to be a highly competitive and risky business in many cities, we examine this category more closely for universal trends. We employ three classes of features: (a) *Static Locality Profiles*, capturing the properties of the locality in which an outlet operates; (b) *Visit Patterns*, reflected in the volume and spatiotemporal patterns of Foursquare check-ins; and (c) *Neighbourhood Mobility Dynamics*, reflected in visitation patterns across distinct neighborhoods. Our specific prediction question is: *given observable features at a point in time, how likely is it that a retail establishment will close down within the next 6 months?*[1]

**Key Research Questions and Contributions:** Our investigations require us to tackle and answer three key questions, enumerated in the sequence they are addressed here:

- *What are some of the key factors that explain business survival?* To address this question, we identify and analyze a range of features and perform a comprehensive study to demonstrate the predictability of survival or failure of an F&B business in

[1]The 6-month duration can, of course, be varied: for now, we choose 6-month as it appears to be a natural time constant for retail businesses deciding whether to close down or not and also because determining an establishment's operating state at finer timescales from Foursquare data is very noisy.

the subsequent 6-month period, in two metropolitan cities, New York and Singapore. Overall, we achieve AUCs of 0.85 and 0.90, for Singapore and New York, respectively, with corresponding precision/recalls at $\approx 80\%$ –this represents an almost 10-15% improvement in accuracy, over similar past work [175]. We also found that the most important factor was the ability of an establishment to draw customers around the clock and not just during specific hours.

- *How generalizable is business survival predictability?* We answer this question in three parts, we: (1) extend our analysis to the broader category of *retail* venues, (2) repeat our analyses on multiple cities across the world, and (3) investigate differences between newly founded and established venues. Our results show consistent performance and also show that certain features have consistently high power (AUCs ranging from 0.82 to 0.84) for predicting survival likelihood, despite geographic differences among cities. Our results show that, across all ten cities considered and using the same classifier, established venues had an area under the curve (AUC) of 0.86 while newer venues had a somewhat lower AUC of 0.81. These results suggest that new venues may have more variability in their underlying causes of failure.

- *How robust are our results?* We perform a series of experiments to validate the robustness of our results. We study the collinearity across the set of features considered and demonstrate with a reduced model consisting of only a subset of features that the accuracy drops only by a few points (e.g., 3% for New York City, from 0.92 to 0.89). Additionally, our results show that for cities with a high volume of check-in data, the prediction accuracy is relatively unaffected even under shorter observation periods.

Overall, we provide compelling new evidence of the power of combining venue-specific, location-related and mobility-based features in predicting the likely demise of retail/F&B establishments across different cities, despite lacking visibility over other factors (e.g., management quality, reviews, and broader economic trends) that plausibly influence such business outcomes. Given the many other un-observable factors (e.g., management quality, reviews, and broader economic trends) that plausibly influence such business outcomes, it is remarkable that our Foursquare and transport based features can act as high-quality, observable proxies for such outcomes.

## 4.2 Related Work

We next explore past works related to the research in this chapter. We categorize the related work along several dimensions.

**Studies on Business Survival and Mortality:** Between 2004 and 2014, Parsa et al. [129, 128, 126, 127], published a four-part series on *Why Restaurants Fail*. In [129], based on a quantitative study of 2400+ restaurants in Columbus, they present a framework for survival, composed of four main areas: environmental factors, family lifecycle, internal factors, and growth stage of the restaurant. In this chapter, we focus on the environmental factors (driven by the retail establishment's location) and additionally include venue-specific visitor dynamics to build a quantitative predictor of an establishment's 6-month failure likelihood.

**LBSN-driven Business Analysis:** The analysis presented by Lei Wang et al. in [175] is closest in spirit to our work. In [175], the authors explore the use of LBSNs in predicting the survival/failure of food establishments using a set of 600+ restaurants in NYC over a 6 month period. The authors study the improvement in prediction offered by six key features (number of competitors within a 3-mile radius, competitors with specials, price range, rating and average daily check-ins of the restaurant and its competitors, and the trend of the average daily check-in growth) respectively. Among other findings, their analyses show that: (i) the competitive analysis matters: considering the check-in information of a restaurant and its neighbors reduces the misclassification rate from 30% to around 10%, and (ii) such competitive effects dissipate beyond a radius of 1 km. While the authors report an overall misclassification rate of 10%, the high imbalance in the underlying data set (only 10% of the FB establishments are classified as failed) implies that the precision and recall of the failed class is only 72% and 66%, respectively. In this chapter, we not only perform a more comprehensive study (10 cities, on multiple categories of venues), and also study the factors which govern business failure at a global level, incorporating several new key features, especially those related to a neighborhood's mobility dynamics. Our work also utilizes a variety of datasets and explores city-specific feature variations of the model.

A variety of works have explored facets of LBSN-based prediction of business demand. In [91], the authors provide empirical evidence on how LBSNs can help find optimal store placements based on a multitude of geographic (e.g., neighborhood, competitiveness, attractiveness of a neighborhood to a particular store category) and mobility features (e.g., inter- and intra-neighborhood transitions). They assume that the optimal placement for a venue is the location which can draw the most amount of check-ins or footfall. Similarly, in [107], the authors develop *ZoneRec*, a framework for recommending zones for a new F&B venue. In this framework, a new venue's features are computed using TF-IDF scores on the venue label, and then matched to zones with similar features–in effect, ZoneRec seeks to mimic the current zonal distribution of FB venues and does not incorporate competition-based features (which may indicate that a new venue should choose a location where that type of venue is currently under-represented). More recently, in [37], the authors used changes in

Foursquare check-in volumes, before and after the opening of a new venue, to study whether certain categories of businesses *cooperate* or *compete* with other close-by businesses in the same category. Finally, in [45], the authors used Foursquare check-ins to investigate whether the stable temporal demand of a new venue (that presently has no historical trends of check-in data) can be predicted from past visitation patterns of existing venues of the same category that are situated in the same ward, and/or wards that are found to be *temporally similar*.

**Urban Dynamics using Social Media:** In this work, we saw that the mobility dynamics of individual locations, at the neighborhood level, turned out to be a key feature in our failure prediction model. Recent work has used social media data to investigate various aspects of such urban dynamics. In [78], the authors carry out a study on gentrification in the wards of London, and more recently in [191], the authors study the impact of cultural investment (e.g., new stadiums/museums) on the businesses and venues in proximity to those new investments. Further, in works such as [61], the authors present the case study of the London Olympics which was held in 2014 where they provide empirical evidence of how local retailers benefited by the increase in footfall to the event-related areas.

**Fusing Social Media and Physical Sensor Data for Urban Analytics:** Our predictive model utilizes features that are derived from a combination of social media and physical mobility (from taxi and bus traces) data. In [161], the authors tackled the problem of unifying multiple streams/modalities of Web-based sensory data by introducing the concept of 'social pixels', which aggregates user interest across multiple channels at a particular geo-location. In [113], the authors introduced a vision for *socio-physical analytics*, and outlined the challenges and opportunities in fusing these disparate data sources, from the social Web and physical sensor streams. As an example of such analytics, in [66], traffic anomalies detected using physical traffic sensors are matched spatiotemporally with anomalies detected on social media (i.e., unusual volumes of keywords) to clarify the cause or source of the anomaly. In [52], the authors applied the concept of socio-physical analytics to wellness profiling, combining data from social media posts (Twitter tweets, Instagram pictures, and Foursquare check-ins) and wearable sensor data (captured by Endomodo, an exercise tracking App) to infer the BMI (Body Mass Index) profile of users. Finally, in [82], the authors demonstrate the fusion of data from social media feeds (Twitter) and physical urban sensors (busses crowdedness and traffic cameras) to improve the spatiotemporal localization of urban events.

## 4.3 Our Approach at a Glance

The motivation for this chapter is to build a predictive model for venue closure. To this end, we first identify a set of candidate features (Section 4.5), pose the problem of predicting closure as a binary classification task and report our findings in Section 4.6. In this section, we formalize the key questions we answer in this work, introduce notations used throughout the work, and define what constitutes *closure*.

We hypothesize that different classes of external factors including the locality a venue operates in, the competitive forces that it faces, and the visit patterns of its customers, play a vital role in its survival. Using a combination of LBSN and transport data, we seek answers for the following questions:

1. Can metrics of the locality profile, visitation patterns, and mobility dynamics of a retail business be used as predictors its success or failure?
2. Do factors that attribute to business failure vary by city, or geographies?
3. Is failure similar for new and established businesses?

### 4.3.1 Notation

We consider the set $\mathbf{V}$ of venues in a city. A venue $v_i \in \mathbf{V}$ is represented with a tuple $< loc, date, gen, spec >$ where *loc* is the geographic location of the venue, *date* is its creation date, *gen* is its general category, and *spec* is its specific category. Further, we define a venue's *neighbourhood*, $\mathbf{N_i}$, as the set of venues that are located within a given radius; we set this distance to 500m as prior work [37] has shown that a venue's operation is affected primarily by conditions within this radial distance. Formally, we define the neighbourhood as:

$$\mathbf{N_i} = \{v_j \in \mathbf{V} : dist(v_i, v_j) < D\} \tag{4.1}$$

where $dist(v_i, v_j)$ represents the distance between $v_i$ and $v_j$ and $D = 500$. We also define a venue's *competitive neighbourhood*, $\mathbf{CN_i} \subseteq \mathbf{N_i}$, as the subset of venues that belong to the same general category, *gen*, as the venue. Similarly, we define the specific competitive neighbourhood, $\mathbf{SN_i} \subseteq \mathbf{CN_i}$, as the subset of venues that share the same specific category, *spec*, within the radius $D$. Further, we define *established* venues as those that have existed for longer than a year and *new* venues as those that have existed for less than one year. Later in Section 4.6.5, we use this distinction to examine the impact of a venue's age on prediction accuracy.

The administrative zone (e.g., Census tract, subzone, ward, etc.) a venue belongs to is referred to as the venue's *locality*, throughout this work. Acronyms used throughout the text are listed in Table 4.1.

### 4.3.2 Defining Closure

For this work, we classify venues as either opened or closed. Prior research on Foursquare data has shown that venues added after June 2011 were highly likely (probability above 0.8) to actually be new venues opening rather than existing venues being added to the system for the first time [37]. To uncover venues that are at risk of closure, we look at check-in metrics between June 2011 - December 2013. For a given month, we define $C_t(v_i)$ as the total number of check-ins to venue $v_i$ in that month. Similar to prior work (Wang et al. [175]), we consider a significant decline in check-in volume as a sign of impending failure. We define a venue, $v_i$, as closed when $RemainsOpen(v_i) = 0$. The formal definition is as follows:

$$RemainsOpen(v_i) = \begin{cases} 0, & \text{if } \frac{\sum_{t=0}^{T} C_t(v_i)}{T+1} < K \times mean(v_i) \wedge \frac{\sum_{t=0}^{T} C_t(v_i)}{T+1} < N \\ 1, & \text{otherwise} \end{cases} \tag{4.2}$$

where $mean(v_i)$ represents the mean number of check-ins for venue $v_i$ prior to June 2013 (starting from its first presence in Foursquare), $T = 5$ represents the 6 month window, and $N = 1$, denoting an average of less than one check-in per month for the venue $v_i$. In other words, we define a venue to be closed if it has less than an average of one check-in per month, for 6 months, and if this average represents a significant decline in demand for this venue. We examine the validity of this definition of closure in Section 4.4.2 where we show it is consistent with ground truth. We experimentally determined the value for $K$ by varying this scaling factor incrementally from 0.15 to 0.30 which resulted in minimal variations in the percentage of closed labels. For London, the percentage of new venues that closed was 6.1% when $K = 0.15$ and 6.7% when $K = 0.30$. Across all ten cities, our closed labels varied marginally by an average of $< 7\%$. As such, we experimentally converged to $K = 0.25$ as this was the optimal value for our analysis.

### 4.3.3 Operationalising the Venue Survival Problem

As previously described, we focus on predicting whether a venue is likely to survive the next six months. We define the "prediction date" (PD) as the fixed date of July 1, 2013 across all venues. We refrain from deciding on a prediction date per venue (based on factors such as age or actual date of failure) as seen in prior work [11] due to practical limitations in

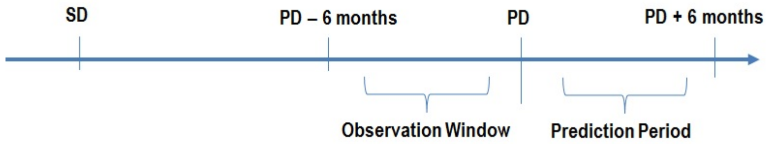



Fig. 4.1 Definition of virtual past and future data used in this work. We use a fixed Prediction Date (PD) across all venues and answer the question, *which of these venues will close during the prediction period* $(PD, PD+6]$*?* using features computed over $(PD-6, PD]$.

acquiring such information (see Section 4.4.1). Using the first Foursquare check-in as a proxy for activity, we define a starting date (SD) for each venue. For each venue, the period $(SD, PD]$, the time from the venue's opening until the prediction date, is considered as the past data and the period $(PD, PD+6months]$ (July'13 to Dec'13) is considered the *virtual* future data. This 6 month period is used in Equation 4.2 to label whether a venue has closed. We depict this in Figure 4.1. The features used in this work (described later in Section 4.5) are all based on data pertaining to an observation window which is uniform across all venues regardless of their SD. In this work, we consider an observation window of 6 months which immediately precedes the prediction date, i.e., $(PD-6months, PD]$ (Jan'13 to June'13), and also investigate the sensitivity of our results to shorter observation periods. All venues considered in this work were operational by the commencement of the observation window.

| Acronym | Detail | Acronym | Detail |
|---|---|---|---|
| F&B | Food and Beverage | SD | Starting Date |
| LBSN | Location based Social Network | PD | Prediction Date |
| AUC | Area under the (ROC) curve | ReLU | Rectified Linear Units |
| ROC | Receiver Operating Characteristic | SELU | Scaled Exponential Linear Units |
| FS | Foursquare | CBD | Central Business District |

Table 4.1 Acronyms used throughout the chapter.

## 4.4 Mobiltiy Datasets

### 4.4.1 Dataset Description

We make use of two types of data for this work: one sourced from Foursquare on a multitude of cities across the world, and the other obtained from transportation authorities in two major cities.

| City | Check-ins | Established Venues | New Venues | % Established, Closed | % New, Closed |
|---|---|---|---|---|---|
| Chicago | 10,600,106 | 8,726 | 556 | 7.3 | 6.5 |
| Helsinki | 4,400,044 | 3,359 | 272 | 5.0 | 5.5 |
| Jakarta | 5,200,052 | 7,135 | 540 | 12.6 | 3.3 |
| London | 4,000,040 | 6,633 | 399 | 2.8 | 6.5 |
| Los Angeles | 3,300,033 | 5,652 | 263 | 6.1 | 2.7 |
| New York | 13,700,137 | 14,733 | 1048 | 8.5 | 7.4 |
| Paris | 3,600,036 | 4,653 | 189 | 5.1 | 6.3 |
| San Francisco | 4,100,041 | 5,407 | 336 | 5.4 | 6.0 |
| Singapore | 12,800,128 | 14,193 | 552 | 23.7 | 3.4 |
| Tokyo | 12,600,126 | 12,385 | 551 | 4.4 | 2.0 |

Table 4.2 Summary of city statistics. For each city, we report the total number of transitions, the number of established venues, the number of new venues, the percentage of established venues that closed, and the percentage of new venues that closed. Venues defined as *new* and *established* had been open for less or more than one year respectively (described in Section 4.3.1). Venue closure was defined using Equation 4.2 (i.e. $RemainsOpen(v_i) = 0$).

**Foursquare Data**: As described in detail above in Section 3.4, the application Foursquare enables users to check in to different locations and share that information with their friend group. In this work, we use the same longitudinal Foursquare dataset as Chapter 3 across ten cities around the world. The data spans three years and over 75 million check-ins. Table 4.2 includes the summary of statistics of the 10 cities we consider in this work–for the sensitivity analysis in Section 4.6.5, we also enumerate the count of new versus established venues. In Figure 4.2 we provide a visualization of the spatial distribution of venues in two cities, New York City and Singapore, that we label as closed and open according to Eq. 4.2 and the timeline described in Section 4.3.3.

**Transport Data**: We rely on two transport datasets, one from New York City and the other from the city-state Singapore, that help us in understanding the movement dynamics and local catchment of *localities* within the city.

From New York City (limited to the Manhattan Borough), we obtain time-stamped records of dropoffs and pickups by yellow taxis for the period of January 2013 - December 2013 (which overlaps partially with the check-ins dataset), made available publicly by the New York City Taxi and Limousine Commission[2]. Each record contains the GPS coordinates

[2]http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

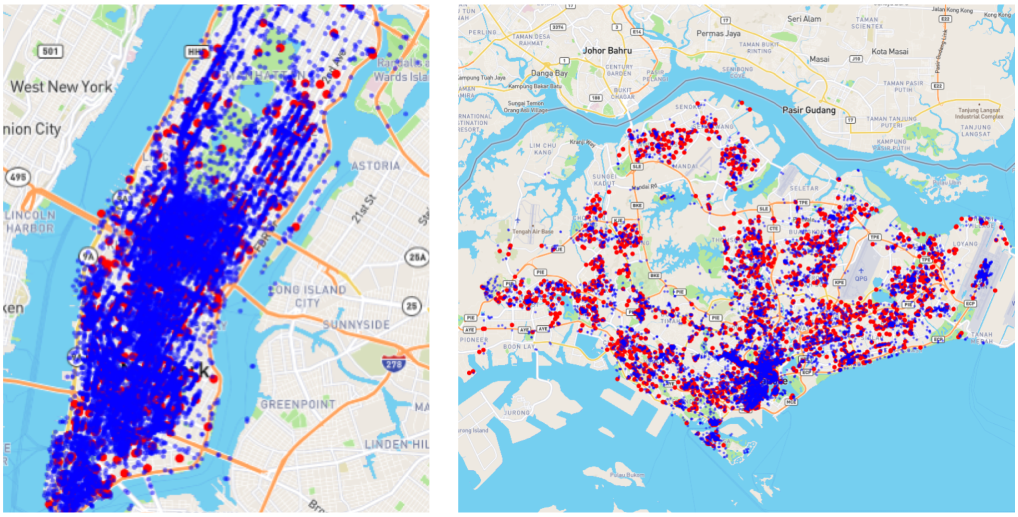



Fig. 4.2 Spatial distribution of venues in New York City (left) and Singapore (right). Blue represents “open" and red represents “closed", as defined by Eq. 4.2.

of the pickup and dropoff points and the corresponding timestamps. We aggregate the pickup and dropoff points to Census tracts [3], where a tract typically houses at most 16,000 residents [4]. In the case of Singapore, we use data from a major taxicab company consisting of all trips occurring between November 2011 through January 2012 whose pickup and dropoff points we map to subzones [5] which are administrative boundaries. Table 4.3 summarises key statistics of these datasets. Each record is of the format $< p_x, p_y, t_p, d_x, d_y, t_d >$ where $(p_x, p_y)$ and $(d_x, d_y)$ are the GPS coordinates of the pickup and dropoff of a single taxi trip, respectively, and $t_p$ and $t_d$ are the corresponding timestamps.

| City | Spatial Aggregation | Total Trips | Observation Period |
|---|---|---|---|
| NYC | 288 Census Tracts | 143 million | Jan 2013 - Dec 2013 |
| SG | 323 subzones | 38 million | Nov 2011 - Jan 2012 |

Table 4.3 Summary of taxi datasets used in the analysis.

### 4.4.2 Venue Closure

As mentioned previously, it is empirically challenging to get the ‘ground truth’ of the closure of retail establishments across cities. In contrast to business openings, which are often advertised and announced on social media, venue closings often happen without fanfare.

[3] http://maps.nyc.gov/census/
[4] https://data.cityofnewyork.us/City-Government/2010-NYC-Population-by-Census-Tracts/si4q-zuzm
[5] https://data.gov.sg/dataset/master-plan-2014-subzone-boundary-web

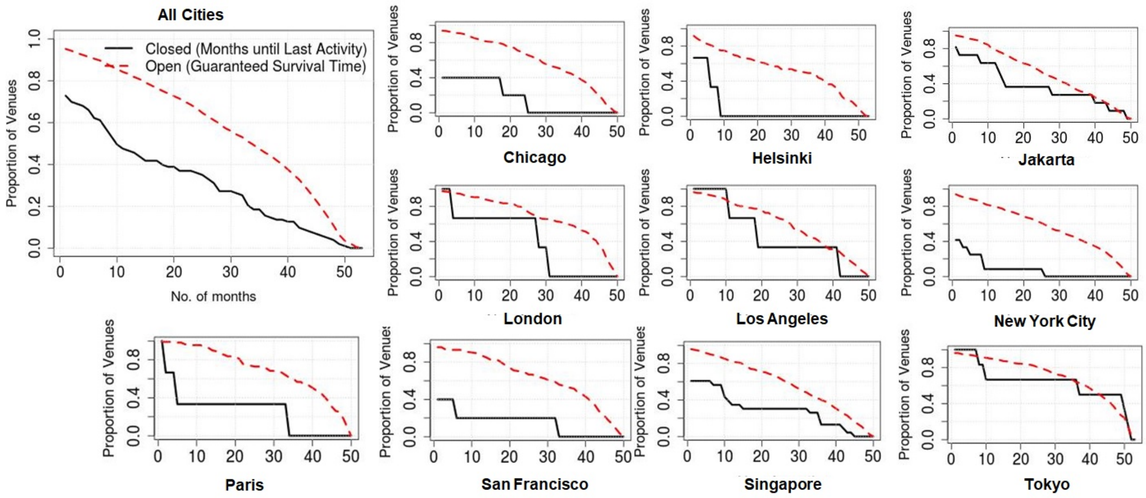



Fig. 4.3 The survival curves (as KM plots) for all F&B venues considered in this work.

Moreover, F&B establishments sometimes exhibit “virtual closure”-a specific venue can simply re-brand itself (e.g., from a coffee shop to a restaurant/lounge), without actually changing owners.

Given such phenomena and due to the fact that the check-ins dataset exists only until the end of 2013, we retrieve additional venue related data using the public Foursquare Venue API [56] to understand how credible the closure labels that we define are. This additional data retrieved consists of longitudinal observations of time-stamped, publicly shared activities at each of these F&B venues as at the query date (i.e., May 2018); these activities consisted of *tips*, public notes or short reviews users can share about a venue, and photo posts from visitors. This allows us to observe the activity at a venue beyond our prediction window and *proxy* the lifetime of the venues considered. We provide the survival plots of the 10 cities considered in this work in Figure 4.3 for F&B venues that started operation after June 2011. We use the Kaplan-Meier estimator [90] to estimate the survival function where the time to event is the time for which a venue remained active (based on tips and photos) beyond the prediction window (i.e, Dec’13). The $x-$axis represents the 53 month-timeline between Jan-14 till May’18 where the black line represents the “time till last activity or *closure*" for venues labeled as “closed" and the red dashed line represents the “guaranteed time alive or *open*" for venues labeled as “open". We see that less than $\approx$10% of restaurants (across NYC, Singapore, and London), supposedly open during June-Dec’13, may be mislabelled, as they see no activity over the next 4 years. Conversely, approx. 50% (Singapore), 70% (NYC), and 65% (London) of restaurants, supposedly closed by Jan’14, cease all activity in the subsequent 2 years. On the other hand, approx. 60% of restaurants in Singapore that we

| **Feature Class** | **Feature** | **Definition** | **Source** |
|---|---|---|---|
| *Locality Profile* | Competition | $\frac{CN_i}{N_i}$ | Foursquare |
| | Specific Competition | $\frac{SN_i}{CN_i}$ | Foursquare |
| | Place Entropy | $\sum_{i=1}^{k} p_i * \ln p_i / \ln k$ | Foursquare |
| | Category Counts | $\|CN_i\|, \quad \|SN_i\|$ | Foursquare |
| | Attractiveness to the Neighbourhood | $\|CN_i\| \times \ln\left(\frac{\|V\|}{\|V_C\|}\right)$ | Foursquare |
| | Catchment of Locality | $\frac{\|D_l\|}{\|D\|}$ | Transport |
| | Temporal Catchment of Locality | $\frac{\|D_{w,l}\|}{\|D_w\|}$ | Transport |
| *Customer Visit Patterns* | Inflow & Outflow | $\frac{\sum_{j=0}^{\|V\|} t(v_j,vi)}{M}$, $\frac{\sum_{j=0}^{\|V\|} t(v_i,vj)}{M}$ | Foursquare |
| | Distance Travelled to Reach Venue | $\frac{\sum_{j=0}^{N} dist(v_j,v_i)}{N}$ | Foursquare |
| | Speed of Travel to Venue | $\frac{\sum_{j=0}^{N} dist(v_j,v_i) \times t_{i,j}^{-1}}{N}$ | Foursquare |
| | Temporal Popularity Skew | $\sum_{i=1}^{24} h_i * \ln h_i / \ln 24$ | Foursquare |
| | Visit Trend | $\frac{c_t(v_i)-b}{t}$ | Foursquare |
| | Temporal Alignment with Competitors | $\sum_{j=1}^{24} (h_i(j) - H_i(j))^2$ | Foursquare |
| *Mobility Dynamics* | Temporal Alignment with Locality | $\sum_{j=1}^{24} (h_i(j) - h_l(j))^2$ | Both |
| | Reachability | $r_{(a,b)}$ | Both |
| | Distance-weighted Reachability | $dr_{(a,b)}$ | Both |
| *Business Attributes* | Cuisine Type | Categoric variables | Foursquare |
| | Price Tier | | Foursquare |

Table 4.4 Summary of Features Investigated in this Work.

labeled as “open" remained active beyond 2 years since the prediction period. This analysis lends credence to the reliability of our “failure" labeling process (i.e., the use of Equation 4.2), but also illustrates the challenge of perfect labeling.

## 4.5 Feature Description

In this section, we describe three classes of features that may play a role in the success of a business. The first set of features help profile the neighborhood (or region) where a particular retail outlet is situated; as some of these features are derived from urban mobility data, they are computed (during our enumeration of results in Section 4.6.2) only for Singapore and NYC. The second set of features are based purely on Foursquare check-in and capture various aspects of visitor dynamics to a specific venue; in Section 4.6.3, we shall evaluate the predictive power of such solely venue-specific features for all ten cities. Finally, the third set of features are based on mobility dynamics (derived from taxi data, and thus evaluated only for NYC and Singapore), and thus capture the mobility-driven characteristics of a venue and its surrounding region.

### 4.5.1 Profile of the Locality

Prior studies on restaurant failure in [129, 128, 126, 127, 175] and the more recent analysis in [175], indicate that a F&B venue's *locality* plays an important role in determining its success. In this work, locality refers to the demarcation of a city into administrative zones–i.e., Census Tracts in the case of NYC, and Subzones in Singapore. More specifically, the locality's centrality affects the volume and visit timings of potential patrons to a specific venue. For example, a cafe that offers breakfast near train stations or bus stops in residential areas, or places that offer quick and easy lunch options in the central business district (CBD) area, may benefit from such a time-dependent influx of residents or visitors, more so than other areas. Similarly, venues in the CBD areas will benefit greatly from the sheer volume of people that commute for work. Furthermore, a venue's neighborhood will also have other venues that have similar offerings (i.e., competitors)–this may result in cooperative or competitive effects affecting the venue's available catchment (the pool of potential visitors). In addition, the variety of venues in a neighborhood may also allure or repel customers–e.g., many cities have ethnic enclaves with a congregation of ethnicity-specific retail outlets. We capture such intrinsic, largely *static*, properties of the locality using the following set of features, computed for each venue.

**Competition**: We define this as the proportion of competitors (in the case of FB, this refers to all 'food' establishments in Foursquare) to the size of the neighbourhood, $\frac{|CN_i|}{|N_i|}$.

**Specific Competition**: Similar to the above definition, we define the competition at a category-specific level. For example, in the F&B case, this is defined as the number of neighbouring venues that serve the same cuisine as venue $i$ (denoted by $SN_i$), normalised by the total number of competing food venues–i.e., $\frac{|SN_i|}{|CN_i|}$.

**Place Entropy**: We exploit semantics information about venues in the city to characterize a venue's neighborhood in terms of diversity of activities available there. The diversity of the area around the venue is measured through the Shannon equitability index [155] from information theory. This metric is calculated as follows:

$$-\sum_{c=1}^{k} p_c * \ln p_c / \ln k \tag{4.3}$$

where $p_c$ denotes the proportion of venues of category $c$ and $k$ is the total number of different categories in $N_i$ within 500 meters of the venue.

**General/Specific category count**: In addition to taking into account the relative proportions of venue categories in the area, we also consider the absolute counts of these venues, as a proxy for the overall size of the neighborhood. Formally, for the FB case, we count the number of food venues, $|CN_i|$, and the number of venues in each specific category (i.e., cuisine) in its neighbourhood, $|SN_i|$.

**Attractiveness of the neighborhood:** We measure this feature at both the general and specific category levels. We borrow the use of the $tf-idf$ weighting scheme from text mining literature, adopting the notion of neighborhoods as *documents* and the venue categories as the *terms* that occur in them. The term frequency $tf$ is simply the count feature defined above, and the Inverse Document Frequency, $idf$ is computed as follows: $\ln\left(\frac{|V|}{|V_c|}\right)$ where $V_c \subset V$ is the set containing all venues belonging to that same category. Then, the attractiveness score of a venue to its neighborhood, $N_i$, is given by, $tf \times idf$. The features defined above capture aspects of the immediate neighborhood (i.e., within a 500m radius) of a given venue. We also define additional features that characterize the $locality$ of the venue.

**Catchment of the Locality:** To capture the overall attractiveness of a locality, $l$, where a venue $v_i$ is situated, we define its catchment using the taxi datasets as: $(\frac{|D_l|}{|D|})$, where $D$ is the total number of taxi drop-offs across the city and $D_l$ is the number of taxi trips that ended in a location within $l$.

**Temporal Catchment of the Locality**: While the above catchment feature is computed over the entire 24-hour day, we also subdivide and compute the catchment over four disjoint time partitions: morning (6 AM to 12 noon), afternoon (12 noon to 6 PM), evening (6 PM to 12 AM) and early morning (12 AM to 6 AM), separately across weekdays and weekends. During a window, $w$, the catchment for that window is then defined as, $\frac{|D_{w,l}|}{|D_w|}$.

### 4.5.2 Visit Patterns

With the advent of LBSNs, it is now possible to obtain finer-grained observations on customers visiting individual retail locations. The variation in trends of customer visits to such locations can reveal important insights into how businesses are faring, and hence should hold predictive power over future performance. Accordingly, we define the following visitation-driven features for further exploration.

**Inflow and outflow**: We employ a number of network metrics as features. We examine the *in-degree centrality* and *out-degree centrality* per month of each venue. In-degree centrality is defined by the number of in-flow transitions a given node receives from other nodes. Conversely, out-degree centrality represents the number of transitions that depart from a given node and end at other nodes. In the context of our work, we calculated the number of Foursquare transitions that arrive at the venue of interest to be the inflow to that venue and the number of corresponding transitions that leave from the venue to be the outflow. Additionally, we examine transitions from/to all venues in the neighbourhood, $N_i$, to compute the **Surrounding Area Inflow** and **Surrounding Area Outflow**. These metrics are normalized by the lifespan of the venue in months, providing a mean inflow and outflow per month. We also calculate the ratio of inflow to outflow transitions for the venue, as well as the corresponding surrounding area ratio. Formally, where $M$ is the lifespan of the venue in months and $t(v_j, vi)$ is the number of transitions from $v_j$ to $v_i$, we define the average monthly inflow to venue $v_i$ as follows:

$$\frac{\sum_{j=0}^{|V|} t(v_j, vi)}{M} \tag{4.4}$$

Additionally, we define outflow as follows:

$$\frac{\sum_{j=0}^{|V|} t(v_i, vj)}{M} \tag{4.5}$$

These features collectively capture the characteristics of transitions between different venue pairs and thus help summarise the potential cooperation effects with other venues in the neighborhood. For example, a high in-degree centrality suggests that a venue is a popular visiting spot for visitors after they have visited other businesses in the neighborhood.

**Distance of travel**: Prior research has shown the distance of travel to and from certain nodes in a network correlates with higher changes of connection between those nodes [151]. As an additional measure of the accessibility of an area, we thus measure the mean distance of travel of transitions to the venue of interest, as well as to the surrounding area(i.e. the venue's neighborhood). Formally, if $dist(v_j, v_i)$ is the distance between $v_i$ and $v_j$ and N is the total number of transitions to the venue, we define this as:

$$\frac{\sum_{j=0}^{N} dist(v_j, v_i)}{N} \tag{4.6}$$

**Speed**: Another important aspect that may influence the popularity of a retail location is how accessible it is to urban dwellers. As a possible measure of such accessibility, we compute the speed with which the venue can be reached from other locations. Formally, we compute the mean **Speed of Travel To** and the mean **Speed of Travel From** the venue of interest. We utilize a similar definition for the mean speed of the surrounding areas. We define speed as:

$$\frac{\sum_{j=0}^{N} dist(v_j, v_i) \times t_{i,j}^{-1}}{N} \tag{4.7}$$

where $dist(v_j, v_i)$ is the distance between $v_i$ and $v_j$, $t_{i,j}$ is the time spent travelling between $v_i$ and $v_j$, and $N$ is the total number of transitions to $v_i$. While the travel distance $dist(v_j, v_i)$ can be computed given the location coordinates of the venue pair $(v_i, v_j)$, the travel time is estimated with the difference in the transition start and end time. For these features, we also compute the standard deviations of these variables following standard definitions.

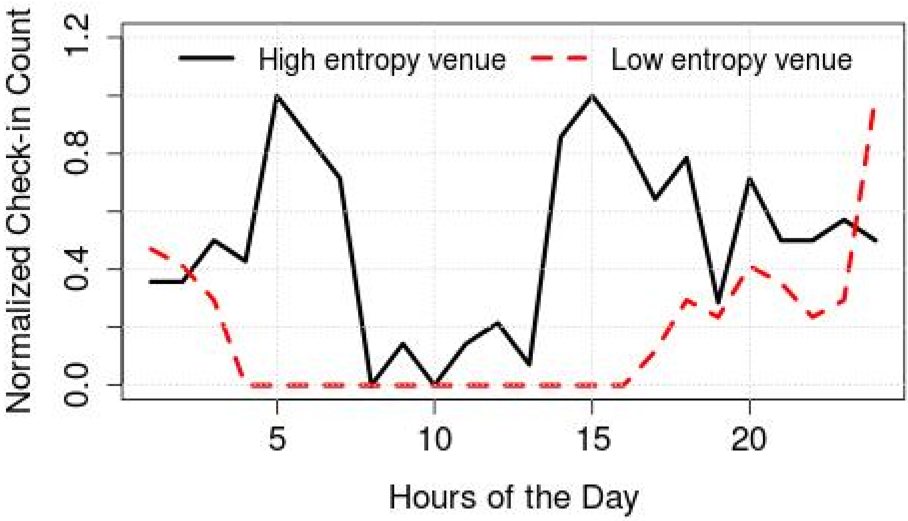



Fig. 4.4 The hourly popularity of a Pizza place in NYC which draws customers around the clock (high entropy) and a Frozen Yogurt place that draws customers mostly towards evening hours (resulting in low entropy).

**Temporal Popularity Skew**: We define the hourly temporal profile of a venue, $h_i$, as a vector of 24 elements, with each element representing the proportion of check-ins the venue has received during that hour as compared to the total check-ins received over all hours. We measure the skew as the entropy of the venue's temporal profile (equivalent to Equation 4.3). As seen in Figure 4.4, a venue that is popular across all hours would have a higher entropy compared to a venue with greater temporal skew (e.g., Frozen Yogurt place).

**Temporal Alignment of Venue to its Competitors**: Past work [119] has utilized the concept of *diurnal synchronization* of a venue to demonstrate how a venue is able to observe a larger set of transitions when its operating hours are more closely aligned with its surrounding venues. Implicitly, a venue is likely to witness spillover to/from other venues, only if their operating hours overlap. Given a venue $v_i$ with a temporal profile $h_i$ and its competitive neighborhood $CN_i$ whose aggregate temporal profile is $H_i$, we posit that venues that operate out of band from their competitors may have a higher likelihood of survival (as they effectively face less competition). Accordingly, we derive this misalignment feature by computing the Euclidean distance between the two vectors:

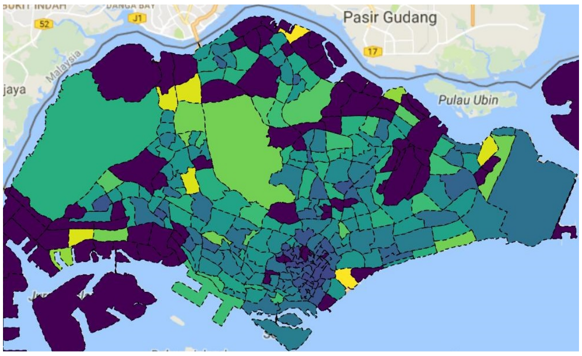
$$\sum_{j=1}^{24} (h_i(j) - H_i(j))^2 \tag{4.8}$$

**Visit Trend**: In [175], the authors show that the growth rate of a venue's check-ins has significant predictive power over its performance in the next three months. Motivated by this observation, we quantify a venue's temporal trend as follows: $C_t(v_i)$, we fit a linear regression model whose slope ($s(v_i) = \frac{C_t(v_i)-b}{t}$) represents the trend, with $b$ being the intercept. Whilst our definition of closure (in Eq. 4.2) and the trend feature both look at the temporal profile of check-ins, they are distinct in that while the trend captures drop/rise in check-ins within the observation window whilst closure is decided based on drop/rise before and after PD (based on average monthly volume of check-ins). To verify that there isn't potential leakage between these two constructs, we computed the correlation between "trend" and the quantity $\frac{\frac{\sum_{t=0}^{T} C_t(v_i)}{T}}{mean(v_i)}$ (see Eq. 4.2) and found that this was weak (varying between -0.018 and 0.06 across all cities, and 0.1 in the worst case for Los Angeles).

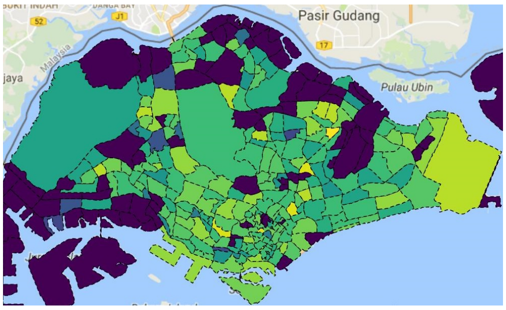
### 4.5.3 Mobility Dynamics

In addition to defining features based purely on a venue's visitation patterns and the static features of its neighborhood, we also define additional mobility-centric features. These features embody our intuition that the temporal pattern of movement of visitors to/from the venue, relative to the temporal pattern of visitors to the broader locality, helps capture the latent preferences of the urban population.

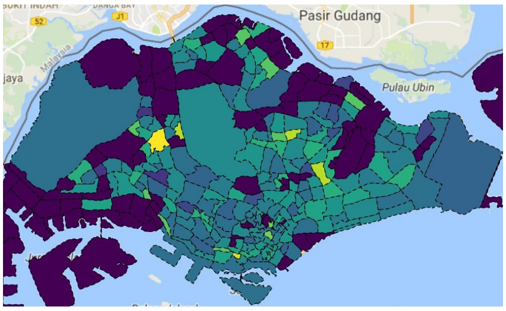
**Temporal Alignment of Venue to its Locality**: We hypothesize that a mismatch between the natural timings of the draw of a locality and the venue of interest may impact a venue's chances of survival. For instance, a cafe operating in a residential area may only see high volumes of customers during the morning or evening hours, due to the working population, who are typically the income generators, moving towards other areas (e.g., CBD) for work. Similarly, a late-night spot in the CBD area, may suffer from a lack of visitors during the




(a) Place Entropy




(b) Temporal Popularity Skew




(c) Alignment to Locality

Fig. 4.5 Select features spatially aggregated over localities across Singapore.

late evening, or early morning hours, and missing out on the opportunity of serving a large population during lunch hours. Similar to the case of venues, we define the hourly temporal profile for a locality as a vector of proportional arrivals to the locality for each hour, $h_l$, during the observation period. Then, we define the (mis)alignment as

$$\sum_{j=1}^{24} (h_i(j) - h_l(j))^2 \tag{4.9}$$

**Reachability of Locality:** We further hypothesize that a locality's accessibility plays a critical role in the survival of its venues. To quantify this, we first construct the transition matrix, $R$, whose elements $r_{a,b}$ represent the total number of trips that originated from locality $a$ and ended at locality $b$ during the observation period. A *reachable* locality is one that attracts trips from many localities. We measure the *reachability* of a locality $a$ as the entropy of the $a^{th}$ column of $R$, $R_a$. Further, a *reachable* locality should attract visits from both distant and local regions. To account for this, we weigh the frequency of transitions inversely by the distance between the regions with $dr_{a,b} = r_{a,b}/d^2{}_{a,b}$ where $d_{a,b}$ is the Haversine distance between the localities $a$ and $b$, and $dr_{a,b}$ are the elements of the modified transition matrix $\hat{R}$. In Figure 4.6, we contrast the two features; the reachability vector shows that the locality

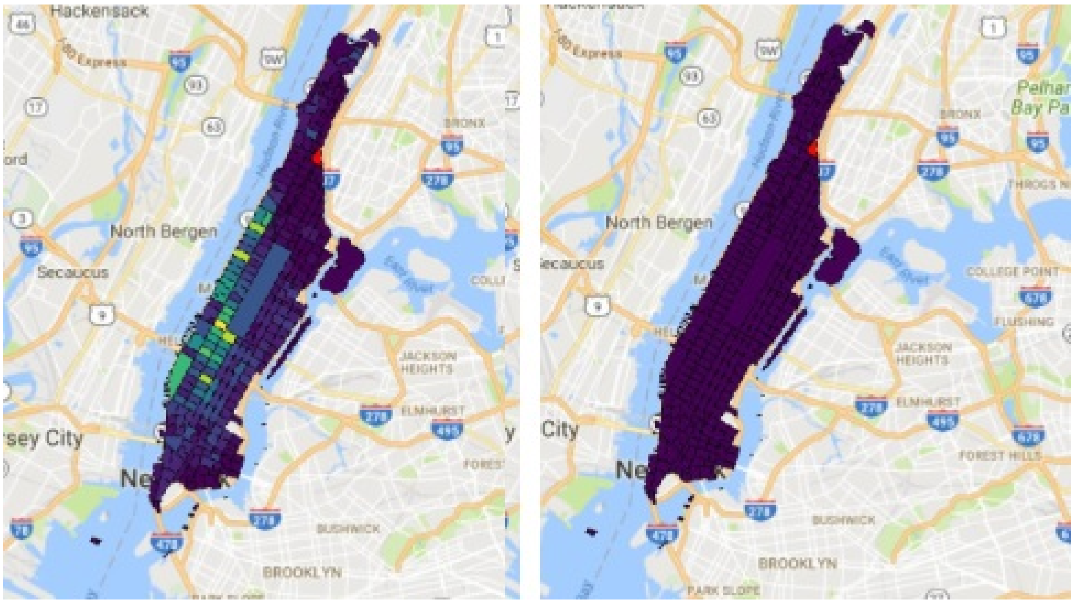



Fig. 4.6 The reachability matrix (left) shows that the locality receives more visits from farther localities whilst its distance-weighted reachability matrix (right) takes the distance into account.

receives much of its footfall from father localities (resulting in low entropy due to such skew) whereas the distance-weighted reachability vector is more uniformly distributed (resulting in high entropy).

Additionally, we also consider a number of control variables such as the **Specific Category** of the venue and the **Price Tier** of the venue. The tiers range from 1 (least pricey) to 4 (most pricey). In Figure 4.5, we visualize the spatial spread of three features aggregated over the different localities. We use this venue characteristic as a categorical feature in our model.

## 4.6 Evaluation

In this section, we report our findings on the predictive ability of individual factors that we consider in this chapter, and the performance of our methodology overall. We first discuss the influence of individual factors on the predictability of survival likelihood in Section 4.6.1, and summarise the overall performance of the combination of features in Section 4.6.2 for the two cities New York and Singapore, for which we have both Foursquare as well as transport data, and further extend our analysis to *Retail* businesses at large. In Section 4.6.3, we discuss the accuracy of our model across different feature classes. Then, in Sections 4.6.4 and 4.6.5, we scale our analysis to ten different cities around the world (focusing on *Visit Patterns*, which can be derived from Foursquare data alone), to answer two additional key questions:

1. do factors that affect business survival vary from city to city?
2. is our prediction framework able to detect the failure in both new and established businesses?

**Prediction task:** We represent the venue closure prediction task as a binary classification task with the closure label (0 – closed and 1 – open) as the dependent variable and the features described in Section 4.5 as independent variables and adopt a Logistic Regression model in all our analyses. Logistic regression also provides the additional benefit of providing an understanding of the relative influence of the features on the prediction outcome.

**Experiment conditions:** As our dataset consists of an unbalanced number of samples of positive (i.e., open venues) and negative (i.e., closed venues) classes with the negative class being much smaller (see Table 4.2), we first create a subset of all the negative samples and randomly sampled, equal-sized positive samples, generating a balanced dataset. We then split the four groups into training and test sets with the training set consisting of 80% of the data on which we perform 10-fold cross-validation to pick the best performing model and report the accuracy of prediction on the test sets. All features described were min-max normalized. The number of training samples in each of the four groups, (1) F&B venues in Singapore, (2) *Retail* venues in Singapore, (3) F&B venues in NYC and (4) *Retail* venues in NYC were 1450, 2794, 552, and 1062, respectively.

**Performance metrics:** In all our analyses, we report the accuracy based on precision, recall, and AUC, following their standard definitions. Precision and recall represent the average over both the positive and negative classes.

**Implementation:** The computations related to logistic regression were performed using R (default package stats [133]) and the ROCR [145] library for performance calculations. The comparison across multiple machine learning models (see Section 4.6.5) was built using Python and the scikit-learn library [153]. The code is open-source and available on Github[6].

### 4.6.1 Feature Selection and Pruning

In order to understand the ability of the features described in Section 4.5 in predicting survival likelihood, we run logistic regression with each feature as the (only) independent

[6]https://github.com/kritts/urban_modelling_closing

variable and report the average $AUC$ over 10-fold cross-validation in Table 4.5 of the top-5 influential features for the two cities, respectively. We also report the correlation between the variables in each case – here, we compute the correlation coefficient as the root of the coefficient of determination ($R^2$), with the sign (positive/negative) based on the estimated coefficient from logistic regression. We avoid the use of the widely used Pearson's correlation coefficient [21] since the two-class dependent variable doesn't fit the linearity assumption that Pearson's requires. We see that the temporal popularity skew, and the temporal alignment with the competitors and the locality itself being top features consistently, each with a high $AUC \geq 0.75$. We apply the Boruta algorithm [99] for feature selection and consider the features that were consistent across F&B venues from both Singapore and New York City in subsequent analyses.

### 4.6.2 Predicting Venue Closure

In this section, we summarise our findings from running logistic regression [58] on the 20 confirmed features resulting from the Boruta search in Table 4.6. For brevity, we only show list features that were found to be statistically significant in the combined model. We run regressions separately for F&B venues, and extend to *Retail* venues in general. *Retail* venues consist of venues that belong to either F&B, Entertainment, Clothing Stores, Nightlife Spots, Food & Drink Shops, Gym/Fitness Centers and other Retail Shops.

We report the following key observations:

1. We see that a number of features consistently appear to have a strong influence on the prediction outcome; namely, the (1) visit trend over the current period, (2) the skew in hourly temporal popularity, (3) temporal (mis)alignment of the venue to its locality, and (4) the (entropy) of the distribution of venue types in the vicinity of a venue, across both cities, and for both F&B and *Retail*, with very few exceptions.

2. Based on the coefficients for the hourly temporal skew feature, it appears that venues that are popular *around the clock*, and not subjected to specific hours, may have a better

| **Feature** | **AUC** | **Correlation** |
|---|---|---|
| Temporal Popularity Skew | 0.788 | 0.528 |
| Alignment with Neighborhood | 0.786 | -0.479 |
| Alignment with Locality | 0.758 | -0.526 |
| Inflow | 0.66 | 0.270 |
| Outflow | 0.624 | 0.268 |

| **Feature** | **AUC** | **Correlation** |
|---|---|---|
| Temporal Popularity Skew | 0.794 | 0.589 |
| Alignment with Neighborhood | 0.782 | -0.575 |
| Alignment with Locality | 0.732 | -0.555 |
| Trend | 0.716 | 0.487 |
| Inflow (Neighborhood) | 0.626 | -0.201 |

Table 4.5 Features with the highest performance in predicting venue closure for Singapore (left) and New York City (right).

| **Feature** | **SG, *Retail*** | **SG, F&B** | **NYC, *Retail*** | **NYC, F&B** |
|---|---|---|---|---|
| Inflow | -58.76 . | -1.61 | -21.01 | -34.01 * |
| Outflow | **175.04** *** | 4.42 | 68.27 ** | 47.02 ** |
| Speed Entering | -3.75 * | **-10.06** *** | -3.51 | -17.17 * |
| Hourly Temporal Skew | **9.80** *** | **10.53** *** | **9.13** *** | **12.97** *** |
| Visit Trend | **0.13** *** | **0.13** *** | **74.34** *** | **58.67** *** |
| Total Visits | 0.82 | 0.55 | 5.41 *** | 1.99 |
| Place Entropy | -2.33 ** | -2.31 * | 4.55 | -3.70 |
| Distance Entering, Sur | -11.32 ** | -1.19 | -6.91 | -3.31 |
| Distance Leaving, Sur | 7.22 * | 1.64 | -0.89 . | -0.48 |
| Temporal Alignment to Locality | **5.67** *** | **6.96** *** | 4.58 . | 8.69 * |
| N | 2794 | 1450 | 1062 | 552 |
| R2CU | 0.356 | 0.349 | 0.580 | 0.577 |

Table 4.6 Coefficients from Logistic Regression for two cities. *** represents $p < 0.001$, ** represents $p < 0.01$, and * represents $p < 0.05$. SG - Singapore, NYC - New York City.

chance at survival. This finding suggests that restaurants that only cater to specific customer segments (e.g., lunchtime office workers or dinnertime visitors) are more likely to experience failure. To further analyze this, we looked at the failure rate, in Singapore, for restaurants in two neighborhoods with skewed visitor dynamics: the Central Business District (CBD) that has a dominant lunchtime presence, and Clarke Quay (CQ) that is geared towards tourist and leisure traffic and is more active at night. We picked all restaurants from CBD and CQ (139 venues in total), and ranked them by their hourly popularity entropy. We compare the top 30 restaurants (highest entropy) and the bottom 30 restaurants (lowest entropy)–i.e., approximately, the top and bottom 20-percentile of such venues. We find a clear difference: whilst only 73% of the bottom-30 restaurants survived the next 6 months, 100% (all 30) of the venues in the top-30 survived.

3. On the contrary, the estimated coefficient of *place entropy* (i.e., negative) suggests that a decrease in entropy improves the likelihood of survival. This seems to suggest that venues that are in the midst of more clustered neighborhoods (such as ethnic enclaves) tend to survive longer. Also, noteworthy, is the significantly large magnitude for NYC under the *F&B* category. This suggests that, in NYC, restaurants have far better survival rates if they are clustered by cuisine (e.g., restaurants in 'Little Italy' or 'Chinatown'), rather than being situated in a less distinctive neighborhood. Upon closer investigation, aggregated at the Neighborhood Tabulation Area granularity [7], we

[7] http://maps.nyc.gov/census/

see that ethnically clustered areas such Chinatown and Little Italy rank low on Place entropy (0.077 and 0.079, respectively) whilst areas such as Washington Heights and Inwood have higher scores in the range of 0.15-0.18. A Pearson correlation coefficient of 0.715 between the place entropy values and mortality rates of those neighborhoods suggests a strong link between low entropy and greater survival likelihood.

4. Not surprisingly, the trend of customers check-in patterns during the current period is indicative of the venue's performance over the following 6-month period. And, as anticipated, we also note that the sign of the coefficient is positive, indicating that venues that experience an upward trend in check-ins have a much higher likelihood of survival.

5. Between the two cities, we observe that in the case of Singapore, several of the features tied to a venue's locality or neighborhood are found to be statistically significant – for instance, the (mis)alignment, measured by the Euclidean distance, suggests that venues that operate outside popular hours of the locality have a distinct advantage over their neighbors.

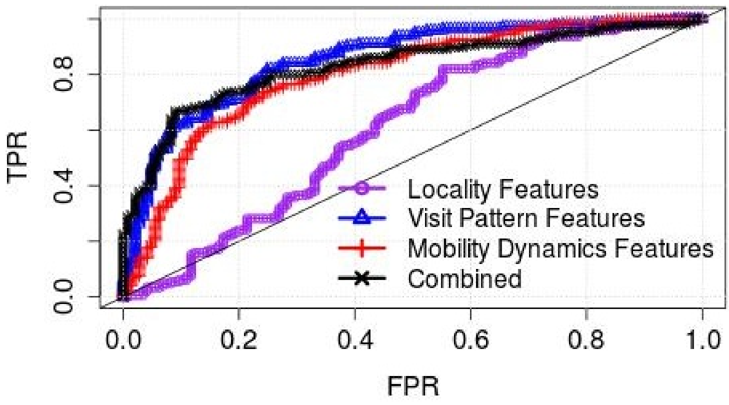
**Comparison with baseline:** As we describe in Section 4.2, the work of Wang et al. [175] is the closest to our work in that they rely purely on LBSN-based features to study the decline in business performance for F&B venues in NYC, over a 3-month window. In Table 4.7, we compare our results against this baseline, reproducing the confusion matrix presented in [175]. The features we consider in this work achieve at least 15% better precision, and a 7-10% increase in recall for venues in NYC. Our evaluation is on a balanced test set, whereas the baseline misclassification rates reported in [175] may be a bit misleading as they are based on a highly imbalanced dataset (with less than 20 samples in their “closed" class).

### 4.6.3 Accuracy Across Feature Classes

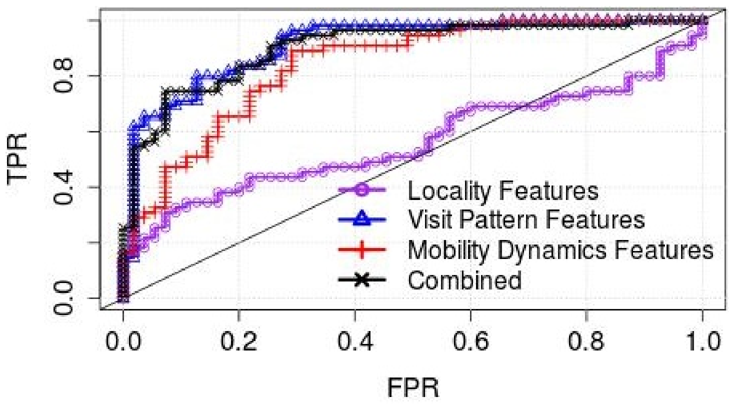
To understand the influence of features classes on the prediction outcome, we run the logistic regression for each class separately, and in combination. In Table 4.8, we report the observed

| | **Wang et al.**[175] | | **NYC, Retail** | | **NYC, F&B** | |
|---|---|---|---|---|---|---|
| *Confusion Matrix* | Labeled 0 | Labeled 1 | Labeled 0 | Labeled 1 | Labeled 0 | Labeled 1 |
| *Observed - 0* | 160 | 5 | 87 | 19 | 46 | 13 |
| *Observed - 1* | 7 | 14 | 19 | 87 | 9 | 42 |
| Precision (closed class) | 66.67% | | 82.07% | | 82.35% | |
| Recall (closed class) | 73.68% | | 82.07% | | 76.36% | |

Table 4.7 Confusion Matrix Comparison against Previous Work [175].




(a) Singapore, *Retail*




(b) New York City, *Retail*

Fig. 4.7 ROC Curves of *Retail* venues in Singapore and New York City. The Curves represent the performance for each class of features and for the combined model, respectively.

AUC scores for *Retail* venues from Singapore and New York, and note that in general, the visit pattern features alone reach AUCs $\geq 0.80$, and the addition of mobility dynamics features lends to only a 4% and 3% increase in accuracy for Singapore and NYC, respectively. Figure 4.7 shows the ROC curves for the same. These results suggest that while mobility features do help inform predictions, visit patterns are more informative with regards to venue forecasting. This may be because visitation patterns are most directly indicative of the demand of venues.

### 4.6.4 Individual Cities Versus All Cities

We next examine how similar the impact of features is for the ten different cities listed in Table 4.2. We analyze venues whose specific category are *Retail*, as defined above and consider features that are extracted using Foursquare alone (see Table 4.4 –i.e., the various *Customer Visit Pattern* features and the bulk of the *Locality Profile* features) due to the unavailability

| | SG, Retail | NYC, Retail |
|---|---|---|
| N | 2794 | 1062 |
| Random Baseline | 0.50 | 0.50 |
| Locality | 0.60 | 0.58 |
| Mobility Dynamics | 0.80 | 0.80 |
| Visit Pattern | 0.82 | 0.89 |
| Combined | 0.86 | 0.92 |

Table 4.8 AUC scores of the different feature classes with Logistic Regression against the Random Baseline. The *Contrast* set consists of venues with the top-5% and bottom-5% values of the reachability feature.

of transport features across all cities. We first examine the per city AUC scores using our Logistic Regression model. We summarise our results of coefficients and their significance level in Table 4.9. The results show a range in the coefficients for all features, suggesting that certain factors differently affect different cities. However, our results show that certain features are consistently significant across many or all our cities. We see that *Visit Trend* acts as a significant feature for all ten cities. Additionally, *Temporal Popularity Skew*, *Temporal alignment to Competitors*, and *Distance of Travel From* are significant features in many of the cities. These results suggest that the dynamics of urban environments have a strong influence on venues. Additionally, we note that the sign of the coefficient is consistent across all cities, suggesting the role these features play in the success of businesses is consistent. This analysis suggests that large-scale multi-city generalizations can be challenging as cities are often unique in their attributes. This suggests that although venue closure is predictable as a task, features vary by city in their contribution to business failure in individual cities.

### 4.6.5 The Impact of Venue Age on Prediction Accuracy

We examine next the predictability of two classes of venues: those that are established, *est*, and those that are new, *new*. We define established venues as those that have existed for longer than a year. Conversely, we define new venues as those that have existed for less than one year. This analysis is performed across all the 10 cities, primarily because the number of new venues in any single city is too small for meaningful analysis. We analyze venues whose specific category is *Retail*, as defined above.

As described in more detail in Section 4.4.2, we examine venue closure between the time period of July to December 2013. For our analysis, we select new venues as those that opened on or after June 1 2012. As prior work has shown that venues added (in Foursquare) after June 2011 were highly likely to actually be new venues [37], we have strong confidence

| | Visit Trend | Temporal Skew | Alignment to Neighborhood | Distance of Travel From | Speed Entering | AUC |
|---|---|---|---|---|---|---|
| Chicago | 0.16*** | 2.05*** | -3.66** | 3.90 | -0.019 | 0.8605 |
| Helsinki | 0.06* | 1.72* | -6.47* | 3.16** | -0.37 | 0.7725 |
| Jakarta | 0.13*** | 1.46*** | -0.80 | 2.21* | -0.093 | 0.8326 |
| London | 0.19** | 1.66 | -7.68* | 1.65* | -0.35 | 0.7689 |
| Los Angeles | 0.32*** | 2.23*** | -3.45* | 0.82 | -0.033 | 0.8001 |
| New York | 0.16*** | 1.96*** | -2.57* | 0.76 | -0.41** | 0.8633 |
| Paris | 0.15*** | 2.07*** | -1.86 | 0.72* | -0.24 | 0.8203 |
| San Francisco | 0.27*** | 1.72** | -3.44 | 1.14 | -0.37 | 0.7775 |
| Singapore | 0.10*** | 3.26*** | -0.34 | 0.32 | -0.041 | 0.8355 |
| Tokyo | 0.13*** | 1.39*** | -2.04* | 0.48* | -0.26** | 0.8155 |
| All cities | 0.12*** | 2.40*** | -0.74 | 0.12* | -0.12* | 0.8803 |

Table 4.9 Per city AUC score and the logistic regression coefficients for multiple cities for the top five most significant features. *** represents $p < 0.001$, ** represents $p < 0.01$, and * represents $p < 0.05$.

in our classification. We compare the prediction performance with multiple classifiers and present our results in Table 4.10.

There is variability across different classifiers; the best performance in terms of AUC is given by a gradient boosting model, which resulted in an AUC of 0.882. Further examination of the results also shows that established venues have a notably higher AUC scores than new venues, suggesting, as intuitively expected, that new venues may be more susceptible to high variations in their causes of failure.

| | **New venues** | | | **Established venues** | | |
|---|---|---|---|---|---|---|
| Classifier | Precision | Recall | AUC | Precision | Recall | AUC |
| Logistic regression | 0.701 | 0.758 | 0.788 | 0.767 | 0.804 | 0.861 |
| **Gradient boosting** | 0.717 | 0.737 | 0.812 | **0.794** | **0.806** | **0.882** |
| Support vector machine | 0.693 | 0.765 | 0.817 | 0.771 | 0.801 | 0.863 |
| Random forest | 0.742 | 0.7 | 0.821 | 0.802 | 0.732 | 0.855 |
| Neural Networks | 0.737 | 0.668 | 0.809 | 0.779 | 0.767 | 0.858 |

Table 4.10 Closure predictions for new and established venues.

| Feature | Combined Model SE | Combined Model VIF | Reduced Model SE | Reduced Model VIF |
|---|---|---|---|---|
| Temporal Popularity Skew | 3.85 | 11.83 | 0.84 | 1.06 |
| Visit Trend | 18.82 | 3.55 | 8.80 | 1.03 |
| Place Entropy | 2.77 | 3.52 | 1.86 | 1.58 |
| Inflow | 18.66 | 9.84 | 4.72 | 1.16 |
| Outflow | 18.17 | 10.00 | | |
| Distance Entering, Surrounding | 2.61 | 3.60 | | |
| **AUC** | 0.92 | | 0.89 | |

Table 4.11 Standard Error of Estimated Coefficients and Variable Inflation Factors of Selected Features for Retail Venues in New York City for the Combined Model (left) and Reduced Model (right). SE- Standard Error.

### 4.6.6 Robustness Checks

In this section, we perform a series of checks to understand the robustness of our analyses under various conditions.

#### Dealing with Collinearity

In addition to the feature selection step carried out in Section 4.6.1, here we test for collinearity across variables - to detect collinearity, we compute the Variable Inflation Factor ($VIF$) [120] of variables – a score greater than 12 suggests that there exists significant correlation across certain variables. We tabulate the standard errors and the $VIF$ values, as an illustration, for the case of F&B venues in New York City. The table compares the *Combined Model* consisting of the select features from Section 4.6.2, against a subset of features whose Pearson's correlation with any other feature is less than 0.5 which we refer to as the *Reduced Model*, in Table 4.11.

We find that: (1) removing the uncorrelated features reduces the standard error in the estimated coefficients and lowers the $VIF$ significantly (all less than 2), and (2) in removing the correlated variables, the resulting $AUC$ drops only marginally.

#### Impact of Amount of Past Data on Performance

In our analyses thus far, we consider an observation window of 6 months (prior to the prediction date) to predict the survival of a venue in the following 6 month period. A natural question then is: *how much data from the past is required to make a reasonable prediction?*

To answer this, we vary the length of the observation window between two (a minimum of two data points are needed for calculating the trend feature) and 6 months, immediately

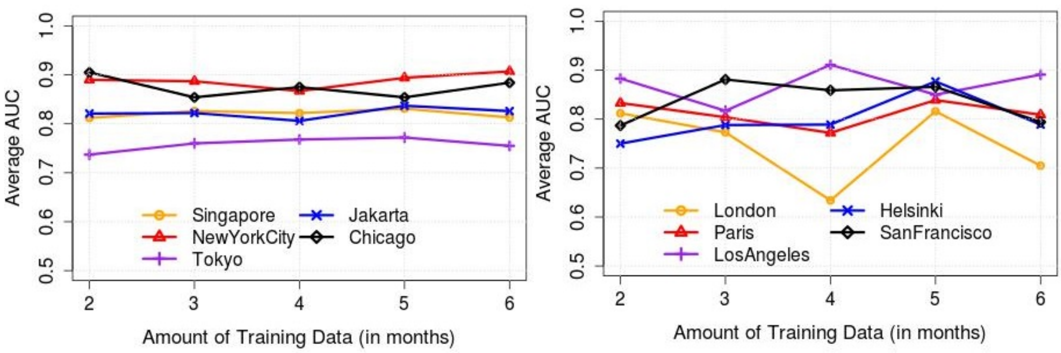



Fig. 4.8 Impact of the Length of Observation Window on Performance.

preceding the prediction date, and repeat our analysis. In Figure 4.8, we plot the amount of training data (in months) on the $x-$axis, and the mean AUC over 10-fold cross-validation on the $y-$axis. We plot our results for all ten cities and note that the AUCs are relatively stable. The worst case of a 20% drop occurs in the case of London besides which the variability is limited for other cities.

## 4.7 Discussion

Here, we describe key implications of our work to business owners and urban planners alike, discuss the limitations in the current study, and our plans for future work.

In our analyses, we saw that both visit patterns to venues and the characteristics of the locality can play a role in the survival likelihood of a venue. While we see $AUCs \geq 0.82$ in general, which demonstrates the theoretical merit of this work, for a practical adoption by stakeholders, we further investigated the precision-recall trade-offs. In the case of impending business failure, a high recall would be warranted as retailers would be less sensitive to false-positives (i.e., the system predicting that the venue is likely to fail, but the venue survives in reality) than vice-versa. For a recall of 0.90, in the case of New York City, for example, a reasonably high precision of 0.83 and 0.73 can be achieved for F&B and over all retail venues, respectively. Whilst in Singapore, the achievable precision drops to 0.74 and 0.70 for the F&B and *Retail* classes. We share key takeaways from interviewing several stakeholders.

**F&B Owners:** The owners (referred to as Owner1 through Owner6) rated an average of 3.5 (on a scale of 1:not useful to 4: very useful) when asked whether understanding their businesses' survival likelihood in the next 6 months is useful to them. They all found the accuracy (of 80-90%) to be either sufficient, or good, for taking precautionary actions except Owner5 who said that the accuracy is low. “Too low to be useful" was also an option that none of the owners chose. Everyone found the prediction horizon (of 6 months) to be appropriate and mentioned that they use a combination of Point-of-Sales data, social media and third-party services like Shopify [157] to monitor current health of their business. Only one respondent (Owner2) said that they perform trend analysis to forecast future performance. All except one respondent said they are either “Likely" (1/6) or “Very Likely" (4/6) to take actions (such as revising the menu or run promotional campaigns) based on reports generated by a future survival prediction system – interestingly, the respondent (Owner2) who chose “Not Likely" runs a franchise of a fast-food company and shared that such changes or decisions cannot be made by franchisees independently, but can only be made by the franchisor which is then implemented across the board by all franchisees.

**Urban Authorities and Planners:** We reached out to an experienced planner at the local authority in Singapore who responded that the agency is interested in knowing survival rates at both the individual (6 on a Likert scale from 1 to 7) and aggregate level (4 on a Likert scale from 1 to 7). The planner also found the prediction horizon of six months to be appropriate, although he felt that the level of accuracy (i.e., 80%) would be too low for the agency to make concrete interventions. He also shared that the agency sees potential in the overall methodology of combining data from LBSNs and urban transportation for informing planning decisions, and in studying people’s behavior, choices, and patterns, in general.

**Public Policy Expert:** We spoke with a Professor of Public Policy in the UK whose response to our precision and recall results ($\approx$ 80%) were positive. He suggested the work could have implications on licensing agreements for new venues by local authorities who currently look for factors such as location and competition [2]. Analyzing the likelihood of failure of that area could be considered as an additional factor in those agreements. He shared that the models could have commercial value for both technology intelligence companies and large retail businesses. Similarly, given the positive impact of aligning a store’s hours with the mobility dynamics of the locality (the *Temporal Alignment* feature in the Mobility Dynamics class), planners could recommend or incentivize restaurants or stores to remain open at certain key hours. In the future, we believe that data-driven analyses such as ours could inform policymaking.

## 4.8 Conclusions

We have presented an approach to the prediction of venue closure. The approach uses a variety of features over spatio-temporal data crowd-sourced from location-based social networks and transport data. Our results show that it is important to go beyond just static features, of a retail establishment and its neighborhood, and include mobility-derived features, related to both the visitor patterns to the venue and the aggregate movement patterns over the venue's neighborhood. Using a variety of such features we show, through an analysis of 10 cities that an appropriate classification model can generate higher prediction accuracy (AUC of $\approx$0.80) than previously reported.

In the following chapter, we forecast the demand of businesses in urban environments. Demand is a function of many complex and nonlinear features such as neighborhood composition, real-time events, and seasonality. However, recent advances in Graph Convolutional Networks (GCNs) networks have had promising results as they build a graphical representation of a system and harness the potential of deep learning architectures. The subsequent chapter presents a model using GCNs to forecast venue demand by representing venues as connected networks of places in a city.

# Chapter 5

# Modelling Urban Dynamics with Multi-Modal Graph Convolutional Networks

In the previous chapter, we used supervised learning models to predict the survival of urban businesses through mobility dynamics, neighborhood profiles, and network features. This chapter builds deep learning models to predict business demand, using Graph Convolutional Networks; it forecasts the future demand of venues by incorporating spatial and topological features into a temporal model.

The main contribution of this chapter is twofold. First, representing cities as connected networks of venues, we quantify their structure and characterize their dynamics over time. We note a strong community structure emerging in these retail networks which highlights the interplay of cooperative and competitive forces within establishments. Second, we present our deep learning architecture which integrates both spatial and topological features into a temporal model which predicts the demand of a venue at the subsequent time-step. Relative to state-of-the-art deep learning models, our model reduces the RSME by $\approx 28\%$ in London and $\approx 13\%$ in Paris. Our approach highlights the power of complex network measures and GCNs in building prediction models for urban environments. The methodology and results can support policymakers, business owners, and urban planners in the development of models to characterize and predict changes in urban settings.

## 5.1 Introduction

Predicting venue demand has long been an active area of research because of its inherent value for numerous stakeholders. Location is known to be highly influential in the success of a new business opening in a city. Where a business is positioned across the urban plane not only determines its reach by clienteles of relevant demographics but more critically, it determines its exposure to a local ecosystem of businesses who strive to increase their share in a local market. The types of businesses and brands that are present in an urban neighborhood, in particular, has been shown to play a vital role in determining whether a new retail facility will grow and blossom, or instead whether it will become a sterile investment and eventually close [84]. Competition is nonetheless only one determinant in retail success. How a local business establishes a *cooperative network* with other places in its vicinity has been shown to also play a decisive role in its sales growth [37]. Local businesses can complement each other by exchanging customer flows with regards to activities that succeed each other (e.g. going to a bar after dining at a restaurant), or through the formation of urban enclaves of similar local businesses that give rise to characteristic identities that then become recognizable by urban dwellers. A classic example of the latter is the presence of many Chinese restaurants in a Chinatown [190].

It is therefore natural to hypothesize that the rise of a business lies in the complex interplay between cooperation and competition that manifests in a local area. Modeling these cooperative and competitive forces in a city remains, however, a major challenge. Today's cities change rapidly driven by urban migration and phenomena such as gentrification as well as large urban development projects, which can lead to shops opening and closing at increasing rates. In 2011, the *fail rate* of restaurants in certain cities, such as New York, was as high as 80% [1] with some businesses closing in only a matter of months. A similar picture has been reported for high street retailers in the United Kingdom with part of the crisis being attributed to the increasing dominance of online retailers [2]. Data generated in location technology platforms by mobile users who navigate the city provides a unique opportunity to respond to the aforementioned challenges. In addition to providing quick updates, in almost real-time, on venue demand in cities - thus accurately reflecting the visitation patterns of local businesses in a given area - they offer a view on urban mobility flows between areas and places at fine spatial and temporal scales. The ability to describe these two dimensions of urban activity - places and mobility - paves the way for predicting venue demand at subsequent time-steps. In this work, we harness this opportunity, building on a longitudinal

[1]https://www.businessinsider.com/new-york-restaurants-fail-rate-2011-8

[2]https://www.theguardian.com/cities/ng-interactive/2019/jan/30/high-street-crisis-town-centres-lose-8-of-shops-in-five-years

dataset by Foursquare that describes mobility interactions between places in London and Paris. There has been limited prior work using temporal Graph Convolutional Networks (GCNs) to model trends in urban areas. This chapter is a first step in harnessing the power of deep learning in conjunction with graph attributes to predict future growth trends for venues in a given area. Our contributions are summarized in more detail in the following:

- **Detecting patterns of cooperation in urban activity networks:** We model businesses in a city as a connected network of nodes belonging to different activity types. We examine the properties of these networks spatially and temporally. With respect to null network models, we observe a higher clustering coefficient, higher modularity, and lower closeness centrality scores which are indicators of strong tendencies for local businesses to cluster and form collaborative communities that exchange customer flows. Strong community structure emerges in these local retail networks and cooperative effects appear naturally, reflecting interactions of nodes within and across communities. In numerical terms, the modularity of urban activity networks is $\approx 0.6$ relative to the corresponding null models with $\approx 0.15$.

- **Novel temporal GCN architecture which combines topological and spatial structure:** We describe our architecture which incorporates both spatial and topological features into a Graph Convolutional Network. We begin by visualizing the temporality of our data and how it can be used to build a network representation. We then detail our architecture which has the potential to be broadly applicable across numerous domains.

- **First temporal GCN applied to urban dynamic prediction:** We employ our spatio-topological temporal model to predict the demand of a given venue in the subsequent month. Our model has a high predictive power over baselines, reducing the RSME of demand prediction by $\approx 28\%$ in London and $\approx 13\%$ in Paris relative to an LSTM. We consistently outperform baseline models across both metrics and both cities of interest. We further show that our model converges more quickly suggesting it is able to quickly learn trends in venue popularity using both topological and spatial features. Our approach shows how complex networks can be used in dynamic deep learning models. Further, it has the potential for broad applications in urban planning and retail through predictions of future growth trends.

Our results are especially important in a digital age with shifting customer preferences as physical businesses are forced to adapt to remain competitive. Our methodology can enable a better understanding of interactions within local retail ecosystems. Modern data and methods, such as those employed in the present work, not only can allow for monitoring these phenomena at scale but also offer novel opportunities for retail facility owners to assess future demand trends though location-based analytics. Similar methods can be applied beyond the scope of the retail sector we study here, namely for urban planning and innovation e.g. through assessing future demand trends of transport hubs, leisure, and social centers, or health and sanitation facilities in city neighborhoods.

## 5.2 Related Work

Understanding retail ecosystems and determining the optimal location for a business to open have been long been questions in operations research and spatial economics [63, 49]. Compared to modern approaches, these methods were characterized by static datasets informing on population distribution across geographies, tracked through census surveys, and the extraction of retail catchment areas through spatial optimization methods [8]. Gravity models on population location and mobility later became a common approach for site placement of new brands [65].

By the early 90s many major brands were already developing their expansion and growth strategy using quantitative methods in a manner that reflected the demographics and preferences of local populations [31]. Multi-national corporations like Starbucks, McDonald's, and 7-eleven fell into this category in their search for international expansion [167].

The availability of spatio-temporally granular urban datasets and the popularization of spatial analysis methods in the past decade led to a new generation of approaches to quantify retail success in cities. In this line, network-based approaches have been proposed to understand the retail survival of local businesses through quality assessment on the interactions of urban activities locally [84]. In addition to networks of places, street network analysis emerged as an alternative medium to understand customer flows in cities, with various network centrality being proposed as a proxy to understand urban economic activities [36, 132]. While this previous work examines the impact of new businesses, it is limited in scope. It considers only homogeneous categories, does not compare network properties across different cities and does not build a prediction model. This is where the primary novelty of the present chapter lies.

More recently, machine learning and optimization methods have been introduced to solve location optimization problems in the urban domain, focusing not only on retail store optimization [91] but also real estate ranking [59] amongst other applications. Location technology platforms such as Foursquare opened the window of opportunity for customer mobility patterns to be studied at fine spatio-temporal scales [46, 44] and moreover, semantic annotations on places presented direct knowledge on the types of urban activities that emerge geographically and led to works that allowed for the tracking and comparing urban growth patterns at global scale [37]. Additionally, the authors in [73] study co-location patterns of urban activities in Boston and subsequently recommend areas where certain types of activities may be missing.

Finally, within the realm of temporal Graph Convolutional Neural Networks there has been limited work applying temporal GCNs to urban environments. The most similar application is that of traffic prediction in which a connected network of roads is used to predict traffic on roads at future time-steps [186]. However, there is a gap in the literature in the application of temporal GCNs beyond traffic prediction. This is where the primary novelty of this present chapter lies. It is the first work using dynamic temporal GCNs to predict urban demand. It shares a novel methodology using both spatial and topological features to inform dynamic predictions.

## 5.3 Dataset Description

As described in detail above in Chapter 3, the basis of our analysis is a longitudinal Foursquare dataset from two cities that spans three years, from 2011 to 2013, and includes over 1.5 million check-ins. For each venue, we have the following information: geographic coordinates, specific and general category, creation date, total number of check-ins, and number of unique visitors. We consider the set of venues $V$ in a city. A venue $v \in V$ is represented with a tuple $< loc, date, category >$ where $loc$ is the geographic coordinates of the venue, $date$ is its creation date, and $category$ is the specific category of the venue.

## 5.4 Urban Activity Networks

We begin by examining transitions between Foursquare venues of different category types that we refer to as urban activities. While we are considering a mix of categories users checkin in the city, our focus from an analysis and modeling point of view will be focusing on urban activities corresponding to retail establishments (e.g. restaurants). In summary, in the following paragraphs we note there exists visible structure in the network of urban activities which varies spatially from city to city as well as temporally across different times of days (e.g. morning versus evenings). We first visualize these trends and then quantitatively measure differences in their structure. We focus on London and Paris for our analysis

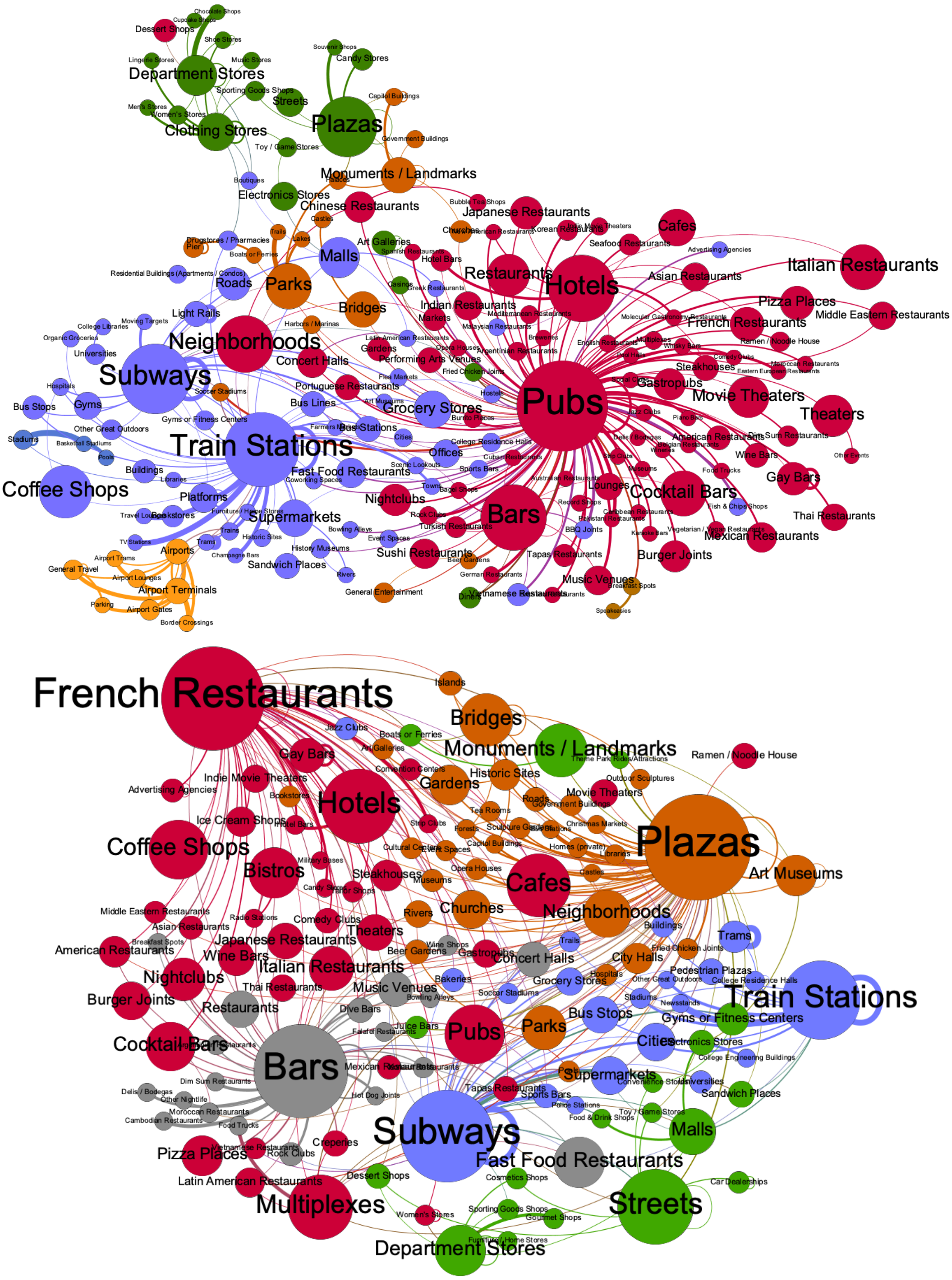



Fig. 5.1 Network visualization of categories during the evening in London (left) and Paris (right). Different colors represent different Louvain communities [22]. We see clusters of travel and transport (blue), nightlife (red), and shopping activities (green).

| | London | | | | Paris | | | |
|---|---|---|---|---|---|---|---|---|
| | AM | Random AM | PM | Random PM | AM | Random AM | PM | Random PM |
| # of nodes | 204 | 204 | 208 | 208 | 180 | 180 | 149 | 149 |
| # of edges | 2271 | 2271 | 3055 | 3055 | 2160 | 2160 | 2184 | 2184 |
| $< C >$ | 0.657 | 0.304 | 0.645 | 0.301 | 0.744 | 0.131 | 0.645 | 0.173 |
| $< C_c >$ | 0.313 | 0.552 | 0.402 | 0.541 | 0.352 | 0.513 | 0.434 | 0.551 |
| $Q$ | 0.583 | 0.150 | 0.608 | 0.131 | 0.588 | 0.148 | 0.641 | 0.161 |

Table 5.1 Network metrics for London and Paris for during the morning AM (6am - 12pm) and evening PM (6pm-12am). These metrics are compared to an Barabási-Albert model (Random).

### 5.4.1 Visualizing Mobility Interactions

To visualize an urban activity network, we create a graph $G_i$ for each city $i$, where the set of nodes $N_{cat}$ is the set of business categories defined previously in Section 5.3. In this network, business categories are linked by weighted directed edges $e_{s \to d}$. A directed link is created from the source category $c_s$ to the destination category $c_d$ if at least one transition happens during the time window we consider (e.g. weekend, weekday, or a period of hours during a day). Thus, the weight of each edge is proportional to the total number of transitions from the source category to the destination category for the particular time period of interest for each city. The weights are then normalized by the total number of check-ins that occurred at $c_d$. Therefore, the weight can be interpreted as the percentage of customers of $c_d$ who come from $c_s$. To eliminate insignificant links, we filter out edges that have less than 50 transitions in total. We examine two time intervals of interest: morning AM (6am-12pm) and evening PM (6pm-12am).

In Figure 5.1 we visualize the network in the evening for two cities, London and Paris. The colors represent different communities, obtained using the Louvain community detection algorithm [22]. Further, the size of the nodes is proportional to their degree. This visualization, as one example, describes similarities and variations in the structure of urban activities in different cities. We observe an underlying common structure for the two cities, even though cultural distinctions can also be noted. We have observed a similar pattern across different cities which we don't visualize due to lack of space. In terms of similarity in network structure, we see a shopping cluster (green) centered around Department Stores; a cluster for travel and transport (blue) centered around categories such as Train Stations and Subways; a leisure cluster (light brown) centered around Plazas and containing outdoor categories (e.g. Parks, Gardens, Soccer Stadiums). On the other hand, differences in network structure become also apparent. We note for instance how recreation activities in the evenings differ across the two cities. London has a considerably large nightlife cluster (red) centered around pubs from which a number of different nightlife categories unfold (e.g. Nightclubs, restaurants of different types, Theater). Paris is more segregated and contains two nightlife clusters: one cluster around French Restaurants (red) linked to Coffee Shops, Theaters, Nightclubs; and another cluster (gray) centered around Bars which contains Food Trucks, Fast Food Restaurants, and Music Venues. This dichotomy translates to the presence of two classes of customers each of which adheres to different types of activity sequences during nighttime. Another observation is regarding variations in network structure over time: the Coffee Shop category in London is separated from the nightlife cluster, which may indicate different kinds of customer behaviors between daytime and evening. Interestingly, we also see associations emerging between types of businesses. Taking Paris as an example, French Restaurants interact a lot with Coffee Shops and Nightclubs and so do Bars with Food Trucks. In both cities, Coffee Shops are drawing crowds from Subways, Toy Stores with Electronics Stores, and Sports Stores.

Overall, these results suggest strong structural characteristics in urban activity networks where different categories of places form interaction patterns of cooperation, where mobile users move from one to the other. Competition, on the other hand, manifests in a more implicit manner in the network in two ways: first, retail facilities that are grouped in the same node (e.g. Bars) have to share customers that have been previously performing a different activity (e.g. going to a Restaurant) and second, through activities that do not share an edge in the network and as a result they do not interact with one another in terms of mobility patterns.

### 5.4.2 Network Properties

We next quantify the structure of these networks in terms of different network properties considering also different time intervals. For our two cities of comparison, we list the network metrics in Table 5.1 and use the following metrics: the average clustering coefficient, $< C >$, the average closeness centrality, $< C_c >$, and the modularity, $Q$. These metrics are defined in more detail in Section 2.3.1.

We compared our network metrics to 3 random baselines: an Erdös-Rényi model[50], a Barabási-Albert[4] model and a configuration model[114] which maintains the same degree distribution. Table 5.1 shows the comparison with the Barabasi-Albert model. However, similar results and significance were observed with the two other random models. This comparison provides an indication of how significant empirical observations are with respect to the random case. First we note that for all three metrics the real networks are very different from the corresponding null models. In general, high clustering coefficient and modularity together with a lower closeness centrality scores point to the tendency of local businesses to form significantly tight clusters that are well isolated from one another. Furthermore, we also investigated variations of these network properties for different periods of the day. We observed in some cases that the closeness centrality was higher in the evening relative to morning hours whereas the average clustering was lower in the evening. However, we could not notice significant changes at the network level. This could mean that some categories are less locally connected to each other during evening hours - potentially due to user routine.

Looking closer at the network modularity scores presented in Table 5.1 we note a clear partitioning of different categories into communities with scores around 0.6 for both cities compared to much smaller values $\approx 0.15$ for the null model. Modularity values increase in the evening in both cities. This translates to a stronger community structure, suggesting customers may be less likely to experience activities in different category communities than in the morning and confirm the dichotomies from previous sections such as *Coffee Shops vs. Pubs* in London and *Bars vs. French Restaurants* in Paris). Finally, the similarity in terms of network properties values between the two cities, as well as the prominent community structure in both suggest that the hypothesis that the organization of the retail business ecosystem is similar across cities is a plausible one. This is true to a certain degree nonetheless, as variations are also noted due to apparent cultural differences.

In this section, we highlighted the dominance of community structure in urban activity networks. This observation suggests that categories gain (or lose) attractiveness as a result of other activities around them. It further raises the question of how businesses affect each other. In the next section, we look at the evolution of the number of check-ins and we introduce our deep learning architecture to model the demand of venues over time to shed light on the aforementioned questions.

### 5.4.3 Temporal Trends

In this section, we investigate how the number of venue check-ins changes from month to month. To do so, for each venue we compute relative change from one month to the next and calculate the standard deviation (STD) of the sequence of changes. For example, given a venue with demand data over a 12-month time-span. This data contains a sequence of 11 changes in demand. For each change, we calculate the relative change in the number of check-ins ($(C_{t+1} - C_t)/C_t$, where $C_t$ is the number of check-ins at time $t$). If we assumed that the number of check-ins remains constant or had very few changes from month to month this

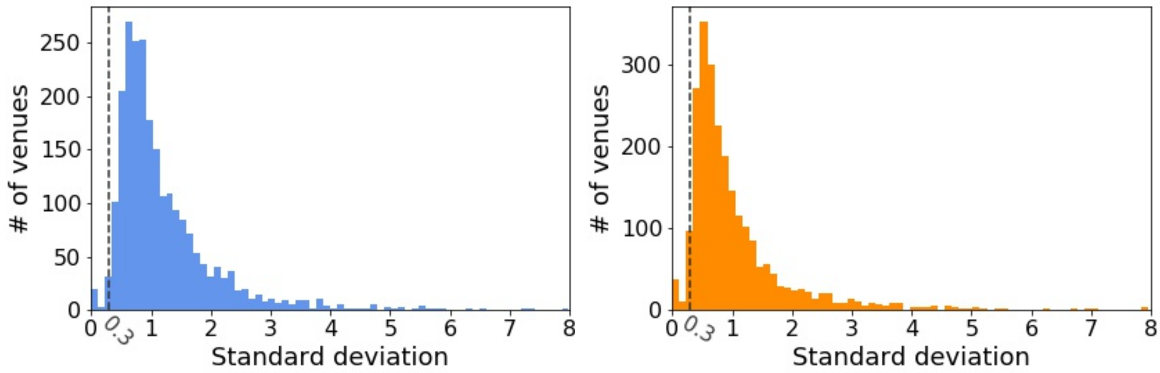



Fig. 5.2 Histogram of the standard deviation of the relative change in checkins for Paris (blue) and London (orange). For each venue we computed the monthly change in checkins relative to the prior month. For instance, a venue doubling its number of checkins from a month to another would be a change of 1. This gave for each venue a sequence of changes. We then computed the standard deviation of each sequence to evaluate the amplitude of changes.

would return a low standard deviation of the relative changes. However, Figure 5.2 rejects this assumption. Figure 5.2 shows the histogram of sequence standard deviations for the city of Paris (in blue) and London (in orange). A venue in the city corresponds to a unit point in the histogram. In order to remove noise due to venues with a low number of check-ins, we filtered out venues that did not reach an average of 20 check-ins per month. We observe in Figure 5.2 that for both cities most venues have an STD over 0.3, with the histogram peak around 0.5. This highlights that individual venues vary significantly from month to month. This motivates our prediction task in the subsequent section.

## 5.5 Methodology

We next describe our model and give an overview of our deep learning approach and architecture. Our model incorporates both spatial and topological features into an enhanced representation which is then fed into a temporal model to predict the next time step. This more complex representation includes attributes of both the spatial and topological domain which enables better predictions. Below, we give an overview of GCNs and LSTMs and describe how our architecture incorporates both.

### 5.5.1 Graph Convolution Networks

Traditional convolutional neural networks (CNNs) are used to represent spatial relationships in Euclidean space for applications to 2D matrices, grid-like structures, or images, amongst

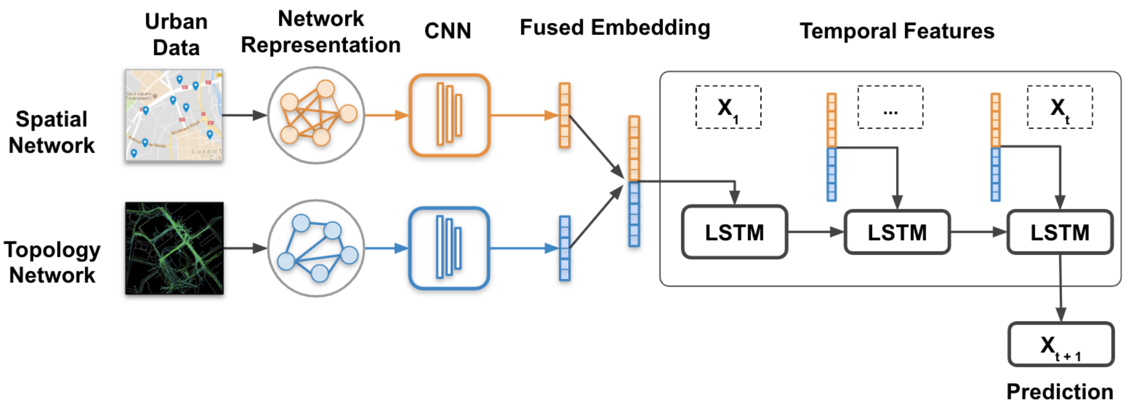



Fig. 5.3 An overview of the architecture to build spatial and topological representation of the venue graph which is then fed into a temporal model. Fed into the LSTM is the input and the previous hidden state.

others [98]. Recent research has worked to extend the principles of a convolutional filter to other applications including graph-structured data. A graph convolutional network (GCN) is a generalization of a CNN and can be applied to capture the topological relationships present between venues urban environments. GCNs harness spectral graph theory by using a spectral filter which is based on the graph Laplacian matrix. Spectral-based GCNs have been increasingly used in numerous applications that involve graph-structured data. These application include image classification, population-based disease predictions, and recommendation systems [135, 125, 179]. To date, the only application of GCNs to urban environments has been that of traffic prediction. [186]. These works represent the network of connected streets in a city as a set of connected nodes in a graph. However, in addition to traffic prediction, GCNs have tremendous potential to model the dynamics and connectivity of venues in urban environments. In this application, the spectral filter captures network features of nodes in the graph as well as in their neighborhood. A GCN model is then built by stacking convolutional layers. For nodes in our network, we define a k-th order neighborhood of a venue $v_i$ as follows:

$$NB_i = \{v_i \in V \ | \ d(v_i, v_j) \leq k\} \tag{5.1}$$

where $d(v_i, v_j)$ is the distance in hops from venue $v_i$ to venue $v_j$. A one-hop neighborhood is defined as the adjacency matrix. Further, the k-th hop adjacency matrix is equal to the k-th product of A. The k-th order adjacency matrix is defined as follows:

$$A^k = \prod_{i=1}^{k} A + I \tag{5.2}$$

A 1-hop graph convolution is defined as follows:

$$f(X,A) = \hat{A}XW \tag{5.3}$$

Where X is our feature matrix, W is the weight matrix and

$$\hat{A} = \widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}} \tag{5.4}$$

Where $\widetilde{D}$ is the degree matrix. We employ two GCNs that build spatial and topological characteristics respectively. Spatial features represent the location of venues relative to others and topological features represent the connectivity of a venue to other venues. These are then fed into a long short-term memory (LSTM) network which is described in detail next.

### 5.5.2 LSTM

At each time-step, the network represented is fed into an LSTM which acquires the dynamics dependence and trends over time [74]. Generally, an LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values of a set time interval and the gates determine the flow of information into the cell and out into the subsequent cell. LSTM models have widely been used on time series data across numerous domains. Within the context of urban environments, LSTMs have been used to model population mobility flow, bicycle availability, and traffic congestion. Very few works have used LSTMs in conjunction with GCNs. A previous work by Zhao et al. [186] built a deep learning model that fed a GCN into a gated recurrent unit (GRU) to better predict traffic on roads in a city. GRUs are very similar to LSTMs in their architecture. Beyond this work by Zhao et al., there a notable gap in the literature in the application of temporal GCNs for urban applications.

Our model aims to predict the number of check-ins to a venue at the subsequent time step. We showed above in Section 5.4.3 that the number of check-ins varies from month to month at the venue level and is not consistently stable. We make the assumption that these fluctuations from month to month are based on numerous factors including seasonality, neighborhood growth, and real-time events. These fluctuations could also be an artifact of our dataset which may not represent a significant sample of users. However, as we see below our results suggest that our features capture some variance. These trends suggest that a prediction task for demand must incorporate data from not only previous time-steps but also from other features such as the neighborhood of a venue. We suggest that an LSTM is capable of incorporating these attributes as it is able to remember previous trends of venue demand when trained on data that incorporates numerous features relating to the venue of interest.

### 5.5.3 Model Architecture

The aim of our model is to predict the demand for a given venue in the subsequent month. Our analysis below focuses on the cities of London and Paris however the methodology can be applied more broadly to other cities worldwide. For our prediction task, we use check-ins to a venue as a proxy for the popularity of that venue. In Section 5.4.2, we built a graph for each city of categories to characterize relationships between them. In this section, we build a network of venues, each venue is represented by a node. An edge is drawn between two nodes if at least one transition occurs between the pair of venues. We hypothesize that venue demand is a nonlinear combination of three sets of attributes: physical location in space, venue connectivity, and temporal trends. Our model incorporates and fuses spatial and topological data which are then used to train a dynamic model. Spatial data represents the physical location of venues and their distance between each other. Topological data represents the interactivity and interconnectedness of different venues. We describe both in more detail below.

**Pre-processing:** To eliminate noise and venues that may not have significant use, we set a threshold of at least a mean of 20 check-ins per month. We do not consider venues that do not meet this threshold.

**Implementation:** The code for the work in this chapter was implemented in Python. The model was built using Tensorflow and Keras, popular and open-source machine learning libraries [1, 34]. Additionally, open-source code written by Zhao et. al [186] was used to create the initial architecture for the model[3]; it was subsequently updated and customized.

**Topological Graph:**

We build a weighted graph $G = (V, E_t)$ to describe the topological structure of venues. The temporal graph consists of a series of graphs over $T$ time steps $G = \{G_1, G_2, ..., G_T\}$. The set of nodes $V = v_1, v_2, ..., v_N$ represent the set of individual venues in our graph. $E_t$ represents the number of transitions between two given venues at time step $t$. When feeding our graph into our model, we represent our graph with two data structures: an adjacency matrix and a feature matrix. The input feature matrix is expressed as $X \in \mathbb{R}^{NxF}$ where $N$ is the number of

[3]https://github.com/lehaifeng/T-GCN

nodes and $F$ is the number of features for each node. The number of node attribute features is equal to the length of the historical time series. The adjacency matrix $A \in \mathbb{R}^{NxN}$ represents the connectivity of the graph $G$. The adjacency matrix contains elements of 0 or 1. The value is 1 if there is transition between two venues in that time step and 0 otherwise. The adjacency matrix is updated at each time step. For our model, each node in our network is a venue and features represent the number of check-ins to each venue at that time step and the adjacency matrix represents whether two venues are connected. We set the order of the adjacency matrix to be $k = 1$ and vary the order below to examine the impact of adjacency order. We found the $k = 1$ was the most accurate value for the model, described below. This topological graph, built from Foursquare check-ins, gives us an indication of how connected a venue is to other venues in a city. This can help inform future demand of a venue as the neighborhood attributes of a venue often relate directly to its own characteristics. However, it does not directly consider the physical location of that venue in space. To this end, we also build a spatial graph of the same set of venues $V$.

**Spatial Graph:**

While the topological graph contains aspects of location, in which venues with a close proximity to each other are likely to have higher rates of movement between the two, it does not consider explicitly the physical location. As such we build a spatial graph $G_s = (V, E_t)$ in which nodes are similarly venues and edges represent the Euclidean distance between the coordinates of a pair of venues. This graph is static as it does not contain temporal features. We calculate the closeness centrality for each node in the graph. As closeness centrality represents how near a given node is to other nodes in the graph, it represents that node's location in physical space. As with the topological graph, we represent our spatial features with two structures: an adjacency matrix and a feature matrix. The input feature matrix is expressed as $X \in \mathbb{R}^{Nx1}$ where $N$ is the number of nodes. The feature for each venue is its closeness centrality. The adjacency matrix $A \in \mathbb{R}^{NxN}$ represents the connectivity of the graph $G_s$. The adjacency matrix contains elements of 0 or 1. The value is 1 if the distance between the two venues was less than 1km. The distance of 1km was determined empirically. This network represents the spatial connectivity of the venues.

## 5.6 Results & Discussion

We build our temporal graph using data beginning on January 1 2012 until December 31 2013. Each monthly graph in $G$ is built using the check-ins at that month. The data was split such that 60% was used as the training set, 20% was used as the validation set, and 20% as

| | London | | Paris | |
|---|---|---|---|---|
| Model | RMSE | MAE | RMSE | MAE |
| ARIMA | 5.9123 | 37.5049 | 7.4211 | 49.9920 |
| SVR | 4.0780 | 33.0480 | 5.5088 | 44.8892 |
| LSTM | 2.4504 | 27.0280 | 3.1793 | 34.4486 |
| Topological TGCN | 2.0731 | 26.9817 | 2.9535 | 34.4081 |
| Topological & Spatial TGCN | 1.7508 | 25.3600 | 2.7387 | 29.8400 |

Table 5.2 Performance comparison of our model relative to baselines in both cities.

the testing set. We set the time steps of the input to be 4 months and run a sliding window through the series. The output of the model, which the model aims to predict, is the next subsequent time-step of the input sequence.

The performance of our model and that of our baselines is evaluated by two metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). We train our model for a maximum of 500 epochs, using Adam optimization [94] with an initial learning rate of $10^5$. We employ early stopping patience of 5, where training is stopped if the loss does not decrease by a threshold of 0.00001 for 5 consecutive epochs. We use a hidden layer of the same size as the number of nodes in the venue network graph (N = 568 for London and N = 482 for Paris). We set the size of hops in the graph convolution to $K = 1$ after experimentation which showed that increasing the value of $K$ decreases the accuracy of the results. This suggests that the direct neighborhood of venues play a larger role in their demand dynamics than venues that are more distant. All of our models were implemented using PyTorch 0.3.1. We compare our model (en-TGC) with the following baseline models:

1. ARIMA: Auto-Regressive Integrated Moving Average [9]. The ARIMA baseline uses solely historical demand fit into a parametric model to predict future demand.

2. SVR: Support Vector Regression [163]. The SVR trains on historical demand data to construct a hyperplane to learn a relationship between the inputs and outputs. We use a linear kernel.

3. LSTM: Long Short-Term Memory recurrent neural network [74]. The LSTM network is a deep learning model which trains on past demand data for individual venues.

4. Topological GCN: The GCN builds a graphical representation of the network using only Foursquare transitions and uses a spectral filter to learn features.

Table 5.2 includes the results of our model relative to other baselines enumerated above in both Paris and London. Our proposed architecture outperforms other models in both

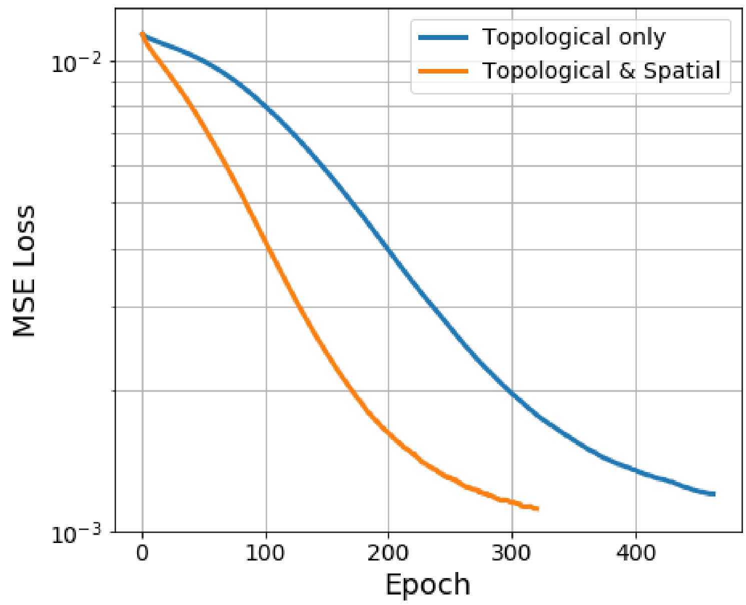



Fig. 5.4 Validation loss over training epoch for London.

RMSE and MAE in both cities. We see that ARIMA has a much higher error rate than machine learning methods, highlighting that venue demand prediction is a complex task that involves multiple factors and that the historical demand trends of a venue are likely not sufficient for future predictions. Our model reduces the RSME by 70.39% relative to the ARIMA prediction in London and by 63.1% in Paris. Further, we see that our model performs better than a simple LSTM. Our model reduces the RSME by 28% relative to an LSTM in London and 13% in Paris. These trends demonstrate the value of network features learned from the spectral graph convolution. It highlights the power of graphical neural networks to model the dynamics of urban environments. We also see a decrease in RSME when integrating spatial features into the model. In London RSME decreases by $\approx$ 15% and in Paris it decreases by $\approx$ 7%. This demonstrates the benefit of integrating spatial features into the representation in the deep learning model. Our experiments demonstrate that our model can learn spatio-temporal trends of venue demand and consistently outperform baseline models. To the best of our knowledge, this is the first application of temporal graph convolutional neural networks applied to mode business dynamics. We see similar trends in the decrease in MAE by our models. An MAE of $\approx$ 25 in London and $\approx$ 29 in Paris are low given the difficultly of the prediction task at hand. With a minimum of at least 20 check-ins per month, venue monthly check-ins in our dataset range from 20 to over 1000. With this context in mind, an MAE of $\approx$ 25 in London and $\approx$ 29 in Paris suggests our model is performing extremely well given the inherent complexity of the task.

We next compare the training efficiency of our model with both topological and spatial features to one in which only topological features are used. Figure 5.4 shows the validation loss across training epoch for both models. The max training epoch was set to 500. However, as we employed early stopping in the training process, the number of training epochs is less than 500. Firstly, we see the combined model converges more quickly than the model with only topological features. This suggests that the fusion of both features enables the model to learn more trends more quickly. Additionally, the loss of the combined model decreases more quickly. We see similar trends for loss when training the models in Paris.

In summary, our model combines spatial and topological features beats all of our baselines reducing the RMSE by up to 28% relative to a deep-learning baseline. Additionally, our model with both feature types converges more quickly than one without, suggesting it is able to learn trends of venue demand at a faster rate.

## 5.7 Conclusion

Our methodology highlights the power of graph convolutional networks in building prediction models within the context of urban cities. This can broadly be applied to many systems in which interactions between agents must be taken into account. We began in Section 5.4 of this work by detecting patterns of cooperation in urban networks and quantifying similarities and differences in network structure across cities. Next in Section 5.4.3, we described the high variability in monthly check-in data, motivating our subsequent prediction task, and also illuminating the inherent complexity and challenges in the task.

Using spatial features in conjunction with topological network measures, we developed a temporal machine learning model to predict future business demand in Section 5.5. The novelty of our approach in methodological terms stems from the use of temporal GCNs with a spatial and topological representation of venues in a city to tackle open modern urban questions.

Our model and results can support policymakers, business owners, and urban planners as they have the potential to pave the way for the development of sophisticated models describing urban neighborhoods and predicting future growth trends for venues in a given area. Our methodologies could also urban planners in better understanding conditions for growth for venues and working to determine the optimal conditions for establishing a venue and more generally new urban facilities. This chapter is a first step in harnessing dynamic graph convolutional networks to working to model urban dynamics.

# Chapter 6

# Reflections & Outlook

This dissertation has taken a step towards building a deeper understanding of and prediction tools for venue dynamics in urban environments worldwide. The preceding chapters presented several approaches to modeling cities and substantiated the thesis that *the spatio-temporal study of urban areas through location-based analytics and complex network theory can advance our understanding of human mobility and enable us to model venue dynamics*.

The research presented in this dissertation is a product of advancements in the availability of geo-spatial data, the computational efficiency of cloud computing, and the breadth of machine learning capabilities. These developments have created numerous opportunities to pursue research questions. For example, although human mobility has been explored and modeled for decades, the emergence of new sources of geographic, socioeconomic, and social data have enabled large-scale studies of urban environments through novel data-driven modeling techniques and contextual domain-specific theories. These empirical computational methodologies have the capacity to quantify our understanding of real-world systems. Their insights constitute a novel resource for the development of mobile applications and services. This dissertation has sought to advance the field of urban computing using novel machine learning architectures and techniques and demonstrate the value in using these datasets to model cities to predict, characterize, and quantify trends. In this chapter, we summarize the major contributions of this dissertation and provide an overview of future directions of research in this space.

## 6.1 Thesis Summary & Contributions

We will next outline our research findings and summarize the key contributions of this dissertation.

**Identifying temporal similarity as a key metric for predicting new venue demand**: In Chapter 3, we introduced a prediction framework that used the characteristic temporal profile of neighborhoods in a city in conjunction with k-nearest neighbor metrics to capture similarities among urban regions. We used a Gaussian Process model to forecast the weekly popularity dynamics of new venues using spatial and temporal features. Our models demonstrated how a particular venue influences the overall temporal profile of the neighborhood in which it is located.

**Creating an on-line prediction task for new venue popularity**: In Chapter 3, we also experimented with an on-line prediction task that predicted relative changes in popularity with respect to historic patterns of a venue. This model presented a novel approach to venue demand predictions.

**Determining factors most influential in the success or failure of a business**: In Chapter 4, we built models to predict whether businesses will close down within a given period. Through careful statistical analysis of Foursquare and taxi mobility data, we uncovered a set of discriminative features, belonging to the neighborhood's static characteristics, the venue-specific customer visit dynamics, and the neighborhood's mobility dynamics. We demonstrated how the survival of such a retail outlet is correlated with not only venue-specific characteristics but also broader neighborhood-level effects.

**Exploring business survival across different venue characteristics**: In addition to creating a set of discriminative features that impacted the likelihood of survival of retail businesses, we examined how these features varied across different geographies and venue maturities. Our research found that new venues were susceptible to higher variations in their causes of failure, likely as a result of the diverse set of challenges new businesses face when opening. Although our results showed that the impact of features varied from city to city, we also noted that certain features were consistently significant across many or all our cities. Features such as *Temporal Popularity Skew*, *Temporal alignment to Competitors*, and *Distance of Travel From* were significant for many of the cities, emphasizing that the dynamics of urban environments have a strong influence on venue demand.

**Quantifying patterns of cooperation and competition across cities**: In Chapter 5, we modeled businesses in a city as connected networks of nodes belonging to different activity types. We saw a strong community structure emerge in local retail networks and that cooperative effects appeared naturally, reflecting interactions of nodes within and across communities. Our analysis examined differences in network metrics across two cities, London and Paris, and found both an underlying common structure as well as notable cultural distinctions.

**Harnessing network frameworks for venue demand prediction**: Our analysis in Chapter 5 employed a spatio-topological temporal model to predict the demand of venues at the subsequent time-step. This work demonstrated how complex networks can be used in dynamic deep learning models. Further, our approach converged more quickly than alternatives suggesting it was able to adeptly learn trends in venue popularity using both topological and spatial features.

## 6.2 Directions For Future Research and Outlook

The research in this dissertation lies at the intersection of machine learning, network science, and urban analytics, a space with many valuable applications of the work. We next explore opportunities for future work that arise from the research pursued in this dissertation.

Our analysis in Chapter 3 predicted the demand characteristics of newly opened businesses. To avoid data sparsity problems, we narrowed our venues of interest to include venues that had at least 100 check-ins over the entire duration of our dataset. Further research efforts could study a broader range of venues since less popular venues could have different characteristics or present unconventional temporal properties. Additionally, further work in the exploration of our analysis could explore the impact of our modeling on different cities around the world to understand their regularity and temporal trends through the application of the methodology discussed in the chapter. The framework proposed can easily be applied to different cities since it is not based on any assumptions regarding the spatial and urban context.

The research pursued in Chapter 4 examined the factors influencing the success or failure of a business. The methodology developed a non-continuous predictor: at it is presented, we computed a variety of features using Foursquare and mobility data and then predict a venue's likelihood of survival over the *next* 6 months. Implicit in our approach is the belief that the majority of our features (such as the hourly temporal profile of localities) are stable, and do not vary significantly with time. As the next step, it would be useful to develop a streaming predictor–one which continually updates the survivability likelihood as time progresses, by appropriately incorporating up-to-date feature values. Additionally, our analysis in the chapter found a set of features that impacted the survival of a business. Building upon this work, further analysis could help to better understand how (i) such features vary across different neighborhoods, based on other factors such as resident demographics, (ii) additional features, from other information sources (e.g., electricity or water consumption) help enhance the accuracy, and (iii) how the classification accuracy will vary with changes to the prediction time horizon.

Lastly, our research in Chapter 5 introduced a novel deep learning framework to model the future popularity of existing businesses. The research examined only the dynamics of businesses that meet our minimum threshold of busyness. While this helps us to focus on only venues with meaningful data, it eliminates venues that may still have interesting trends, online and offline. However, this present study sets the frame for further general studies. Additionally, in this work, we conducted a preliminary examination of the variations in network trends across two cities. Future work could expand upon this to explore the duality between general network trends and cultural consumer idiosyncrasies across cities. It could also incorporate more attributes of the road network of the city, such as traffic congestion, busyness, and speed.

More broadly, there are many opportunities for future work in urban computing. Our research in Chapter 4 established that the combination of social media and urban mobility data together provided a high predictive power. Similarly, future work in urban computing can benefit from the power of many types of data sources in informing a prediction. For example, work combining spatial, temporal, and mobility data in conjunction with social data could yield new insights and inform novel prediction tasks. Additionally, an inherent limitation in LBSN data is that it only represents the online experience of individuals which may differ from their offline experiences. Data from offline experiences can be gathered in numerous ways (e.g. passive sensing modalities). Future work examining these characteristics may reveal new takeaways for businesses. Lastly, prediction models rarely take into account unprecedented events. COVID-19, for example, impacted business around the world and lead to a significant decline in retail and Food and Beverage demand [112]. Researching the impact of COVID-19, the recovery of businesses, and changes in customer visitation patterns will be of paramount importance in the years to come.

To summarize, the research presented in the preceding chapters worked to demonstrate that innovative machine learning architectures and techniques can advance the field of urban computing. This dissertation has put forward research that takes a step towards better understanding urban movements and developing effective frameworks for business decision systems. Our findings offer new insights into human mobility patterns and their impact on business dynamics. They also open new research directions in the prediction of spatio-temporal urban mobility trends. We hope our contributions encourage researchers to develop urban computing methodologies that incorporate deep learning architectures, complex network metrics, and contextual insights from cities.

# References


[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[2] Adminstration, U. S. B. (2018). Apply for licenses and permits. https://www.sba.gov/business-guide/launch-your-business/apply-licenses-and-permits.

[3] Aguirre, A. J. L., emanjski, I., and Gautama, S. (2016). Forecasting travel behaviour from crowdsourced data with machine learning based model.

[4] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

[5] Alessandretti, L., Karsai, M., and Gauvin, L. (2016). User-based representation of time-resolved multimodal public transportation networks. *Royal Society Open Science*, 3(7).

[6] Aleta, A., Meloni, S., and Moreno, Y. (2017). A Multilayer perspective for the analysis of urban transportation systems. *Scientific Reports*, 7:44359.

[7] AlSonosy, O., Rady, S., Badr, N., and Hashem, M. (2018). *Business Behavior Predictions Using Location Based Social Networks in Smart Cities*, pages 105–122. Springer Singapore, Singapore.

[8] Applebaum, W. (1966). Methods for determining store trade areas, market penetration, and potential sales. *Journal of marketing Research*, pages 127–141.

[9] Asteriou, D. and Hall, S. (2016). *ARIMA Models and the Box-Jenkins Methodology*, pages 275–296.

[10] Authority, G. L. (2014). LSOA Atlas. https://data.london.gov.uk/dataset/lsoa-atlas.

[11] Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., and Shah, N. H. (2017). Improving palliative care with deep learning. *arXiv preprint arXiv:1711.06402*.

[12] Balan, R. K., Nguyen, K. X., and Jiang, L. (2011). Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th International conference on Mobile systems, Applications, and Services (MobiSys'11)*, pages 99–112. ACM.

[13] Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge University Press, Cambridge.

[14] Barthelemy, M. (2017). *Morphogenesis of Spatial Networks*.

[15] Barthélemy, M. (2010). Spatial networks. *PHYSICS REPORTS*, page 2011.

[16] Battaglia, P., Pascanu, R., Lai, M., Rezende, D., and Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics.

[17] Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Physical Review E*, 89(3).

[18] Batty, M. (2007). Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals.

[19] Baumann, P., Kleiminger, W., and Santini, S. (2013). The influence of temporal and spatial features on the performance of next-place prediction algorithms. pages 449–458.

[20] Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82.

[21] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

[22] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[23] Brockmann D, Hufnagel L, G. T. (2006). The scaling laws of human travel. *Nature*, 439.

[24] Buono, C., Alvarez-Zuzek, L. G., Macri, P. A., and Braunstein, L. A. (2014). Epidemics in partially overlapped multiplex networks. *PLOS ONE*, 9(3):1–5.

[25] Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151.

[26] Capponi, A., Vitello, P., Fiandrino, C., Cantelmo, G., Kliazovich, D., Sorger, U., and Bouvry, P. (2019). Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6.

[27] Chai, D., Wang, L., and Yang, Q. (2018). Bike flow prediction with multi-graph convolutional networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '18, page 397–400, New York, NY, USA. Association for Computing Machinery.

[28] Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., and Scott, J. (2007). Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620.

[29] Chen, B., Hall, D., and DB, C. (2005). Wiring optimization can relate neuronal structure and function. *PNAS*.

[30] Chen, S., Claramunt, C., and Ray, C. (2014). A spatio-temporal modelling approach for the study of the connectivity and accessibility of the Guangzhou metropolitan network. *Journal of Transport Geography*, 36:12 – 23.

[31] Cheng, E. W. and Li, H. (2004). Exploring quantitative methods for project location selection. *Building and Environment*, 39(12):1467–1476.

[32] Chiang, M.-F., Hoang, T.-A., and Lim, E.-P. (2015). Where are the passengers?: A grid-based gaussian mixture model for taxi bookings. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, pages 32:1–32:10, New York, NY, USA. ACM.

[33] Cho, E., Myers, S., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. pages 1082–1090.

[34] Chollet, F. et al. (2015). Keras.

[35] Chou, J.-S., Ho, C.-C., and Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics*, 44:57–75.

[36] Crucitti, P., Latora, V., and Porta, S. (2006). Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125.

[37] Daggitt, M., Noulas, A., Shaw, B., and Mascolo, C. (2016). Tracking urban activity growth globally with big location data. *Royal Society Open Science*, 3(4).

[38] Dai, X., Sun, L., and Xu, Y. (2018). Short-term origin-destination based metro flow prediction with probabilistic model selection approach. *Journal of Advanced Transportation*, 2018.

[39] Daily, B. N. (2018). How and why businesses collect consumer data.

[40] Daqing, L., Fu, B., Wang, Y., Lu, G., Berezin, Y., Stanley, H., and Havlin, S. (2014). Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proceedings of the National Academy of Sciences of the United States of America*, 112.

[41] De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015). Structural reducibility of multilayer networks. *Nature Communications*, 6:6864.

[42] Django Software Foundation (2019). Django. https://djangoproject.com.

[43] Dong, C., Xiong, Z., Shao, C., and Zhang, H. (2015). A spatial–temporal-based state space approach for freeway network traffic flow modelling and prediction. *Transportmetrica A: Transport Science*, 11(7):547–560.

[44] D'Silva, K., Jayarajah, K., Noulas, A., Mascolo, C., and Misra, A. (2018a). The role of urban mobility in retail business survival. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):100:1–100:22.

[45] D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C., and Sklar, M. (2017). If I build it, will they come? Predicting new venue visitation patterns through mobility data. In *Proceedings of the ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*.

[46] D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C., and Sklar, M. (2018b). Predicting the temporal activity patterns of new venues. *EPJ Data Science*, 7(1):13.

[47] Eagle, N. and Pentland, A. S. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

[48] Eckmann, J.-P., Moses, E., and Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101:14333–7.

[49] Eiselt, H. A. and Laporte, G. (1989). Competitive spatial models. *European Journal of Operational Research*, 39(3):231–242.

[50] Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297.

[51] Facebook (2017). Discover great places in every city. https://www.facebook.com/places/.

[52] Farseev, A. and Chua, T. (2017). Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 95–101.

[53] Fletcher, P., Hoyle, H., and Patty, C. W. (1991). *Foundations of Discrete Mathematics*, page 463.

[54] for National Statistics, O. (2011). Number of Electoral Wards/Divisions in the United Kingdom. https://www.ons.gov.uk/.

[55] Fortuna, R. and Marovt, U. (2016). Link prediction in foursquare network.

[56] Foursquare (2020). Get details of a venue. www.developer.foursquare.com/docs/api/venues/details.

[57] França, U., Sayama, H., McSwiggen, C., Daneshvar, R., and Bar-Yam, Y. (2016). Visualizing the "heartbeat" of a city with tweets. *Complexity*, 21(6):280–287.

[58] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[59] Fu, Y., Ge, Y., Zheng, Y., Yao, Z., Liu, Y., Xiong, H., and Yuan, J. (2014). Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM'14)*, pages 120–129. IEEE.

[60] Gallotti, R. and Barthelemy, M. (2015). The multilayer temporal network of public transport in great britain. *Nature*, 2.

[61] Georgiev, P., Noulas, A., and Mascolo, C. (2014). Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*

[62] Gepp, A., Kumar, K., and Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting*, 29(6):536–555.

[63] Ghosh, A. and Craig, C. S. (1983). Formulating retail location strategy in a changing environment. *The Journal of Marketing*, pages 56–68.

[64] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, pages 695–719.

[65] Gibson, M. and Pullen, M. (1972). Retail turnover in the east midlands: A regional application of a gravity model. *Regional Studies*, 6(2):183–196.

[66] Giridhar, P., Amin, M. T., Abdelzaher, T., Wang, D., Kaplan, L., George, J., and Ganti, R. (2017). Clarisense+. *Pervasive Mob. Comput.*, 41(C):381–396.

[67] González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

[68] Google (2017). Popular times and visit duration. https://support.google.com/business/answer/6263531.

[69] Granell, C. and Ostermann, F. O. (2016). Beyond data collection: Objectives and methods of research using vgi and geo-social media for disaster management. *Computers, Environment and Urban Systems*, 59:231–243.

[70] H. Davenport, T. (2014). How strategists use “big data” to support internal business decisions, discovery and production. *Strategy  Leadership*, 42(4):45–50.

[71] Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs.

[72] Hecht, B. and Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 14:197–205.

[73] Hidalgo, C. A. and Castanẽr, E. E. (2015). Do we need another coffee house? the amenity space and the evolution of neighborhoods.

[74] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[75] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

[76] Hristova, D., Noulas, A., Brown, C., Musolesi, M., and Mascolo, C. (2015). A multilayer approach to multiplexity and link prediction in online geo-social networks. *CoRR*, abs/1508.07876.

[77] Hristova, D., Panzarasa, P., and Mascolo, C. (2016a). Multilayer brokerage in geo-social networks.

[78] Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., and Mascolo, C. (2016b). Measuring urban social diversity using interconnected geo-social networks. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16.

[79] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.

[80] Hyman, M. and Steiner, S. (1996). The vignette method in business ethics research: Current uses and recommendations.

[81] IBM (2020). Benefits of cloud computing. www.ibm.com/uk-en/cloud/learn/benefits-of-cloud-computing.

[82] Jayarajah, K., Subbaraju, V., Weerakoon, D., Misra, A., La Thanh, T., and Athaide, N. (2017). Discovering anomalous events from urban informatics data. page 101900F.

[83] Jayasuriya, R. and Wodon, Q. (2020). Explaining country efficiency in improving health and education indicators: the role of urbanization.

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

[85] Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., and González, M. C. (2016). The timegeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences*, page 201524261.

[86] Johnson, N. E., Ianiuk, O., Cazap, D. D., Liu, L., Starobin, D., Dobler, G., and Ghandehari, M. (2017). Patterns of waste generation: A gradient boosting model for short-term waste prediction in new york city. *Waste management*, 62.

[87] Jones, C. and Volpe, E. (2011). Organizational identification: Extending our understanding of social identities through social networks. *Journal of Organizational Behavior*, 32:413 – 434.

[88] Jordahl, K. (2014). Geopandas: Python tools for geographic data. *URL: https://github.com/geopandas/geopandas*.

[89] Kaiser, E. J., Godschalk, D. R., and Chapin, F. S. (1995). Urban land use planning.

[90] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[91] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. (2013). Geospotting: 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*.

[92] Karsai, M., Perra, N., and Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific reports*, 4:4001.

[93] Keegan, S., O'Hare, G., and O'Grady, M. (2012). Retail in the digital city. *International Journal of e-Business Research*, 8:18–33.

[94] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

[95] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

[96] Kong, X., Gao, H., Alfarraj, O., Ni, Q., Zheng, C., and Shen, G. (2020). Huad: Hierarchical urban anomaly detection based on spatio-temporal data. *IEEE Access*, PP:1–1.

[97] Krause, J., Croft, D., and James, R. (2007). Social network theory in the behavioural sciences: Potential applications. *Behavioral Ecology and Sociobiology*, 62:15–27.

[98] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

[99] Kursa, M. B., Rudnicki, W. R., et al. (2010). Feature selection with the boruta package. *J Stat Softw*, 36(11):1–13.

[100] Kwon, Y. D., Chatzopoulos, D., ul Haq, E., Wong, R. C.-W., and Hui, P. (2019). Geolifecycle: User engagement of geographical exploration and churn prediction in lbsns. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3).

[101] Levine, R. and Norenzayan, A. (1999). Pace of life in 31 countries. *Journal Of Cross-Cultural Psychology*.

[102] Li, S. and Ma, Y. (2014). Urbanization, economic development and environmental change. *Sustainability*, 6.

[103] Li, Y., Zheng, Y., Zhang, H., and Chen, L. (2015). Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM SIGSPATIAL 2015.

[104] Liang, X., Shen, X., Feng, J., Lin, L., and Yan, S. (2016). Semantic object parsing with graph lstm.

[105] Lichman, M. and Smyth, P. (2014). Modeling human location data with mixtures of kernel densities. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[106] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1).

[107] Lin, J., Oentaryo, R. J., Lim, E.-P., Vu, C., Vu, A., Kwee, A. T., and Prasetyo, P. K. (2016). A business zone recommender system based on facebook and urban planning data. In *European Conference on Information Retrieval*, pages 641–647. Springer.

[108] Liu, K., Gao, S., and Lu, F. (2019). Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling. *Comput. Environ. Urban Syst.*, 74:50–61.

[109] Lu, X., Yu, Z., Sun, L., Liu, C., Xiong, H., and Guan, C. (2016). Characterizing the life cycle of point of interests using human mobility patterns.

[110] Lussier, R. (2005). A success versus failure prediction model for the real estate industry. *American Journal of Business*, 20:46–53.

[111] Marr, B. (2018). How much data do we create every day? The mind-blowing stats everyone should read. www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read.

[112] McKinsey (2020). COVID-19: Implications for business. https://www.mckinsey.com/business-functions/risk/our-insights/covid-19-implications-for-business.

[113] Misra, A., Jayarajah, K., Nayak, S., Prasetyo, P. K., and Lim, E.-p. (2014). Socio-physical analytics: Challenges & opportunities. In *Proceedings of the 2014 Workshop on Physical Analytics*, WPA '14.

[114] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms*, 6(2-3):161–180.

[115] Nations, U. (2018). Population division. The world's cities in 2018, data booklet. www.un.org/en/development/desa/population/index.asp.

[116] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

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

[118] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11).*

[119] Noulas, A., Shaw, B., Lambiotte, R., and Mascolo, C. (2015). Topological properties and temporal dynamics of place networks in urban environments. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion.

[120] O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

[121] Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

[122] Pan, B., Zheng, Y., Wilkie, D., and Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'13)*, pages 344–353. ACM.

[123] Pappalardo, L. and Simini, F. (2017). Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829.

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

[125] Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., and Rueckert, D. (2017). Spectral graph convolutions for population-based disease prediction. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, Cham. Springer International Publishing.

[126] Parsa, H., Gregory, A., and Terry, M. (2011a). Why do restaurants fail? part iii: An analysis of macro and micro factors. *Dick Pope Sr. Institute Publications*, 15.

[127] Parsa, H., van der Rest, J.-P. I., Smith, S. R., Parsa, R. A., and Bujisic, M. (2015). Why restaurants fail? part iv: The relationship between restaurant failures and demographic factors. *Cornell Hospitality Quarterly*, 56(1):80–90.

[128] Parsa, H. G., Self, J., Sydnor-Busso, S., and Yoon, H. J. (2011b). Why restaurants fail? part ii - the impact of affiliation, location, and size on restaurant failures: Results from a survival analysis. *Journal of Foodservice Business Research*, 14(4):360–379.

[129] Parsa, H. G., Self, J. T., Njite, D., and King, T. (2005). Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly*, 46(3):304–322.

[130] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[131] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*.

[132] Porta, S., Latora, V., Wang, F., Rueda, S., Strano, E., Scellato, S., Cardillo, A., Belli, E., Cardenas, F., Cormenzana, B., et al. (2012). Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488.

[133] Project, R. (2020). The R stats package. https://www.r-project.org/.

[134] Puissant, A., Rougier, S., and Stumpf, A. (2014). Object-oriented mapping of urban trees using random forest classifiers. *International Journal of Applied Earth Observation and Geoinformation*, 26:235–245.

[135] Qin, A., Shang, Z., Tian, J., Wang, Y., Zhang, T., and Tang, Y. (2018). Spectral-spatial graph convolutional networks for semisupervised hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, PP:1–5.

[136] Raconteur (2019). A day in data. www.raconteur.net/infographics/a-day-in-data.

[137] Rahimi, S. M., Far, B., and Wang, X. (2019). Behavior-based location recommendation on location-based social networks. *GeoInformatica*.

[138] Rajagopalan, H. K., Saydam, C., and Xiao, J. (2008). A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research*, 35(3):814–826.

[139] Rasmussen, C. E. and Williams, C. (2006). Gaussian processes for machine learning.

[140] Ratti, C., Frenchman, D., Pulselli, R. M., and Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

[141] Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., and Strogatz, S. (2010). Redrawing the map of great britain from a network of human interactions. *PloS one*, 5:e14248.

[142] Reades, J., Calabrese, F., Sevtsuk, A., and Ratti, C. (2007). Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3).

[143] Rhee, S., Seo, S., and Kim, S. (2018). Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. pages 3527–3534.

[144] Riketta, M. and Nienaber, S. (2007). *Multiple Identities and Work Motivation: The Role of Perceived Compatibility between Nested Organizational Units*.

[145] ROCR (2020). ROCR: Visualizing the performance of scoring classifiers. www.cran.r-project.org/web/packages/ROCR/index.html.

[146] Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS ONE*, 6(1):e15923.

[147] Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., and Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 111(37):13290–13294.

[148] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009a). Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20(1):81–102.

[149] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009b). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

[150] Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A. (2011a). Nextplace: A spatio-temporal prediction framework for pervasive systems. volume 6696, pages 152–169.

[151] Scellato, S., Noulas, A., and Mascolo, C. (2011b). Exploiting place features in link prediction on location-based social networks. KDD ’11. ACM.

[152] Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.

[153] Scikit-learn (2020). scikit-learn machine learning in python. www.scikit-learn.org/stable.

[154] Shaw, B., Shea, J., Sinha, S., and Hogue, A. (2013). Learning to rank for spatiotemporal search. In *Proceedings of the 6th ACM international conference on Web Search and Data Mining (WSDM'13)*, pages 717–726. ACM.

[155] Sheldon, A. L. (1969). Equitability indices: Dependence on the species count. *Ecology*, 50(3):466–467.

[156] Sheu, J.-B. (2007). An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research Part E: Logistics and Transportation Review*, 43(6):687–709.

[157] Shopify (2020). Shopify. www.shopify.com.sg.

[158] Silva, T. H., de Melo, P. O., Almeida, J., Musolesi, M., and Loureiro, A. (2014). You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in Foursquare. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*.

[159] Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., and Loureiro, A. A. (2013). A comparison of Foursquare and instagram to the study of city dynamics and urban social behavior. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 4. ACM.

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

[161] Singh, V. K., Gao, M., and Jain, R. (2010). Social pixels: Genesis and evaluation. In *Proceedings of the 18th ACM International Conference on Multimedia*.

[162] Smith, C., Quercia, D., and Capra, L. (2013). Finger on the pulse: identifying deprivation using transit flow analysis. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work (CSCW'13)*. ACM.

[163] Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3).

[164] Solé-Ribalta, A., Gomez, S., and Arenas, A. (2016). A model to identify urban traffic congestion hotspots in complex networks. *Royal Society open science*, 3.

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

[166] Stella, M., Beckage, N., and Brede, M. (2016). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, 7.

[167] Sternquist, B. (1997). International expansion of US retailers. *International Journal of Retail & Distribution Management*, 25(8).

[168] Szell, M., Lambiotte, R., and Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641.

[169] Tang, J., Musolesi, M., Mascolo, C., and V, L. (2009). Temporal distance metrics for social network analysis. *Proc. 2nd Workshop on Online Social Networks*, pages 31–36.

[170] Tricentis (2020).

[171] Täuscher, K. and Laudien, S. M. (2018). Understanding platform business models: A mixed methods study of marketplaces. *European Management Journal*, 36(3):319 – 329.

[172] Vermeiren, K., Van Rompaey, A., Loopmans, M., Eria, S., and Mukwaya, P. (2012). Urban growth of kampala, uganda: Pattern analysis and scenario development. *Landscape and Urban Planning*, 106:199–206.

[173] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

[174] Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. pages 1100–1108.

[175] Wang, L., Gopal, R., Shankar, R., and Pancras, J. (2015). On the brink: Predicting business failure with mobile location-based checkins. *Decision Support Systems*, 76.

[176] Weber, H. (2015). Foursquare by the numbers.

[177] Williams, M. J. and Musolesi, M. (2016). Spatio-temporal networks: reachability, centrality and robustness. *Royal Society Open Science*, 3(6).

[178] Yang, D., Qu, B., Yang, J., and Cudre-Mauroux, P. (2019). Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach. In *The World Wide Web Conference*, WWW '19, page 2147–2157, New York, NY, USA. Association for Computing Machinery.

[179] Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18. ACM.

[180] Yu, B., Yin, H., and Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

[181] Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1):22–32.

[182] Zarsky, T. Z. (2003). "Mine your own business!": Making the case for the implications of the data mining of personal information in the forum of public opinion.

[183] Zhang, H., Zheng, Y., and Yu, Y. (2018). Detecting urban anomalies using multiple spatio-temporal data sources. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1).

[184] Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5165–5175. Curran Associates, Inc.

[185] Zhang, X. Q. (2016). The trends, promises and challenges of urbanisation in the world. *Habitat International*, 54:241–252.

[186] Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., and Li, H. (2019). T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*.

[187] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3).

[188] Zhong, C., Arisona, S. M., Huang, X., Batty, M., and Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.*, 28(11).

[189] Zhong, C., Huang, X., Arisona, S., and Schmitt, G. (2013). Identifying spatial structure of urban functional centers using travel survey data: A case study of singapore. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, COMP '13, pages 28:28–28:33.

[190] Zhou, M. (1992). *Chinatown: The Socioeconomic Potential of an Urban Enclave*. Temple University Press.

[191] Zhou, X., Hristova, D., Noulas, A., Mascolo, C., and Sklar, M. (2017). Cultural investment and urban socio-economic development: a geosocial network approach. *Royal Society Open Science*, 4(9).

[192] Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313.