SAN DIEGO STATE UNIVERSITY AND

UNIVERSITY OF CALIFORNIA

Santa Barbara

Understanding Human Mobility and Urban Dynamics with Big Geospatial Data Analytics

A Dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Geography

by

Chanwoo Jin

Committee in charge:

Professor Atsushi Nara, Chair

Professor Ming-Hsiang Tsou

Professor Alan T. Murray

Professor Richard L. Church

December 2022

The dissertation of Chanwoo Jin is approved.

_____

Ming-Hsiang Tsou

_____

Alan T. Murray

_____

Richard L. Church

_____

Atsushi Nara, Committee Chair

December 2022

ACKNOWLEDGEMENTS

I am sincerely thankful to the members of my committee, Dr. Atsushi Nara, Dr. Ming-Hsiang Tsou, Dr. Alan T. Murray, and Dr. Richard L. Church Atsushi, for their guidance, encouragement, and support during the research and preparation of this dissertation. I especially express my sincere gratitude to Dr. Nara, my adviser, for his unwavering support and direction. He continuously provided encouragement and was always willing and enthusiastic to assist in any way he could throughout my Ph.D. program. It would not have been possible to complete my Ph.D. study without his thoughtfulness and patience in guiding my research.

I would also like to thank Dr. Tsou for providing a lot of opportunities to collaborate with brilliant scholars through the HDMA Center. I also really thank Dr. Murray for providing a lot of feedback, discussions, and support when I was at UCSB. It was my great pleasure to discuss my research with him every week and publish a peer-reviewed journal paper. I am also thankful to Dr. Church for his valuable feedback and comments on my research. I believe that all the knowledge and attitudes toward research I learned from my committee will be a great asset for my future career.

Finally, and importantly, I would be remiss to overlook the support of my wife, Sujin Sung, and my son, Logan Doyoon Jin. This dissertation would not have been possible without their warm love, continued patience and endless support. I am grateful to my parents, parents-in-law, and many friends for always standing by me, encouraging me, and believing in me.

This dissertation is supported in part by the National Science Foundation under Grant No. 1634641, IMEE, "Integrated Stage-Based Evacuation with Social Perception Analysis and Dynamic Population Estimation", Grant No. 1837577, CS4All, "Encoding Geography: Building Capacity for Inclusive Geo-Computational Thinking with Geospatial Technologies", and Grant No. 2031407, CS4All, "Collaborative Research: Encoding Geography-Scaling up an RPP to achieve inclusive geocomputational education". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

VITA OF CHANWOO JIN

December 2022

EDUCATION

Ph. D., Geography **2022**
University of California, Santa Barbara / San Diego State University
Dissertation: Understanding Human Mobility and Urban Dynamics with Big Geospatial Data Analytics

M.A., Geography **2016**
Seoul National University, Seoul, Korea
Thesis: Analysis of the Spatio-temporal Interaction Patterns of the Housing Sales and Rent Price Using SpVAR-Lasso

B.A., Geography **2014**
Seoul National University, Seoul, Korea

PROFESIONAL EMPLOYMENT

Fall 2022 – present	**Assistant Professor**
Northwest Missouri State University	Geography / Geographic Information Sciences
Fall 2017 – Spring 2022	Teaching Assistant
San Diego State University	Geography

PUBLICATIONS

Peer-reviewed Journals

Jin, C., Park, S., Ha, H. J., Lee, J., Kim, J., Hutchenreuther, J., and Nara, A., (In Review). Predicting Households' Residential Mobility Trajectories: A Geographically Localized Interpretable Model-agnostic Explanation, *International Journal of Geographical Information Science.*

Fernandez, G., Maione, C., Yang, H., Zaballa, K., Bonnici, N., Carter, J., Spitzberg, B., **Jin, C.**, and Tsou M. H. (2022). Social network analysis of Covid-19 sentiments: 10 metropolitan cities in Italy, *International Journal of Environmental Research and Public Health, 19(13)*. 7720. DOI: 10.3390/ijerph19137720

Jin, C., and Murray, A. T. (2021). Exploring Public Open Data: Spatiotemporal Dynamics of Restaurant Entrepreneurships in Seoul, Korea, *International Journal of Geospatial and Environmental Research. 8*(3). 5. https://dc.uwm.edu/ijger/vol8/iss3/5

Fernandez, G., Maione, C., Zaballa, K., Bonnici, N., Spitzberg, B. H., Carter, J., Yang, H., McKew, J., Bonora, F., Shraddha S. Ghodke, S. S., **Jin, C.**, Ocampo, R., Kepner, W., and Tsou, M. H., (2021) The Geography of Covid-19 Spread in Italy Using Social Media and Geospatial Data Analytics, *The International Journal of Intelligence, Security, and Public Affairs*, DOI: 10.1080/23800992.2021.1994813

Jin, C., Nara, A., Yang, J. A., and Tsou, M. H. (2020). Similarity Measurement on Human Mobility Data with Spatially Weighted Structural Similarity Index (SpSSIM). *Transactions in GIS*, 24(1), 104-122.

Jin, C., and Lee, G. (2020). Exploring spatiotemporal dynamics in a housing market using the spatial vector autoregressive lasso. *Transactions in GIS*, 24(1), 27-43.

Lee, G., **Jin, C.**, Kim, J., and Kim, W. (2016). A Study on the Characteristics of the Spatial Distribution of Sex Crimes. *Journal of the Korean Geographical Society*, 51(6), 853-871.

Jin, C., and Lee, G. (2016). Analysis of the Spatio-temporal Interaction Patterns of the Housing Sales and Rent Price Using SpVAR. *Journal of Real Estate Analysis*, 2(2), 23-42.

Choi, S., Lee, S., **Jin, C.**, and Lee, C. (2016). Exploring Availability of Real-estate Market Consumer Sentiment Index to Estimate Real-estate Overheating Regions. *The Korea Spatial Planning Review*, 88, 25-41.

Jin, C., and Lee, G. (2015). Optimal Location Modeling for Happy Houses. *Journal of the Korean Urban Geographical Society*, 18(2), 81-95.

Jin, C., and Lee, G. (2014). Spatial Hedonic Modeling Using Geographically Weighted LASSO Model. *Journal of the Korean Geographical Society*, 49(6), 917-934.

Book Chapters

Nara, A., Machiani. S. G., Luo, N., Ahmadi, A., Robinett, K., Tominaga, K., Park, J., **Jin, C.**, Yang, X., and Tsou M-H. (2021). Learning Dependence Relationships of Evacuation Decision Making Factors from Tweets. In Nara, A., and Tsou, M. H. (Eds), *Empowering Human Dynamics Research with Social Media and Geospatial Data Analytics* (pp. 113-138). Cham, Switzerland: Springer International Publishing.

Fernandez, G., Maione, C., Zaballa K., Bonnici, N., Spitzberg, B. H., Carter, J., Yang, H., Jack McKew, J., Bonora, F., Ghodke, S. S., **Jin, C.**, Rachelle De Ocampo, R. D., Kepner, W., Tsou M. H. (2021). Sentiment analysis of social media response and spatial distribution patterns on the COVID-19 outbreak: The case study of Italy. In *Empowering Human Dynamics Research with Social Media and Geospatial Data Analytics* (pp. 167-184). Cham, Switzerland: Springer International Publishing.

Conference Paper

Embury, J., Nara, A., & **Jin, C.** (2022). Spatially weighted structural similarity index: a multiscale comparison tool for diverse sources of mobility data. *In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility* (Nov. 1). Association for Computing Machinery. DOI:3557921.3565542

Jin, C., and Thompson, C. A. (2018). Identifying geographic disparities in breast cancer mortality in California. In *APHA's 2018 Annual Meeting & Expo (Nov. 10-Nov. 14).* American Public Health Association.

AWARDS

2022 Graduate Student Travel Award, *San Diego State University.*

2019 3rd Place, Seoul Research Competition, *The Seoul Institute*.

2019 Cotton Bridge Award for GIS Emphasizing Techniques, *San Diego State University*.

2019 Finalist, Student Paper Competition, Geographic Information Science and Systems Specialty Group, American Association of Geographers.

2018 Graduate Student Travel Award, *San Diego State University*.

2017 Presidential Graduate Research Fellowship, *San Diego State University*.

2016 1st Place, Research Paper Competition, *Korea Appraisal Board*.

2015 2nd Place, Research Paper Competition, Korea Research Institute for Human Settlements.

ABSTRACT

Understanding Human Mobility and Urban Dynamics with Big Geospatial Data Analytics

by

Chanwoo Jin

Human mobility and urban dynamics are the keys to understanding diversity and complexity in cities. With advancement of technologies, a significant amount of geospatial data is generated, shared, and analyzed. Big geospatial data analytics highlights the importance of geography in data-driven knowledge discovery. The goal of this dissertation is to progress the fundamental understanding of human mobility and urban dynamics by developing novel methodological frameworks and models that utilize micro-scale spatiotemporal big data and encourage knowledge creation. To achieve this goal, this dissertation includes three studies that focus on developing new methods to understand human mobility, urban dynamics, and their interactions. In the first study (Chapter 2), I propose a novel index measuring similarities between human mobility patterns from different data sources such as social media and traditional survey. The second study (Chapter 3) shifts focus to urban dynamics. Applying a series of spatiotemporal exploratory studies, an efficient method for examining spatiotemporal patterns at a micro-scale in restaurants is proposed. The third study (Chapter 4) investigates the relationships between human mobility and urban dynamics with a novel explainable deep learning approach

enhancing predictivity and interpretability of neural network models within geographic context. Throughout this dissertation, a comprehensive framework for understanding complexity of human mobility and urban dynamics is suggested through incorporating detailed spatial data into big data analytic models from geographic perspectives. This will provide individuals and governments with fundamental knowledge for better decision-making associated with economic growth and development.

TABLE OF CONTENTS

LIST OF FIGURES

1. Introduction

1.1. Motivation

Understanding human mobility and urban dynamics is fundamental to enhancing decision-making associated with socioeconomic development in urban systems. Human mobility, including physical travel of people and goods and virtual transactions of information, plays a significant role in (re)distributing uneven resources in cities and (re)organizing urban structures. Modern cities, homes of more than half of the world's population, are not simple spaces where various activities occur, but complex places where human activities continuously interact with urban structures. Impacts of human activities on urban environments are non-linearly intertwined with various unobserved and unknown factors. Urban dynamics, as the study and understanding of forces and their effects on changes in urban structures such as land uses, socioeconomic functions of locations, and relationships between areas, have been highlighted to understand the complexity of cities (Forrester, 1969; Batty et al., 1999). They suggested bottom-up approaches that emphasize impacts of heterogeneous and autonomous individuals on spatiotemporal changes in urban structures; however, insufficient micro-scale data on human mobility and urban dynamics have resulted in major challenges.

The recent advancement of technologies, along with a new culture of data creation and sharing is enabling a generation of large and diverse data in real-time, making them accessible for unveiling various patterns of human mobility and urban dynamics (Shaw et al., 2016). For example, location-based social media and consumer review services provide information on individuals, revealing their current locations and impressions on places

through check-in, messages, and ratings. In addition, open data initiatives encourage governments to share individual-level data routinely collected for public purposes such as welfare, taxation, and business licensing (Arribas-Bel, 2014; Lansley et al., 2018). Although the availability of these new spatially and temporally fine-granular data can help fill research gaps in understanding of micro-scale human mobility and urban dynamics, it remains a challenge to utilize these new types of data in urban studies due to population biases, privacy concerns, uncertainties, and concurrence of scales from multiple data sources (Longley et al., 2015).

Various methodologies, from traditional statistical models to cutting-edge big data analytics, have been applied to comprehend the complex relationships between human mobility and urban dynamics. The increasing amount of georeferenced data enhances the importance and the necessity of geospatial exploratory data analysis, which aims to identify underlying spatiotemporal patterns and trends hidden in datasets beyond priori knowledge (Miller & Han 2009; Miller & Goodchild, 2015). Furthermore, using geospatial big data has also improved performance of Artificial Intelligence (AI) techniques, especially deep neural networks, in geographic studies as GeoAI (Janowicz et al., 2020). Despite the substantial progress in applying deep neural networks to geographic studies, for example, to classify spatial objects from remote sensing images and textual data, GeoAI remains in an early stage with many technical and theoretical challenges such as a lack of explainability of deep learning models (Li, 2020). Explainable AI is an ongoing research effort to increase the transparency and the interpretability of models to verify its modeling process and outcome and to provide human-understandable justifications for supporting decision making in practical applications. However, it is still challenging to implement explainable neural

network models for geographic studies because of the absence of a methodological framework and case studies.

1.2. Big Geospatial Data Analytics: GeoAI

The term Artificial Intelligence (AI) is not a completely new term, but one that has been re-highlighted and is now ubiquitous, not only in academia but also in our daily lives. Automatic subtitling and recommendation systems in YouTube, autonomous driving cars produced by Tesla, and virtual assistant systems in smartphones are prominent examples of the successful utilization of AI techniques. AI is simply defined as an attempt to enable a computer to have some of the same intellectual capabilities and thought processes as human beings (Openshaw & Openshaw, 1997). It is an extremely broad concept encompassing many other technical terms such as machine learning (ML), deep learning (DL), artificial neural network (ANN) and deep neural network (DNN). For instance, ML is a sub-field of AI, referring to the study of improving computer algorithms through iterating experiments or processes, which are, in turn, implemented by the use of techniques such as neural networks.

AI development has been ongoing since the 1930s, but there have been several fluctuations in its popularity due to computing, algorithm, and data issues. For example, in the 1950s, neural networks were conceptually developed, but researchers at the time lacked sufficient tools to support the idea. It was not until the 1980s that advancements in computer hardware and software refueled the study of the field. However, AI faced another "winter" in the early 1990s because there were insufficient data to solve difficult problems to which complex AI techniques could be applied. DNNs that stacked multiple neural networks as layers, were developed in the 1990s, but they did not flourish until the 2010s. In addition to technical advances, significant changes in culture have played a key role in advancing AI,

particularly in the case of deep learning techniques (Janowicz et al., 2020). Opening data to the public has successfully attracted enormous attention from diverse fields. For example, public competitions, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) for classifying images (ImageNet, 2017) and the Netflix Prize for the best prediction of movie ratings in 2009, encouraged people to develop better models with advancements in prediction.

In contrast to the diverse applications of AI in computer science and other fields, AI has attracted less attention from geographers, even geographic information (GI) scientists utilizing quantitative approaches. However, there have been some attempts to adapt AI to geography since the 1990s (Openshaw & Openshaw, 1997). The concept of AI has included diverse notions such as machine learning, methodologies developed by GI scientists can also be regarded as AI techniques. Geographically weighted regressions (GWR) (Brunsdon et al., 1998) are another good example of integrating geospatial models and machine learning techniques. As a kind of local regression, it is an efficient method to explore spatial heterogeneity such as varied relationships between independent variables and a dependent variable in a space. One challenge in this process is defining the range of neighborhood due to the bias–variance trade-off problem. When a small neighborhood range is defined, the model will be accurate but with little flexibility (overfitting). On the other hand, with a large neighborhood, it will be less accurate but more flexible (underfitting). As an example, GWR models highlight the importance of definitions of locality to control model fitting problems and provide interpretable explanations in terms of geographic context.

Although there have been several successful integrations of GIS and machine learning, these seem insufficient to reflect recent advancements of AI, a particularly deep learning

process or deep neural networks. There are many obstacles hindering development of AI in GIS. Tsou (2017) points out that the lack of programming skills and suitable high-performance computing (HPC) frameworks is a significant challenge for integrating AI and GIS. Moreover, uncertainty and complexity of spatial data have also impeded opening spatial data to the public (Kedron et al., 2021). For example, the practice of standardizing datasets, like the enormous number of labeled images that have been obtained by ImageNet has the power to attract researchers and ideas from diverse fields and provide ground truth to compare and evaluate new models and techniques. GI scientists have been less devoted to these efforts. However, since 2017, after the first international workshop on the theme of "GeoAI", academic interest in this area has spiked and the amount of data from diverse sources such as social media, crowdsourcing maps and street images continues to grow (Janowicz et al., 2020).

GeoAI, or geospatial artificial intelligence, is an attempt to combine AI, geospatial big data, and high-performance computing to provide a better understanding of complex geospatial processes (Li, 2020). GeoAI does not refer to a single analytic tool or technique, but a new research agenda encompassing a variety of research topics from data acquisition and storage to analysis and visualization. Although there are, as of yet, few studies explicitly addressing geospatial context in AI techniques, it is applied to a wide range of fields including health (VoPham et al., 2018; Boulos et al., 2019), mobility (Yin et al., 2019; Xing et al., 2019; Yin et al., 2019), neighborhood conditions (Yen et al. 2018), disaster management (Tien et al., 2018; Peng et al., 2019), urban dynamics (Dorji et al., 2019; Snyder et al., 2019) and land use changes (Gebru et al., 2017; Shi 2019; Law & Leira, 2019).

GeoAI remains in an early stage with many technical and theoretical challenges including an insufficient explainability of deep learning models (Li, 2020). Although many studies have successfully applied popular deep learning techniques such as convolutional neural networks (CNNs) to classify geospatial objects and to detect changes in spatial structures from remote sensing images (Dorji et al., 2019; Snyder et al., 2019; Li & Hsu, 2020), it is still challenging to explain the relationships between inputs and outputs and the reason why the models provide better prediction. Improved predictive accuracy has been achieved by GeoAI models, but explanation remains as the next question of GeoAI (Papadakis et al., 2022).

As deep learning techniques employ a large number of hidden layers to improve their performance, the architecture of a network becomes complex, with a number of layers, thus hindering humans from understanding these processes. For example, one of the earliest models of convolutional neural networks for image detection, named *LeNet-5*, used six hidden layers and about 60 thousand parameters to identify a handwritten letter with a 32 by 32 image (LeCun et al., 1998). Within the architecture, even a single feature generates a considerable number of random values, which improves model performance such as prediction, but they are hardly observed because they have too many parameters to be interpreted one by one and do not have any contextual meanings. Because of the interpretability issues with complex structures of hidden layers and neurons, deep neural networks are frequently referred to as "black-box" models. (Gilpin et al., 2018).

The explainability of a model is essential for validating the modeling process and results as well as for providing human-understandable knowledge to support decision-making in critical applications, including medicine, urban planning, and disaster management.

Explainable AI has been widely explored as a new research priority in several disciplines, with the goal of improving the transparency and interpretability of deep learning models (Samek et al., 2018). Generally, two approaches are recommended: (1) visualizing hidden layers and their relationships and (2) contextualizing the network's component parts. Researchers can identify more crucial values than others by visualizing the process of how input values change but attempting to understand a model in the context of geography is less focused. To summarize, GeoAI studies employing deep neural networks are required to consider the explainability of their models in order to enhance our understanding of complex relationships between human mobility and urban dynamics.

1.3. Research Objectives

In this dissertation, I introduce novel methodological frameworks and models that utilize micro-scale geospatial big data to advance knowledge in human mobility, urban dynamics, and their relationships through big data analytic approaches. Three key research challenges are: (1) evaluating human mobility patterns imprinted in various data sources including social media data and public data; (2) understanding micro-scale urban dynamics through publicly available individual-scale data; and (3) explaining urban dynamics with human mobility through highly accurate and interpretable neural network models.

The first study in this dissertation addresses a challenge of understanding diverse characteristics of human mobility patterns across geographic scales by utilizing fine-scale data extracted from diverse sources. Due to discordance of spatiotemporal resolution between data sources, aggregation at a certain level is required. An origin–destination (OD) matrix provides mobility patterns among spatial units within a given temporal scale. This study proposes a novel method to measure similarity of origin–destination (OD) matrices by

considering spatial contexts that usually determine mobility patterns. It will provide an underlying knowledge of human mobility extracted from social media data, which can ultimately facilitate the understanding of the complexities of human mobility (Jin et al., 2020).

The second study in this dissertation investigates the value of public open data in urban studies. In this study, individual-level data on restaurant business licenses is explored to identify underlying spatiotemporal trends in micro-scale urban dynamics. Among many businesses in the tertiary industry, restaurants are one of the most common services and they open and close more frequently than other types of service-based firms. This study explores spatiotemporal changes in restaurant locations with hot spot analyses, trends analysis on spatial clusters, and space-time scan statistics. This study presents a new analytical framework to discover spatial hot spots of commercial activities and their temporal fluctuations. Identifying not only spatial patterns but also temporal fluctuations is key to deepening our understanding of urban dynamics, providing insights on drivers of economic growth and development (Jin & Murray, 2021).

The third study in this dissertation identifies the relationships between urban dynamics and human mobility with recurrent neural networks based on the survival analysis framework. To enhance explainability of the neural network model, this study proposes Geographically Localized Interpretable Model-agnostic Explanation (GLIME) by extending Local Interpretable Model-agnostic Explanation (LIME) within geographic context. The geographically explainable deep neural networks provide significant improvements in the predictability of nonlinear relationships between human mobility and urban dynamics, as well as shed light on complex mechanisms underlying human mobility and urban dynamics.

Most importantly, the findings emphasize the importance of spatial heterogeneities in impacts of human mobility on urban dynamics.

1.4. Significance

This dissertation provides a comprehensive framework for studying human mobility and urban dynamics, particularly from a spatiotemporal perspective through big geospatial data analytics. Moreover, this study contributes to the literature in the fields of urban geography, economic geography, geospatial data science, and complex system science by addressing challenges around the use of micro-scale big data, integrating heterogeneous multi-scale data, modeling spatial and spatiotemporal interactions and processes, and applications of AI to geographical studies. By emphasizing the spatial perspective, in particular, this study contributes to geographic information sciences by extending aspatial analytic tools geographically with consideration of spatial context (Chapter 2 and 4). Also, it advances the study of explainable AI and GeoAI for social sciences, aiming for a comprehensive understanding of the processes of social phenomena (Chapter 4). It can be used to answer a number of research questions such as identifying spatial and spatiotemporal disparities in socioeconomic developments.

The knowledge gained through the framework provides critical insights not only to researchers in academia but also governments, industries, and citizens, thereby enhancing their geospatial decision-making processes associated with socioeconomic development. By utilizing publicly accessible data and analytic tools, the study lowers barriers to data analyses, which are essential for better decision-making (Chapter 3 and 4). This study provides critical guidance to individuals and governments, explaining complexity of human

mobility and urban dynamics with better understanding of complexities in human and urban dynamics associated with economic growth and development.

2. Evaluating Human Mobility Patterns in Big Data[1]

2.1. Introduction

Urban dynamics are built up on diverse types of human movements including migration, commuting, and leisure activities and hence the study of human mobility is essential for understanding complex urban systems (Dodge et al., 2016; Miller & Shaw, 2015). To examine human mobility and urban dynamics, various forms of data sources exist; for instance, census surveys, activity diaries, Call Detailed Records (CDRs) from cellphones, Global Positioning System (GPS) points, and new media data. Among these, researchers have recently adopted social media data, which has produced a massive amount of publicly available individual-scale timestamped geo-referenced data, for discovering unrevealed travel patterns and activities. For example, finding major human flows (Gao et al., 2018; Poon & Pandit, 1996; Xu et al., 2015) can support decisions across multiple fields including public health, epidemiology, and disaster response (Martín et al., 2017; Nara et al., 2017; Panigutti et al., 2017; Wesolowski et al., 2015). While previous studies have often examined and analyzed spatial distributions of movements, less research has focused on measuring similarities and differences between multiple data sources (Gao et al. 2018). Each data source has its unique characteristics to describe human mobility due to diversities in survey participants or platform users who provide their mobility data, and methodologies to collect, manage, and publish geo-referenced data. For instance, in the case of social media data

[1] This chapter represents a revised version of a paper published in *Transactions in GIS*.
Jin, C., Nara, A., Yang, J. A., and Tsou, M. H. (2020). Similarity Measurement on Human Mobility Data with Spatially Weighted Structural Similarity Index (SpSSIM). *Transactions in GIS*, 24(1), 104-122.

demographic characteristics, Instagram users are more popular among female and younger populations than Twitter (Table 2.1), the variations of which may produce significant differences in mobility patterns and play a significant role in shaping the complexity of human mobility.

Table 2.1 Demographic characteristics of Instagram and Twitter in the U.S. (Greenwood et al., 2016).

		Instagram	Twitter
All online adults		32	24
Gender	**Men**	26	24
	Women	38	25
Age	**18-29**	59	36
	30-49	33	23
	50-64	18	21
	65+	8	10

% of online adults who use social media

Therefore, it is crucial to understand the capabilities and limitations of each data source for describing human mobility. This can further help in grasping comprehensive human mobility patterns, where multiple mobility data sources complementarily explain different types of human movements. There have been attempts to develop new methodologies to measure similarities between mobility patterns from diverse sources (Xia et al., 2011; Yuan & Raubal, 2014) and to address those demographic biases on multiple social media data (Crooks et al., 2015; Gao et al., 2017). However, effectively and quantitatively measuring mobility similarities from multiple data sources continues to be a research challenge.

To address this research gap, this chapter proposes a new method, Spatially weighted Structural SIMilarity index (SpSSIM), to measure the similarity of Origin-Destination (OD) flow matrices to compare mobility patterns from multiple data sources. SpSSIM adopted Structural SIMilarity index (SSIM), an image quality assessment technique to measure the similarity between two images (Wang et al., 2004). SSIM has been applied to measure the

similarity of human mobility by mapping the OD matrices into arrays of image values; however, previous works do not consider the spatial configuration of flows on the OD matrices. We extended SSIM by incorporating spatial adjacency by utilizing a series of spatial weight matrices. A key advantage of SpSSIM is that local similarities can be properly measured and investigated.

While SSIM utilizes a square moving window to calculate the similarity of images or matrices, SpSSIM employs a geographically defined range with spatial weights. This enables SpSSIM to calculate similarities of flows in a certain geographic boundary. As our case study, we compared each pair of OD matrices of human daily mobility extracted from three mobility data sources; U.S. Census-based Longitudinal Employer-Household Dynamics Origin-Destination employment statistics (LODES), Twitter, and Instagram, and aggregated at the sub-regional areas (SRAs) scale in San Diego County, CA. Two geo-referenced social media data, Twitter and Instagram, were collected via publicly available Application Programming Interfaces (APIs) in 2015. The case study demonstrated the capability of SpSSIM to measure the mobility similarities between each data source and to provide an underlying knowledge of human mobility extracted from social media data, which can ultimately facilitate the understanding of the complexities of human mobility. The remainder of this article is organized as follows. Section 2 describes related works on measuring the similarity of human mobility and studying human mobility with social media data. In Section 3 introduces the SpSSIM methodology and Section 4 describes data used in this study, respectively. Section 5 details the results of comparative experiments between SSIM and SpSSIM and interpretation of SpSSIM values with a case study of San Diego

County, CA. The final section discusses implications of the results and conclusions with future works.

2.2. Related Work

2.2.1. Methodological Approaches for Quantifying Similarity of Mobility

Methodological approaches to characterize and compare mobility patterns have been, in essence, quantifying similarity in mobility data, which can identify major trends in movements and allow comparing the trends from diverse data sources. As a traditional approach, dominant flow analysis (Nystuen & Dacey, 1961) counted the amount of flow and detected major trends of human mobility such as trading (Smith, 1970; Xu et al., 2015) and tourist traveling (Pearce, 1996). Another approach was employing principal component analysis to identify centers of mobility network (Garrison & Marble, 1964). Components derived from PCA represent the similarity of regions regarding the amount of flow (Poon & Pandit, 1996). For example, Clayton (1977) categorized US states in terms of the similarity of origins and the numbers of inter-state immigrants. More recently, Gao et al. (2018) employed spatial scan statistics (Kulldorff, 1997) to compare major patterns of taxi trips in the morning and afternoon in New York City and county-to-county migration flows between age-groups in the United State by clustering origins and destinations. While these methodological approaches have been effective to summarize mobility patterns with a few major trends, the number of clusters was potentially arbitrary, and the comparison was limited to detected clusters (Salvador & Chan, 2004). In other words, the similarities measured by clusters can be sensitive to the number of clusters and clustering methods.

Various methodologies have been explored to calculate the similarity between mobility data, which are often treated as trajectories, i.e., two sets of temporally sequenced location points. Common trajectory similarity measurements calculate distances between points on each trajectory. For example, Euclidian distance has been widely used to measure geographic gaps between two points from each trajectory (Zheng & Zhou, 2011). Meanwhile, Levenshtein distance, or edit distance, developed from informatics to measure distance between two strings has also been applied to geographic trajectories by matching each of their intermediate points and calculating the differences (Yuan & Raubal, 2014; Yuan & Nara, 2015).

These approaches provide similarity measurements of individual sequential movements. More recently, Behara et al. (2018) proposed Mean Normalized Levenshtein distance for OD matrices (MNLdOD) by applying Levenshtein distance to measure similarity of two OD matrices. This measurement compares the descending order of destinations by the normalized flow volume from each origin (i.e., a row of OD matrices) as strings and calculates the similarity row by row to capture differences in the order of destination choices from the same origin. Since MNLdOD employs the orders by flow volumes rather than the actual number of flows, it is limited in fully incorporating flow volume variations in its similarity index.

2.2.2. Human Mobility and Social Media

Human mobility has been a long-discussed issue in social and geographic sciences and complex mobility dynamics and behaviors have been studied in a variety of applications including migration, traveling, and evacuation (Cresswell, 2012; Noulas et al., 2012). Focusing on the daily human mobility, fundamental activities of human living such as

commuting, shopping, and leisure trips often accompany movements, which further contribute to form complex urban dynamics through human-human and human-environment interactions in space and time (Huang & Wong, 2016; Sun et al., 2015; Wu et al., 2014). Thus, investigating human mobility is a key to comprehend complex urban systems, yet it has been challenging (Larsen et al., 2006; Yuan & Raubal, 2014) especially since traditional data collection methods (e.g., census survey and travel diary) were limited to observe and recode human mobility at the full-scale due to their high cost (Miller & Shaw, 2015). Nowadays, recent advancement and adaptation of mobile Information and Communication Technologies (mICTs) and location-aware technologies (LATs) have enabled the recording of human mobility via mobile devices. This larger and finer scale spatiotemporal data can fulfill the investigation of daily human mobility and reveal un-discovered patterns not captured in the data collected through traditional methods (Hawelka et al., 2014; Liu et al., 2012; Wu et al., 2014).

Social media can be one of data sources capturing human mobility by taking advantages of mICTs and LATs. Recent studies have explored the potential of social media data to reconstruct individual trajectories and describe dynamic human movement behaviors in detail (Nara et al., 2017). For example, traveling patterns and behaviors have been studied using check-in data to detect hotspots and unusual visiting places (Sun et al., 2015), and geo-tagged Twitter posts to estimate the volume of country-to-country flows (Hawelka et al., 2014). Geotagged social media posts have been applied to investigate human mobility and evacuation behavior during disastrous events (Li et al., 2018; Martín et al., 2017; Wang & Taylor, 2014). Despite the usefulness of social media data to investigate human and urban dynamics, they are known to be biased by the unevenness of demographic, geographic, and

temporal distributions. Furthermore, since each social media platform has its own unique usages and demographics, human mobility patterns extracted from social media potentially differ by platforms; however, few studies have investigated the similarity and difference in human mobility by social media platforms.

To solve the bias of a single social media data source, some studies have integrated multiple social media data sources with other spatial and aspatial data to gain profound knowledge in human activities and urban contexts. For instance, Gao et al. (2017) synthesized multiple data sources from Flickr, Instagram, Twitter, Travel Blogs, and Wikipedia to extract semantics and identify cognitive regions based on a belief that each source represented different user groups. They assumed that Flicker was more tourism-oriented than other social media such as Twitter and Instagram, which showed daily activities, news and visited places. Crooks et al. (2015) utilized open-source crowdsourcing datasets ranging from Global Positioning System (GPS) tracks and Foursquare to Twitter, Flickr, and weblogs to demonstrate how social media, trajectory, and traffic data could be analyzed to capture the evolving nature of a city's form and function. They argued that each data source represented a part of dynamic and complex urban functions. These approaches highlighted the importance of integrating multiple data sources to gain deeper insights into urban dynamics. Despite the importance and necessity of data integration in mobility studies, it has been less utilized due to limited understanding of the capabilities and limitations of each data source and their similarity and difference.

2.3. Methodology

This research proposes a novel index, spatially weighted SSIM (SpSSIM), to measure the similarity of two OD flow matrices to compare mobility patterns. Our method adopted

the concept of structural similarity index (SSIM), which originally assess image quality by comparing local patterns of image structure. We extended SSIM spatially by utilizing a series of spatial weight matrices that define the range of neighborhood geographically. The following sub-sections demonstrate the algorithm of SSIM, the process of spatial extension, and verification of SpSSIM.

2.3.1. Spatially Weighted Structural Similarity Index (SpSSIM)

Wang et al. (2004) developed SSIM to measure the similarity between two images for assessing the quality of copied or generated images from an original image. As the human visual system is familiar with the overall arrangement of images rather than single values of cells to compare images, SSIM considered the arrangement of image values by quantifying the local patterns of pixel intensities consisting of three components - luminance, contrast, and structure – with a moving window. SSIM calculates image similarity between two images X and Y is expressed as Equation 2.1.

$$SSIM(x,\ y) = f(l(x,\ y)^{\alpha} \cdot c(x,\ y)^{\beta} \cdot s(x,\ y)^{\gamma})$$
$$l(x,\ y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2+\mu_y^2+C_1},\ c(x,\ y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2},\ s(x,\ y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y+C_3} \quad (2.1)$$

where $l(x,\ y)^{\alpha}$, $c(x,\ y)^{\beta}$, and $s(x,\ y)^{\gamma}$ represent the three components luminance, contrast, and structure, respectively. *x* and *y* denote black-white color values of pixels (0 to 255) in images X and Y. μ, σ^2, and σ_{xy} refer mean, variance, and covariance, respectively. C_1, C_2, and C_3 are constants to enforce SSIM to be less than 1. Therefore, the value of SSIM equals 1 when two images are exactly same, and it gets close to 0 when they are less similar. When we regard the importance of each component is identical ($\alpha = \beta = \gamma = 1$), and C_3 is equal to $C_2/2$, Equation (1) can be simplified as the Equation 2.2.

$$SSIM(x,\ y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad 2.2$$

Once SSIM is calculated in a window, it moves to the next cells and computes SSIM by the last cell of images. Then, the overall similarity of images X, Y is represented by the mean of local SSIM values when M denotes the total number of local windows (Equation 2.3).

$$MSSIM(X,\ Y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j, y_j) \quad 2.3$$

SSIM has been widely utilized to assess the quality of images and to evaluate the performance of image processing due to its simplicity and accuracy (Brunet et al., 2012). The index has been recently employed to compare movements because the amount of flow in OD matrices can be considered as values of images (Djukic, 2014; Pollard et al., 2013). For example, Djukic (2014) used SSIM with a square window to evaluate the estimation of OD demands in traffic flows (Figure 2.1) since traditional statistics such as MSE are limited to consider the spatial correlation between OD pairs. Yet, SSIM is still problematic in terms of the sensitivity of OD pairs ordering because the order in a matrix is not always arranged by spatial contiguity or distances. For example, the contiguous cells in an OD matrix can be far from each other in real space when the order is based on the size of population or is randomly distributed. In this case, a square window is limited to filter out the spatial correlation between OD pairs because contiguous cells are not spatially adjacent. Moreover, the same values of OD matrices with different orders result in different SSIM values. To solve the problem, Djukic (2014) suggested to find the best way to represent spatial dependency of OD pairs by testing various window sizes and Behara et al. (2017) re-ordered

OD pairs to group to the upper level and calculated SSIM within the new level; however, selecting the optimal window size and the optimal order of OD pairs remains challenging.

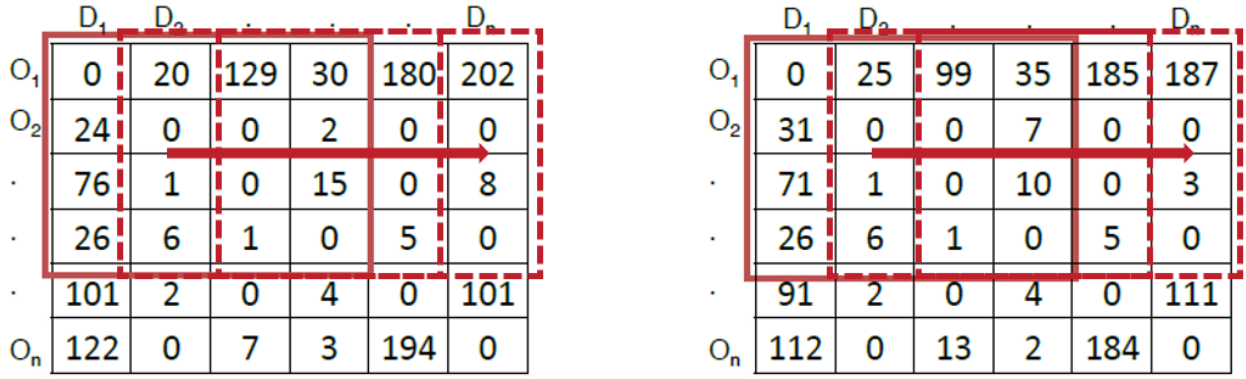


Figure 2.1 Comparison of two OD matrices using SSIM (Pollard et al. 2013).

To overcome the SSIM ordering issue to compare OD matrices, SpSSIM (Equation 2.4) utilizes a series of spatial weight matrices instead of the moving window of SSIM. The range of locality is defined by multiplying a spatial weight matrix with OD matrix using Hadamard product (Equation 2.5). The spatial weight matrix consists of 0 and 1 where flows (f_{ij}) in the distance range (d_{ij}) has the weight value of 1 (Equation 2.6). In other words, SpSSIM computes a similarity between two OD matrices only in a specific geographic range by blocking values from outside of the range with multiplying 0 to the values (Figure 2.2). As a result, SpSSIM will have the value between 0 and 1 and be close to 1 when two matrices are similar.

$$SpSSIM(x, y, w) = \frac{(2\mu_{wx}\mu_{wy} + C_1)(2\sigma_{wx,wy} + C_2)}{(\mu_{wx}^2 + \mu_{wy}^2 + C_1)(\sigma_{wx}^2 + \sigma_{wy}^2 + C_2)} \quad 2.4$$

$$WF = W * F = \begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix} * \begin{bmatrix} f_{11} & \cdots & f_{1j} \\ \vdots & \ddots & \vdots \\ f_{i1} & \cdots & f_{ij} \end{bmatrix} = \begin{bmatrix} w_{11} * f_{11} & \cdots & w_{1j} * f_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} * f_{i1} & \cdots & w_{ij} * f_{ij} \end{bmatrix} \quad 2.5$$

$$w_{ij}^{D_{max}^{min}} = \begin{cases} 1, & D_{min} \le d_{ij} < D_{max} \\ 0, & Otherwise \end{cases} \quad 2.6$$

	D_1	D_2	.	.	.	D_n
O_1	0	20	129	30	180	202
O_2	24	0	0	2	0	0
.	76	1	0	15	0	8
.	26	6	1	0	5	0
.	101	2	0	4	0	101
O_n	122	0	7	3	194	0

	D_1	D_2	.	.	.	D_n
O_1	0	25	99	35	185	187
O_2	31	0	0	7	0	0
.	71	1	0	10	0	3
.	26	6	1	0	5	0
.	91	2	0	4	0	111
O_n	112	0	13	12	184	0

$w_{ij}^{D} = 0$

$w_{ij}^{D} = 1$

Figure 2.2 Comparison of two OD matrices using SpSSIM.

For the global similarity between two OD matrices, a series of spatial weight matrices is required to encompass the whole study area (Equation 2.7). For instance, if the first spatial weight matrix takes 1 when the distance between two regions is less than 10km (bin 1), the flows in the range of 0 to 10km are included for a local SpSSIM value. Then, the second matrix represents spatial relationships between two regions in the range of 10 to 20km to calculate another SpSSIM value in the next bin two. When the series of bin covers the whole region (bin *n*), the average of SpSSIM values in each bin is calculated to denote the overall similarity between two OD matrices. As SSIM, the SpSSIM value equals to 1 when two OD matrices have the exact same patterns.

$$Global\ SpSSIM(X,\ Y) = \frac{1}{n}\sum_{b=1}^{n} SpSSIM(X, Y, W^b) \quad 2.7$$

Moreover, SpSSIM can compare local inbound flows (in-flows) and outbound flows (out-flows) (Localized SpSSIM). The similarity of local directions of movements is measured by calculating SpSSIM with only columns or rows in a spatial weight matrix (Equation 2.8). When the *i*-th row of two matrices are compared in a geographic bin, the value of localized SpSSIM represents the similarity of out-flows starting from the *i*-th region. On the other hand, the similarity of flows moving into the *j*-th region can be calculated by comparing the *j*th columns of two matrices.

$$Localized\ SpSSIM(X,\ Y) = \begin{cases} SpSSIM(X_i, Y_i, W^b), & outflow \\ SpSSIM(X_j, Y_j, W^b), & inflow \end{cases} \quad 2.8$$

2.3.2. Bootstrap Verification

To verify the statistical significance of SpSSIM, we employed bootstrap to estimate the distribution. Bootstrap generates a random distribution of a parameter by iteratively resampling the observed data (Westfall & Young, 1989). It simulates samples with the same size of observation and allows replacement (Efron, 1979). For example, when a set of observation is $X = \{x_1, x_2, x_3, \dots, x_n\}$, a sample can be $\tilde{X}_1 = \{x_3, x_2, x_2, \dots, x_n\}$. Then, the process of sampling is iterated a large number of times and statistics of the created samples are computed. It is similar to Monte Carlo simulation regarding repeating a process, but Monte Carlo simulation generates random cases under null hypothesis rather than resamples from the existing dataset. In this study, we resampled the observed number of flows from each OD matrix due to difficulty to assume a null hypothesis for mobility patterns. We randomly resample 999 times to generate the probability distribution of SpSSIM and

estimate p-values of the index based on the mean and standard deviation of resampled values. The p-values verify whether the observed mobility patterns from two data sources are randomly different or not.

2.4. Data

2.4.1. Build OD Matrices from Social Media Data

As a case study to demonstrate SpSSIM to measure similarity in human mobility, we compared OD matrices generated from three data sources, Twitter, Instagram, and LODES. First of all, we collected Twitter and Instagram georeferenced posts using APIs from 12/07/2014 to 05/17/2015 (161 days) in San Diego County. To avoid duplicated posts possibly generated by cross-posting from Instagram to Twitter, we considered tweets only posted from mobile-based Twitter application sources (e.g., Twitter for Android, Twitter for iPhone, etc.). This removed tweets cross posted from other social media platforms including Instagram and Foursquare. The numbers of posts in the period were 1,916,580 for Twitter and 4,362,176 for Instagram, respectively. From these social media posts, we extracted individual daily trip segments by connecting temporally adjacent two georeferenced points within the same day. For example, if a person posted a message on a social media at a location A in the morning and another one at a location B at night, the segment from A to B is regarded a movement. If the person posts at three locations A, B, C sequentially in a day, the segment is regarded as two movements, A-B and B-C.

These extracted trip segments include irrelevant data such as no movements and movements with an unrealistically high velocity. To further clean up the irrelevant trip segments, we removed segments with zero distance where their origin and destination are at

a same location. We also removed segments with the average velocity greater than sixty-five miles per hour (mph), which is the state's general maximum speed limit in California (California Department of Transportation). Then we aggregated these trip segments to build OD matrices based on 41 San Diego sub-regional areas (SRAs) as a spatial unit. SRAs represent local communities/neighborhoods that is suitable to describe regional contexts of human mobility. From the social media data, we extracted 116,253 and 297,339 individual daily movements between SRAs from Twitter and Instagram, respectively. **Error! Reference source not found.** summarizes the number of data collected, processed, and generated for each social media platform.

Table 2.2 Data summary.

Source of social media	Number of posts originally collected	Number of posts after removing cross-platform data	Number of extracted daily movements
Twitter	2,202,719	1,916,580	116,253
Instagram	4,362,176	-	297,339

To compare human mobility extracted from social media data with non-social media-based mobility data, we used the LODES data. It represents commuting patterns, or home-to-work flows, based on employer reporting records at census block level that can cover more than 90 % of all employment categories except self-employment or military personnel (Horner & Schleith, 2012). We aggregated LODES flows into the SRA level to investigate the regional context of human mobility patterns. A flow between two SRAs was defined as f_{ij} indicating a person moved from the *i*-th SRA to the *j*-th SRA in a day. We define origin regions as rows and destination as columns in an OD matrix, where the sum of the *i*-th row denotes the total amount of out-flows from the *i*-th SRA and the sum of the *j*-th columns represents the total amount of in-flows to the *j*-th SRA. To understand flows between

neighborhood, we removed internal flows, which their origins and destinations are the same region (f_{ii}). Then, we normalized the flows between SRAs through probability of flows, which scale the value of flows from all datasets to be between zero and one.

2.4.2. Data Description

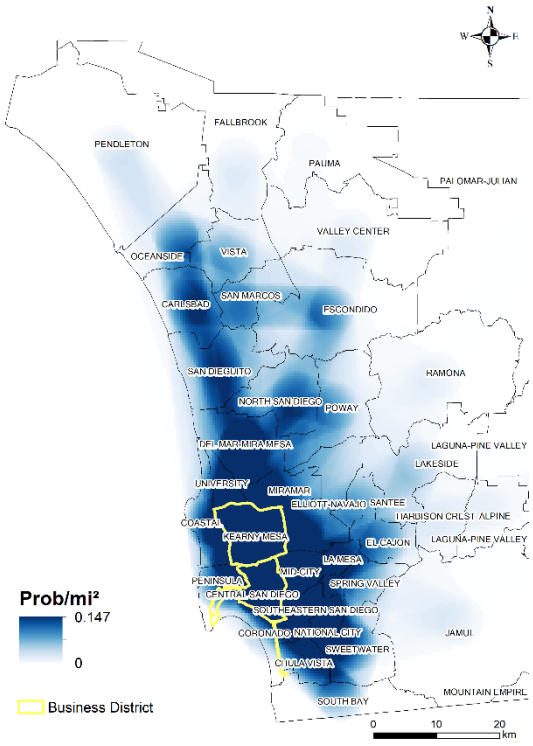
Table 2.3 describes descriptive statistics of flows in three data sources. Generally, all probability distributions of flows were positively skewed. Almost all movements from three sources were traveled within 50 km. Twitter and Instagram have more movements within 20 km (78.8%, 73.0% respectively) than LODES (57.5%). However, LODES has the largest total amount of flows and the lowest percentages of zero cells, which refer to no movements between two SRAs. In LODES, there are 23 (1.4%) OD pairs of SRAs with no flow out of 1,640 pairs (41x40) excluding internal flows. On the other hand, there are many OD pairs with no flow in Twitter and Instagram. This describes that the commuting-based mobility (LODES) was more ubiquitously distributed in San Diego County than social media-based mobility. This further indicates that social media were more frequently used within geographically confined regions rather than overall regions. Between Twitter and Instagram, Instagram-based flows (zero OD flows=27.5%) were geographically sparser than Twitter-based flows (zero OD flows=13.4%) even though the number of flows in Instagram is 2.5 times more than that in Twitter. One potential explanation for these patterns is that each social media has different usages and purposes. For example, Twitter users are more likely post messages at their ordinary locations such as home and work, while Instagram users are more willing to share their extraordinary experiences by posting pictures at places for social activities and entertainment. Section 5 provides further discussions on the difference in flows between Twitter and Instagram.

Table 2.3 Descriptive statistics of flows.

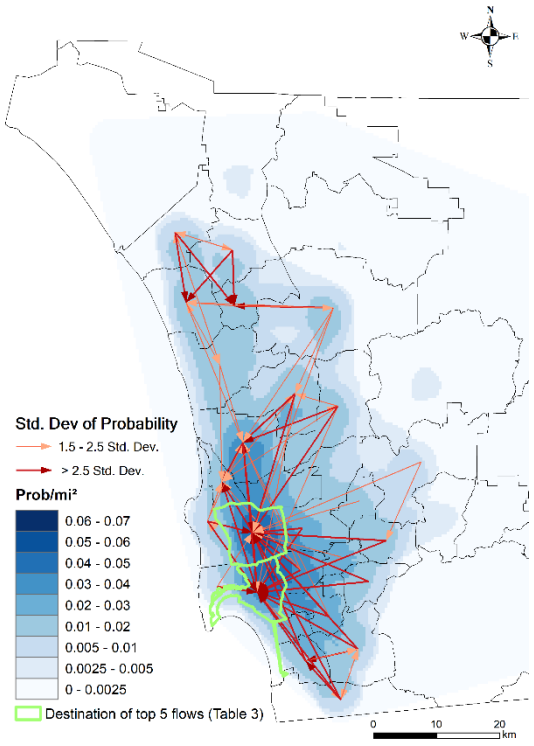
	LODES	Twitter	Instagram
Total amount	836,974	116,253	297,339
Mean	497.9	69.2	176.9
Median	91	8	14
Std. D	1,135.601	180.722	551.491
Max	10,438	1,539	7,163
1st quartile	19	2	0
3rd quartile	420	44	108
# of Zero (ratio)	23 (1.4%)	219 (13.4%)	451 (27.5%)
Skewness	4.444	4.605	7.418
By 20km	57.535%	78.849%	73.040%
By 50km	96.490%	98.042%	98.116%

Table 2.4 Top 5 flows between SRAs by data sources.

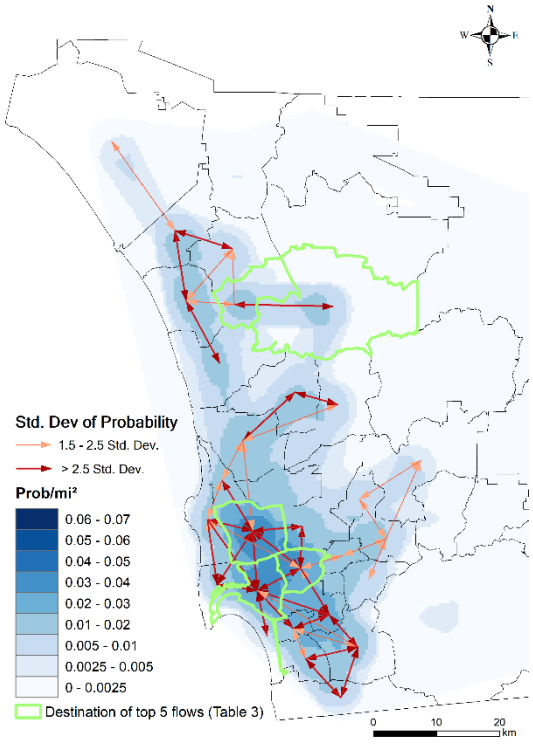
Rank	LODES		Twitter		Instagram	
	Origin	Destination	Origin	Destination	Origin	Destination
1	Southeastern	Central	Central	Kearny Mesa	Coastal	Central
2	Kearny Mesa	Central	San Marcos	Escondido	Kearny mesa	Central
3	Mid-city	Central	Kearny Mesa	Central	Central	Coastal
4	Central	Kearny Mesa	Kearny Mesa	Mid-city	Peninsula	Central
5	Del Mar-Mira Mesa	Kearny Mesa	Escondido	San Marcos	Central	Peninsula

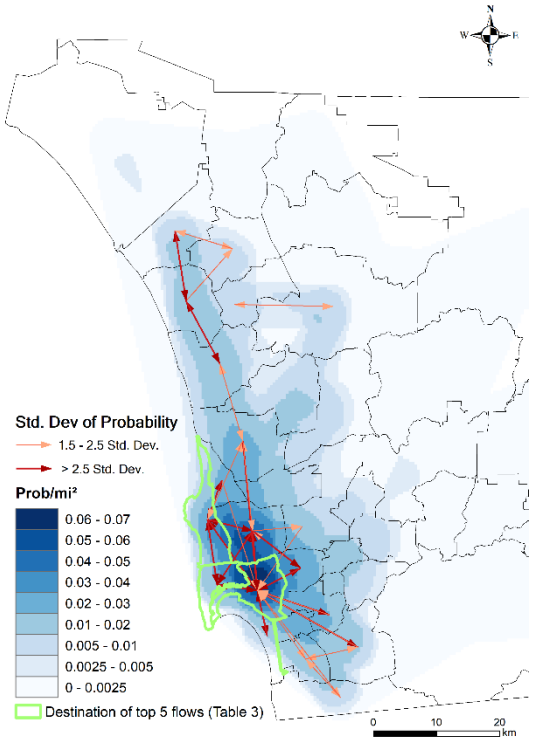

FALLBROOK
PENDLETON
PALOMA
PALOMAR-JULIAN
VALLEY CENTER
OCEANSIDE
VISTA
CARLSBAD
SAN MARCOS
ESCONDIDO
SAN DIEGUITO
RAMONA
NORTH SAN DIEGO
POWAY
MIRAMAR MESA
LAGUNA-PINE VALLEY
LAKESIDE
UNIVERSITY
MIRAMAR
ELLIOTT-NAVAJO
SANTEE
COASTAL
HARBISON CRLST
ALPINE
KEARNY MESA
EL CAJON
LAGUNA-PINE VALLEY
LA MESA
MID-CITY
PENINSULA
SPRING VALLEY
CENTRAL SAN DIEGO
SOUTHEASTERN SAN DIEGO
CORONADO
NATIONAL CITY
JAMUL
SWEETWATER
CHULA VISTA
MOUNTAIN EMPIRE
SOUTH BAY
Prob/mi²
0.147
0
Business District
0
10
20
km

(a)


Std. Dev of Probability
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.
Prob/mi²
0.06 - 0.07
0.05 - 0.06
0.04 - 0.05
0.03 - 0.04
0.02 - 0.03
0.01 - 0.02
0.005 - 0.01
0.0025 - 0.005
0 - 0.0025
Destination of top 5 flows (Table 3)
0
10
20
km

(b)


Std. Dev of Probability
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.
Prob/mi²
0.06 - 0.07
0.05 - 0.06
0.04 - 0.05
0.03 - 0.04
0.02 - 0.03
0.01 - 0.02
0.005 - 0.01
0.0025 - 0.005
0 - 0.0025
Destination of top 5 flows (Table 3)
0
10
20
km


Std. Dev of Probability
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.
Prob/mi²
0.06 - 0.07
0.05 - 0.06
0.04 - 0.05
0.03 - 0.04
0.02 - 0.03
0.01 - 0.02
0.005 - 0.01
0.0025 - 0.005
0 - 0.0025
Destination of top 5 flows (Table 3)
0
10
20
km

(c) (d)

Figure 2.3 Spatial distribution of probability of flows: (a) total; (b) LODES; (c); Twitter; and (d) Instagram.

Four maps in Figure 2.3 describe the density of total probability of flows in San Diego County and the spatial distributions of out-flows from LODES, Twitter, and Instagram, respectively. The majority of flows were concentrated in the western region of San Diego County corresponding with population distribution. Flows represented as an arrow in the maps are major flows where the number of flows is over 1.5 standard deviations from the mean of each source. To avoid over cluttering, only flows larger than +1.5 standard deviation are displayed. The most frequent flows from LODES moved into two regions (highlighted by the yellow border in Figure 2.3a), Central San Diego known as the Central Business District (CBD) of San Diego and Kearny Mesa known as a new business district. As compared to LODES, flows from social media revealed that frequent destinations are not limited to those two business districts. Twitter users visited comparatively diverse destinations whereas Instagram users preferred traveling to coastal areas such as Coastal and Peninsula SRAs (Table 2.4).

2.5. Results

2.5.1. SSIM and Sensitivity

To test the sensitivity of the OD pairs ordering, we generated two sets of reordered OD matrices of each data source and compared the results of SSIM and SpSSIM. We resampled the orders of pairs two times based on 1) the population size and 2) the alphabetical order of the SRA name. For SSIM, we tested 8 square windows, where the window size was increased from 5 to 40 cells by 5 cells. For SpSSIM, we set twelve distance bins where the

bin width is 10km and the distance range is from 0 to 120km. For each bin, we calculated SpSSIM using a spatial weight matrix determined by Equation 2.6. Cells of the nine heatmaps in Figure 2.4 represent the probabilities of movements and rows and columns are origins and destinations, respectively. The LODES heatmaps illustrate an apparent tendency of central places to move in with distinct lines, whereas Twitter maps reveal relatively diverse patterns and Instagram maps look mixed versions of the other two.

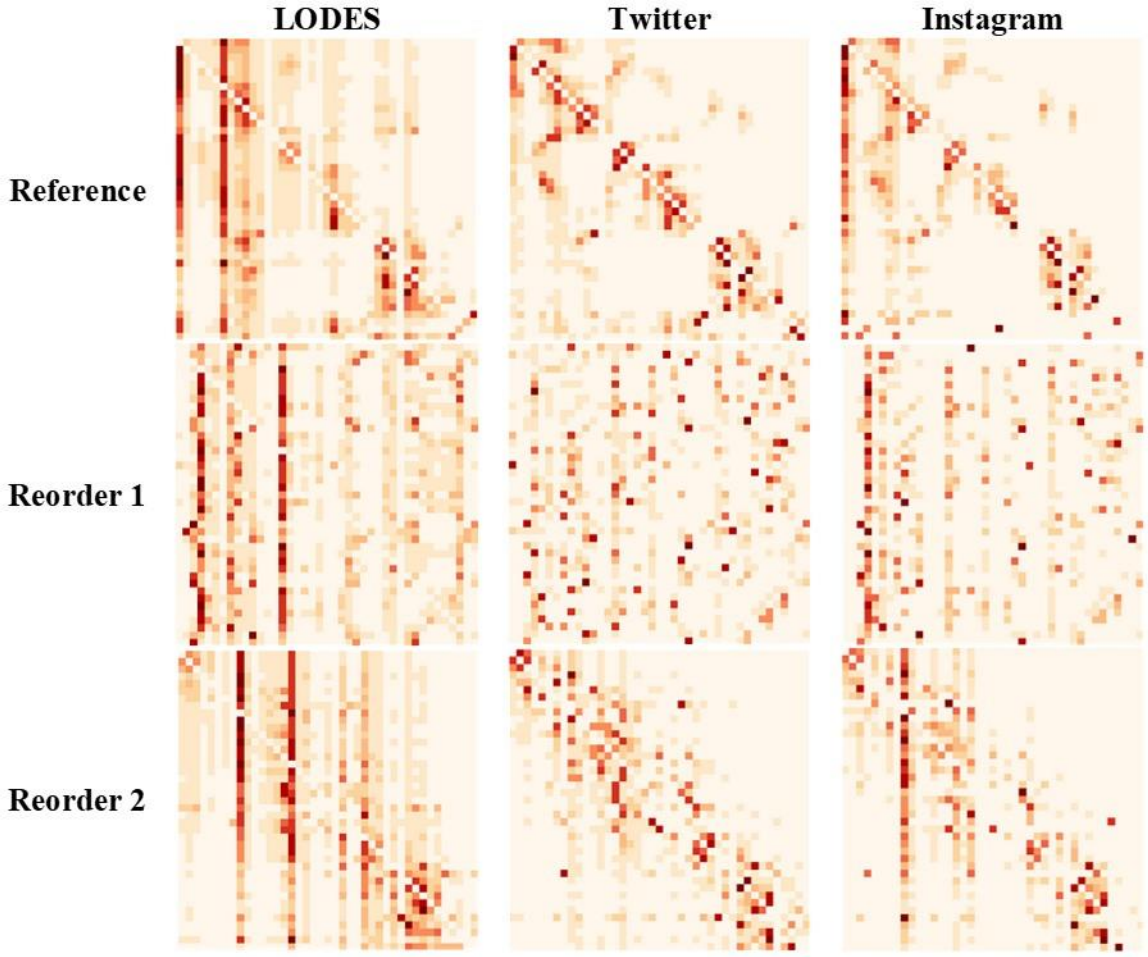


Figure 2.4 Heatmaps of OD pairs.

We calculated SSIM by varying sizes of windows to test the sensitivity of SSIM (Table 2.5). LODES and Twitter (L-T) have the lowest similarity while Twitter and Instagram (T-I) show similar patterns. However, the SSIM values differ by orders. For example, L-T have all different values of SSIM when the orders were shuffled regardless of the window sizes. Reordered OD matrices basically represent the same phenomenon, but SSIM is incapable to identify that they are the same patterns although the differences are slight. Moreover, SSIM fails to consider spatial correlations between OD pairs. Since the two matrices were reordered by population and names, it is hard to guarantee that contiguous cells in a window are spatially close to each other. The sensitivity issues of pair ordering make SSIM less reliable, and it is challenging to define pair orderings.

Table 2.5 The values of SSIM by window sizes.

OD \ Win Size		5	10	15	20	25	30	35	40
Reference	**L-T**	0.675	0.663	0.684	0.692	0.697	0.701	0.681	0.675
	L-I	0.715	0.729	0.754	0.758	0.760	0.743	0.703	0.668
	T-I	0.823	0.797	0.785	0.776	0.788	0.803	0.806	0.793
Reorder1	**L-T**	0.617	0.635	0.637	0.642	0.658	0.666	0.667	0.679
	L-I	0.682	0.692	0.694	0.692	0.683	0.670	0.650	0.644
	T-I	0.778	0.776	0.779	0.782	0.788	0.785	0.780	0.779
Reorder2	**L-T**	0.637	0.642	0.660	0.682	0.695	0.695	0.691	0.681
	L-I	0.687	0.676	0.658	0.646	0.636	0.632	0.630	0.644
	T-I	0.779	0.737	0.758	0.763	0.767	0.770	0.772	0.780

Contrary to SSIM, SpSSIM produces the same value regardless of the order of OD pairs. Because the weight matrices in SpSSIM define the spatial relationships between SRAs, reordering does not affect the spatial arrangements of OD pairs whereas SSIM compares contiguous cells in OD matrices regardless of OD pairs' spatial configurations. Table 2.6 shows the SpSSIM values calculated for twelve travel distance bins. Most SpSSIM values up to 80km travel distance ranges were significant at a 95% confidence level or better when

compared to a random distribution. It demonstrates that the similarity between mobility patterns from each pair of sources is statistically significant. These results imply that SSIM is less suitable to be a measurement of similarity than SpSSIM because the former fails to calculate the same values from the same events while the latter succeeds.

Table 2.6 The values of SpSSIM by spatial weight distances.

D_{min}^{max} (km)	L-T			L – I			T – I		
	SpSSIM	Mean	Std.dev	SpSSIM	Mean	Std.dev	SpSSIM	Mean	Std.dev
0 - 10	0.655***	0.335	0.085	0.682**	0.371	0.113	0.841***	0.290	0.107
10 - 20	0.744***	0.257	0.054	0.680***	0.328	0.076	0.732***	0.205	0.064
20 - 30	0.467***	0.237	0.049	0.617***	0.318	0.069	0.793***	0.192	0.060
30 - 40	0.377*	0.252	0.052	0.610***	0.333	0.079	0.657***	0.206	0.066
40 - 50	0.466**	0.271	0.060	0.717***	0.344	0.087	0.623***	0.221	0.070
50 - 60	0.523***	0.285	0.062	0.704***	0.348	0.093	0.713***	0.228	0.073
60 - 70	0.520**	0.298	0.068	0.395	0.357	0.101	0.774***	0.251	0.085
70 - 80	0.594**	0.339	0.088	0.715**	0.377	0.117	0.878***	0.290	0.107
80 - 90	0.568	0.371	0.115	0.207	0.385	0.134	0.458	0.328	0.133
90 - 100	0.637	0.401	0.147	0.515	0.390	0.157	0.772*	0.371	0.157
100 - 110	0.753	0.419	0.246	0.000	0.411	0.257	0.000	0.406	0.251
110 - 120	0.000	0.425	0.316	0.000	0.392	0.328	0.000	0.000	0.000
Global	0.525**	0.344	0.018	0.487**	0.366	0.005	0.603***	0.249	0.098

*****: 99.9%, **: 99%, *: 95% significance level**

2.5.2. SpSSIM in San Diego County

The SpSSIM values in Table 2.6 represent the degree of similarity in the mobility patterns derived from two different data sources by distances. Overall, the mobility patterns between social media in San Diego County were more similar to each other (Global SpSSIM (T-I) = 0.603) than to LODES flows (Global SpSSIM (L-T) = 0.525; Global SpSSIM (L-I) = 0.487). The mobility similarity between LODES and Twitter is relatively higher under 20km (0.655 and 0.744). It describes that Twitter users were more likely to make short trips where origins and destinations were similar to home and work locations reported in LODES in the travel distance range from 0 to 20km. The SpSSIM values, however, steeply decrease from 20 to 40km (0.467 and 0.377). The dissimilarity increases since there are much fewer

Twitter-based flows than LODES commuter-based flows in this distance range (Table 2.3). In addition, the majority of flows in LODES were heading to business districts such as Central San Diego and Kearny Mesa (see Figure 2.3b) while the destinations of Twitter users were diverse including beach areas (South Bay, Oceanside, and Del Mar-Mira Mesa) and parks (Sweetwater and Poway) as well as business districts (see Figure 2.3c). From 40km, the SpSSIM value gradually increases again by 110km since the probabilities in longer distance trips were close to zero in both data sources. To note, the SpSSIM values are not statistically significant over the range of 80km. Compared to Twitter, mobility patterns from Instagram were less similar to LODES. Similar to other comparison, the SpSSIM values of LODES and Instagram are relatively higher within 0 to 20km. Unexpectedly, however, the similarity between LODES and Instagram within 40 to 60km peeked. A potential explanation of this pattern is that remarkable places in San Diego are concentrated in downtown and coastal area, where also have many jobs. This similarity is also observed from Table 2.3 and Figure 2.3b and 2.3d. The most frequently visited destinations from two datasets are quite similar. It indicates that Instagram users are more willing to move longer distance than Twitter if attractions are far away.

On the other hand, travel patterns derived from Twitter and Instagram resemble each other. The SpSSIM value of Twitter and Instagram in the range of 0 to 10km (0.841) is the highest in the same distance range bin and the second highest among all SpSSIM values. This explains that short daily trips observed from Twitter and Instagram users share similar origins and destinations, which are clustered in Central San Diego, Kearny Mesa, Peninsula, and Mid-City (Figure 2.3c and 2.3d). The SpSSIM value of Twitter and Instagram slightly decreased in the range of 10 to 20km. Travel destinations of Twitter users in this distance

range included suburb regions such as Escondido, San Marcos, Pendleton, and North San Diego, whereas those of Instagram users were concentrated in the downtown region of San Diego City such as Central San Diego, Coastal, and Peninsular. The SpSSIM values further decreased as the distance range increased to 30 to 50km because the total probability of mobility derived from Twitter in the range (0.06) was smaller than that from Instagram (0.09). In addition, destinations of Twitter movements were more scattered than those of Instagram (Table 2.4).

2.5.3. Localized SpSSIM

Localized SpSSIM helps investigating similarity in terms of local in-flows and out-flows. Figure 2.5 demonstrates an example of comparing localized in-flow SpSSIM between LODES and Twitter from 10 to 40km. Since in-flows describe the number of people moving into a region, it represents the popularity of places. The mobility patterns from LODES and Twitter were less similar in Southeastern San Diego (highlighted by the blue border) within 20km than other regions indicating the lowest SpSSIM values in each range (0.074 and 0.120 respectively) while the global SpSSIM is the highest in the range 10–20km (0.744). On the other hand, the Del-Mar-Mira Mesa region (green border in Figure 2.5) showed different patterns. While the SpSSIM value was high in the range of 0 to 20km, the value dropped from 30km. In the range of 20 to 40km, SpSSIM values were 0.289 and 0.181 respectively while the global values are 0.467 and 0.377. Although the dissimilarities were not as large as Southeastern San Diego, it denotes that the flows moving into the region dramatically changed from 30km.

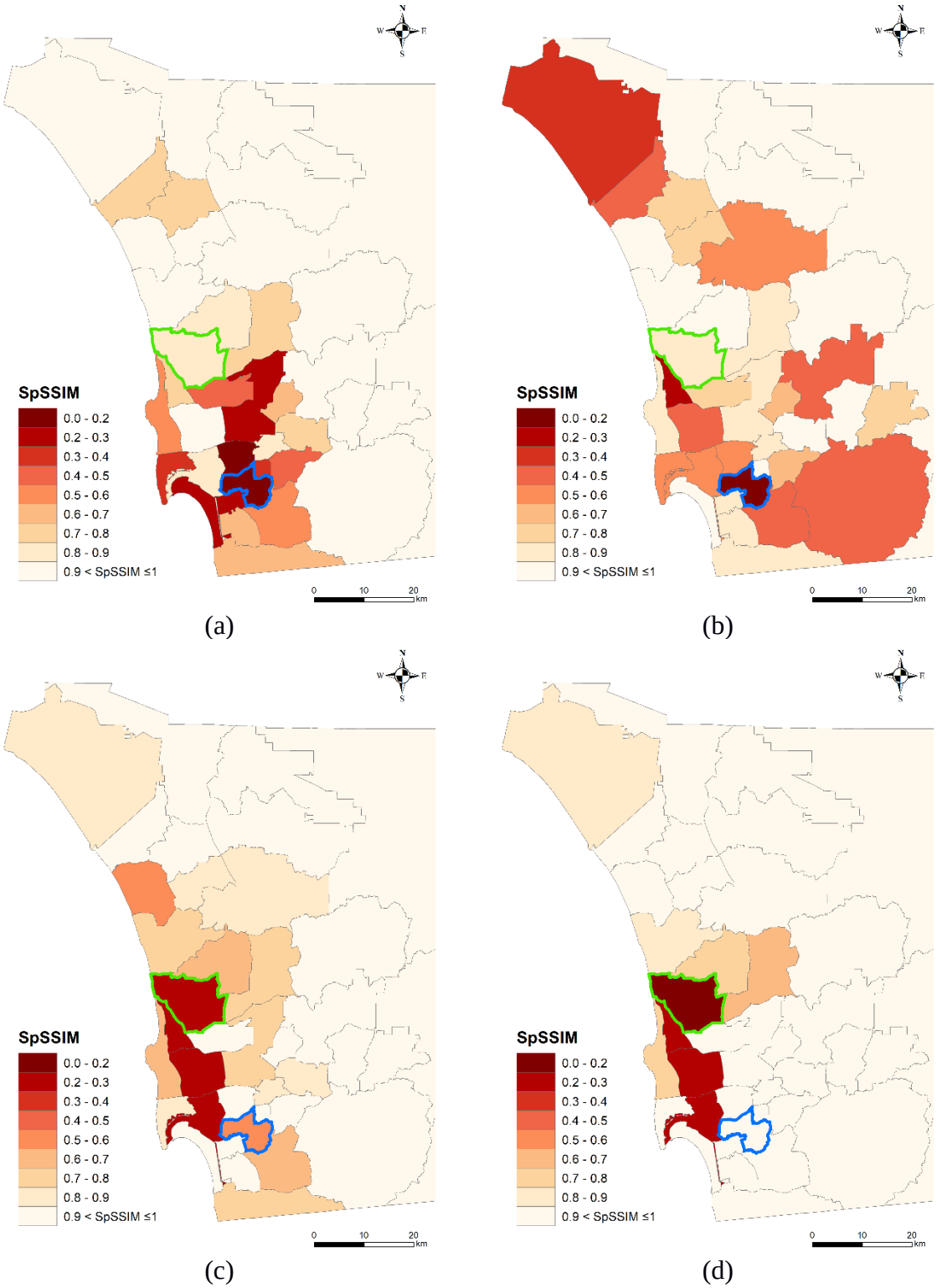


Figure 2.5 Localized SpSSIM (in-flows of LODES-Twitter): (a) 10km; (b) 20km; (c) 30km; and (d) 40km.

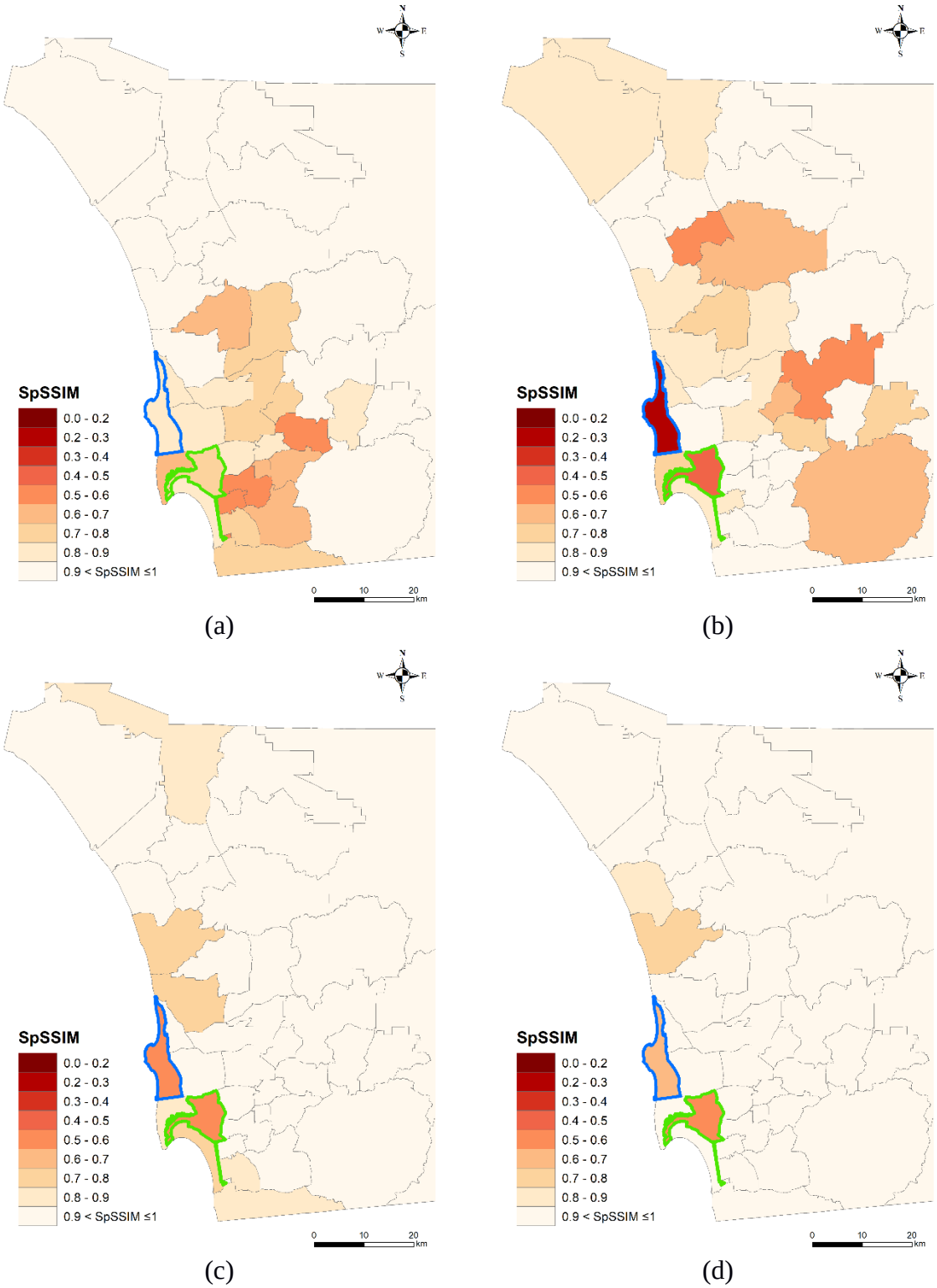
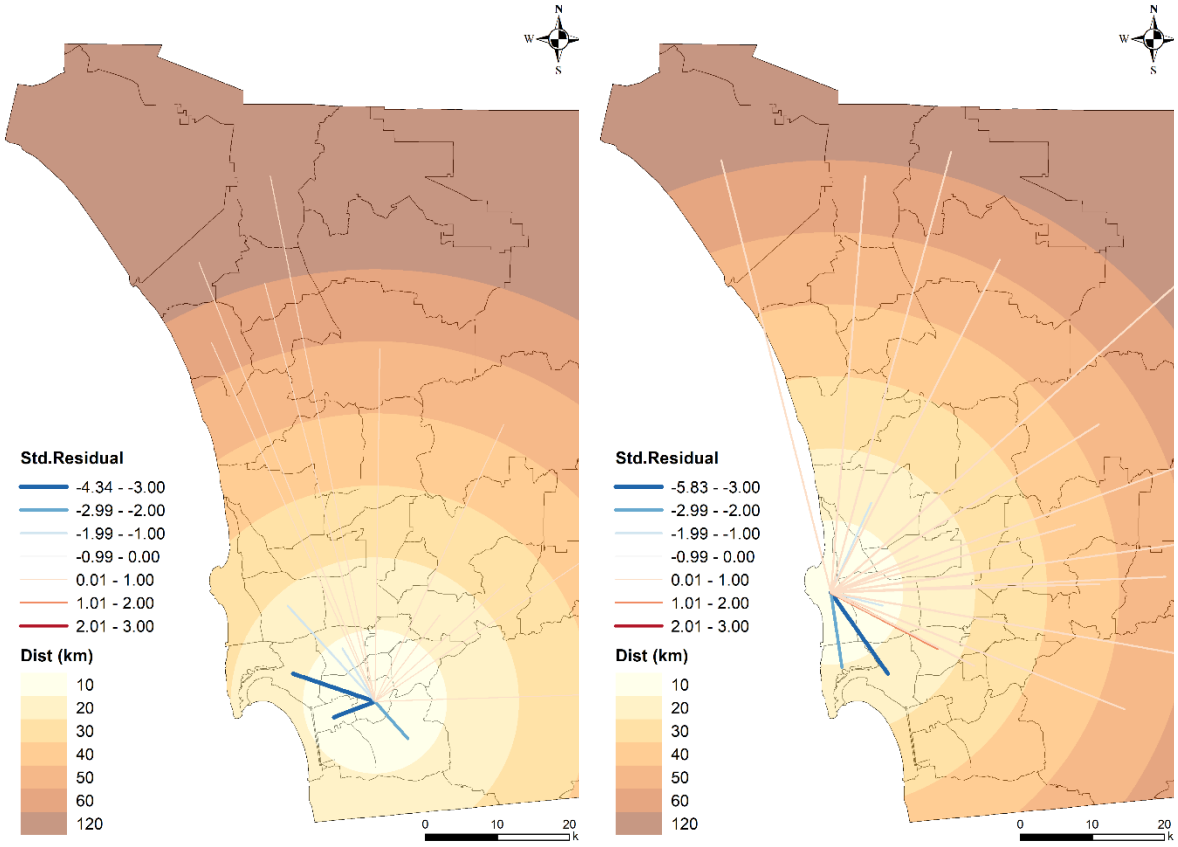


Figure 2.6 Localized SpSSIM (in-flows of Twitter-Instagram): (a) 10km; (b) 20km; (c) 30km; and (d) 40km.

Figure 2.6 illustrates localized SpSSIM of in-flows between Twitter and Instagram in the range of 0 to 40km. Unlike the comparison of LODES and Twitter, most regions show relatively high values indicating that Twitter and Instagram have similar mobility patterns. Within 10km, the lowest localized SpSSIM value is 0.577 at Southeastern San Diego SRA denoting that short trip patterns of Twitter and Instagram users are relatively very similar. However, in the range of 10 to 20km, Coastal (blue border in Figure 6, localized SpSSIM = 0.298) and Central San Diego (green border in Figure 6, localized SpSSIM = 0.404) display dissimilarity patterns compared to the global value in the range (0.732). This suggests that in-flow mobility patterns from two social media data in these two regions present notable dissimilarity when travel distances are 10-20 km (Figure 2.6b) or longer (Figure 2.6c and 2.6d).

To further investigate the dissimilarity of the localized SpSSIM in those two regions, we mapped standardized differences of in-flows between two data sources by dividing the differences by the standard deviation of the difference. Figure 2.7a illustrates the standardized differences between LODES and Twitter regarding in-flows into Southeastern San Diego SRA. The negative values in Figure 2.7a represent that the probability of movements derived from Twitter was larger than LODES. In other words, more Twitter users entered into Southeastern San Diego than LODES-based commuters within 20km. In particular, the movements of Twitter users between Central San Diego and National City, within 20km from the origin, outnumbered LODES movements. The localized SpSSIM detects the large differences with significantly low values (0.074 and 0.120). Similarly, Figure 2.7b demonstrates the differences between Twitter and Instagram flows moving into Coastal SRA. Blue lines with negative values describe more Instagram users entered into the

region than Twitter users. The localized SpSSIM also depicts a large inflow of Twitter users from Mid-City to Costal SRA (orange line) when traveling distances are within 20km (Local SpSSIM=0.298) while most shorter distance inflows to Costal SRA are dominated by Instagram users.



(a) (b)

Figure 2.7 Standardized difference of in-flows: (a) Southeastern San Diego (LODES-Twitter); and (b) Coastal (Twitter-Instagram).

Although the localized SpSSIM values measure the differences between mobility patterns formed from OD matrices of diverse sources, the measurement itself does not provide the contexts behind the (dis)similarity. Here we provide potential explanations for the discovered patterns. Dissimilarities between LODES and Twitter in Southeastern San Diego (Figure 2.7a) can be explained by the socioeconomic backgrounds. Southeastern San

Diego SRA has been one of the most impoverished areas in San Diego County (Joassart-Marcelli et al. 2014). Due to its economic decline, there have been fewer job opportunities and attracting places. Furthermore, the region has been historically more ethnically diverse than other areas. According to ACS 5-Year (2007 – 2011) estimation, the Hispanic population took over 50% of the total population followed by the African and African-American population (18.2%). Due to lack of job opportunities in the region, LODES commuter-based in-flows was identified as low. On the other hand, Twitter has been more popular among Hispanic African and African-American than other ethnic groups (Krogstad, 2015), which can explain relatively larger in-flows into the region by Twitter users.

Discovered patterns can be also described by the differences in the social media platform usage. Gao (2015) and Steiger et al (2015) pointed out that Tweets are most likely associated with home and workplace activities. Instagram, on the other hand, has more geotagged pictures (18.8%) as compared to Twitter (0.6%). Since Twitter posts are more related to home and workplace activities, the similarity of LODES-Twitter is larger than LODES-Instagram (Table 2.6). Moreover, the dissimilarity of Twitter and Instagram in terms of in-flows into Coastal SRA (Figure 2.6) can come from the fact that there are many photogenic spots in the coastal region, where people are willing to share their experiences.

2.6. Conclusion

In this research, we introduced a new method, SpSSIM, to measure the similarity between two OD matrices. We demonstrated the capability of SpSSIM by conducting a case study to compare three mobility data sources, LODES, Twitter, and Instagram in San Diego County. In addition, we assessed the OD matrix ordering problem in SSIM when it is applied to spatial datasets. Our sensitivity test verified that SpSSIM is robust regardless of

the arrangement of OD pairs while conventional SSIM is sensitive. This is achieved by employing a series of spatial weight matrices to resolve the sensitivity to the spatial configuration. SpSSIM is also statistically verified through bootstrap which generates the hypothetical distribution of SpSSIM.

Our case study revealed notable similarities and differences in the mobility patterns from three data sources. In general, the mobility patterns of two social media, Twitter, and Instagram, resembled each other more so than when compared to LODES. The most frequent destinations of LODES were distributed in business districts while social media users were traveling to diverse points of interests. This is expected results since LODES flows are specifically comprised of employment-based home-work trips whereas Twitter and Instagram flows are based on social media users who have various purposes for trips. SpSSIM can depict similarity over travel distances. For example, the similarity between LODES and Twitter increased in the range of 0 to 20km and steeply decreased in the range of 20 to 40km, which is explained by the sparseness of mobility probability and diversity of destinations in Twitter. Furthermore, SpSSIM can help discovering local outliers by mapping localized values. We demonstrated in-flow (dis)similarities of LODES-Twitter and Twitter-Instagram. The localized SpSSIM values quantify and characterize the local mobility from different data sources by geographic distances.

While SpSSIM can successfully measure the similarity between two mobility datasets in the form of OD matrices, SpSSIM has two limitations. The first issue is that the distance ranges are arbitrary defined and SpSSIM values are sensitive on them. SpSSIM utilizes a series of pre-defined distance bins to overcome the sensitivity of OD pairs and window sizes in SSIM. Further research is required to understand the sensitivity of defined distance bins.

However, we argue that the use of the distance instead of the window size in SSIM can provide an explainable frame in terms of spatial context. Thus, SpSSIM helps measuring the (dis)similarity of movements occurring in a specific spatial boundary in which researchers are interested. As another limitation, SpSSIM does not provide the amount of flow difference. Therefore, it is necessary to map the differences to understand the contexts of discovered mobility patterns (Figure 2.7). Nevertheless, SpSSIM, as an exploratory tool, provides spatial distribution of similarity with localized values and better understanding of the human dynamics and complexity in urban system. By detecting outliers, researchers can selectively focus on investigating regions with high (dis)similarity and further study mobility contexts in those regions. In sum, this study provides a methodological approach to comparing mobility patterns in spatial context and deepens our understanding of social media data in mobility analysis.

3. Exploring Micro-scale Spatiotemporal Dynamics in Urban Space[2]

3.1. Introduction

Competition in commercial activities has a significant impact not just on the success of individual businesses, but also on urban structure and economy (Jung and Jang 2019; Vandell and Carter 1994). Within the tertiary industry, restaurants are one of the most common services, with an interesting tendency to open and close more frequently than other types of businesses. The diverse locations and popularity of restaurants aid in understanding changes in urban space since they regenerate or revitalize neighborhoods by attracting consumers and investment (Zukin et al. 2017; Zhai et al. 2015; Hyde 2014). Although many socioeconomic factors can contribute to the success and failure of restaurants, the precise location of the restaurant usually plays a crucial role in entrepreneurship (Dock et al. 2015; Ghosh and Craig 1983). Some locations can be more accessible and profitable than other locations (Church and Murray 2009). Moreover, they compete and seek to gain strategic advantages, consequently generating agglomerations when successful (Li and Liu 2012; Hotelling 1929). However, locational advantages are not necessarily permanent (Prayag et al. 2012). Over time, once popular districts may become dilapidated, have outdated designs, and fail to reflect current trends or preferences of consumers. Therefore, identifying spatial

[2] This chapter represents a revised version of a paper published in the International Journal of Geospatial and Environmental Research.

Jin, C., and Murray, A. T. (2021). Exploring Public Open Data: Spatiotemporal Dynamics of Restaurant Entrepreneurships in Seoul, Korea, *International Journal of Geospatial and Environmental Research*, 8(3). 5.

patterns and temporal fluctuations is key to deepening our understanding of urban dynamics, providing insights on drivers of economic growth and development.

Big data analysis has been recently important in better understanding human and environmental complexities (Singleton and Arribas-bel 2021) and capable of discovering new knowledge about urban dynamics (Miller and Goodchild 2015). The advancement of information and communication technologies, computation technologies, and location-aware technologies enables the generation of large and diverse data in real-time, making them accessible for broad usage (Shaw et al. 2016). For example, location-based social networks and consumer review services provide information on people, revealing where they are and their impressions through check-in, messages, and ratings. New types of data help to fill research gaps on micro-scale urban dynamics, including shop preference and restaurant popularity (García-palomares et al. 2018; Steiger et al. 2015; Tsou 2015). However, it remains a challenge to utilize these data in urban studies when temporal changes are significant due to limited history and a lack of detailed information, such when the businesses opened, closed, etc.

On the other hand, public open data, or government administrative data, is routinely collected by authorities for public purposes, including welfare, taxation, and licensing (Lansley et al. 2018). As a governmental tool, it generally covers an entire population and is regularly updated. Although public data has a long history and is considered "officially approved," its use has been limited in geographic studies at the individual level because of spatiotemporal aggregation. However, as interest in open data has increased, governments have made some raw individual data available to the public (Arribas-bel 2014). As an example, the U.S. and U.K. recently launched websites to share the governmental data with

other countries, including Japan and South Korea. It is expected that such data can help improve the effectiveness of public policies involving socioeconomic issues.

To understand dynamic patterns given increasing amounts of data, the importance and necessity of exploratory data analysis has become apparent. Exploratory approaches are generally grounded in statistics, supporting the identification of unique patterns in data before assuming hypotheses based on theories (Tukey 1962). Such an approach enables detection of underlying spatiotemporal trends in complex urban dynamics as the amount of fine scale data grows (Miller 2010; Mazur and Manley 2016). Many studies have recently utilized Big Data, including call records and social media, to understand diverse patterns of human activities in urban spaces, effectively overcoming limitations associated with data reported via traditional aggregated geographic units (Tu et al. 2017). For example, using the check-in information from a consumer reviewing service, Zhai et al. (2015) and Zhang et al. (2021) discerned popular places in a city through exploratory analyses involving kernel density estimation and cluster detection. With public open data in Seoul, Korea, Kim et al. (2021) and Lee et al. (2020) explored spatiotemporal patterns of restaurants by analyzing the annual number of openings and closings of restaurants. However, exploring spatial patterns of restaurants as temporal snap shots provides only a partial understanding of spatiotemporal dynamics because business lifespans are continuously changing.

In this chapter we explore spatiotemporal dynamics in the entrepreneurship of restaurants through the application of three exploratory approaches using public open data. While several studies have analyzed social media and consumer review services data with points of interest to identify the popularity of places (Zhang et al. 2021; Widaningrum et al. 2020; Zukin et al. 2017; Li et al. 2013), they have not focused on the survivability of

businesses. Because government agencies manage licensing of businesses, it is possible to explore the spatiotemporal changes in not only openings and closings of restaurants but also their lifespans. Among many businesses in a city, we focus on restaurants because of constant change and the significant impacts they have on local structure (see Zukin et al. 2009). We also advance replicability efforts in research by analyzing freely accessible public datasets, doing so using methods available through open software (Kedron et al. 2021; Murray et al. 2013; Newman 2010). This study will investigate continuously changing spatiotemporal patterns of restaurant clusters in Seoul, Korea with public open data at micro scales to enhance our understanding of urban dynamics, providing a foundation for diverse planning and decision-making in cities.

3.2. Location of restaurant business

3.2.1. Location theories for restaurants in cities

Restaurants have played a major role in economic growth of cities by providing jobs, tourism, and regenerating and revitalizing neighborhoods by attracting more consumers and investment (Self et al. 2015; Hyde 2014; Zukin 2009). Therefore, understanding their location patterns is a key issue for not only individual businesses but also for public policy and decision-makers. Seminal theories have been developed, often supported by empirical studies associated with retail location success in urban environments (Hurst 1972). Central place theory highlights that there is a maximum distance that people are willing to travel in consuming a good or service as well as requirements for a minimum level of demand to support a business. These two essential concepts explain why service providers prefer certain locations in central business districts with a larger customer base (Church and

Murray 2009; Austin et al. 2005; Mulligan 1984). However, restaurant patronage is more complex, relying on intra-urban patterns of travel (Smith 1985). While the gravity law explains some aspects of trade area behavior (Huff 1964), particularly retail and restaurant locations (Li and Liu 2012), it cannot fully explain clustering effects.

Competition between vendors based on Hotelling (1929) offers some insights for the restaurant industry. Co-location in accessible areas results in profit maximization, provided total demand is sufficient. That is, agglomeration of businesses generates positive externalities, so restaurants gain an advantage by clustering together as long as the market is not saturated (Jung and Jang 2019). Since restaurant clusters provide favorable environments for enhanced food options and reduced costs of shared facilities, the spatiotemporal patterns of clusters are important (Prayag et al. 2012). Sun and Paule (2017), for example, detected restaurant clusters from Yelp reviews. As another example, Minner and Shi (2017) argued that spatial clusters of locally owned restaurants in commercial strips are signs of redevelopment. Recently, point of interest (POI) data enabled Zhang et al. (2021) to distinguish areal characteristics based on differences in clusters of local and non-local restaurants and Widaningrum et al. (2020) found spatial clusters of fast-food restaurants. Although these studies highlight the importance of restaurant clusters, rarely illustrated is their temporal change such information is lacking in social media and consumer review services data.

3.2.2. Restaurants in Seoul, South Korea

The South Korean restaurant sector represents a relatively large percentage of industry compared to other developed countries. According to the 2017 economic census in South Korea, 12.36% (496,915) of all establishments were restaurants while in the USA they

represent 6.05% (598,656) of all establishments. However, the portion of employees in this sector was only slightly larger in South Korea (7.29%) than the USA (6.68%). Thus, the average number of employees per business is only 3.17 in Korea whereas in the USA it is 16.80. Furthermore, 95.81% of restaurants in Korea were operated by sole proprietors and 96.81% were single stores, having no headquarters or other locations. Low entry barriers encourage starting new businesses, but causes a saturated market, with economic fluctuations posing a risk to marginal operations. As small businesses comprise a large portion of the national economy in South Korea, a significant number of failures can lead to not only individual but also nationwide socioeconomic issues, such as a high unemployment (Kim and Lee 2019).

Many studies have investigated spatiotemporal patterns of restaurants in Seoul, the capital and socioeconomic center of South Korea. Competition in Seoul is greater than most other cities. Shin and Shin (2009) found that restaurants tend to be concentrated in the central business district (*Jung-gu* and *Jongno-gu*) and the other centric regions, including *Gangnam* and *Seocho-gu* and *Mapo-gu*, relying on large demand from nearby office workers and young adults. Yu and Lee (2017) also investigated restaurant clusters in Seoul and categorize them by factors contributing to their level of agglomeration. While restaurants in the central business district depend on commuter demand, those in *Mapo-gu* are oriented to more diverse types of consumers such as tourists. Although these studies explain urban structure through the spatial configurations of restaurants, they are limited to recent changes, lacking sufficient temporal information about evolving spatial patterns.

The South Korean government made data on business licenses available, encompassing location and opening/closing information, making research possible for exploring

spatiotemporal patterns of all businesses, no just samples. However, many studies have been limited in explaining citywide changes because they have focused solely on well-known commercial areas. Jeong and Yoon (2017) compared the survivability of restaurants along main streets and back streets in the *Itaewon* region, a prominent multicultural district in Seoul. Kim et al. (2018) illustrated expansions of commercial areas in *Hongdae* region, a well-known campus town with four prestigious universities. At a broader scale, Ryu and Park (2019) classified five popular commercial areas based on changes in types of businesses. Lee et al. (2020) attempted to follow temporal changes in restaurant clusters with yearly changes in density of operating restaurants. Kim et al. (2021) compared spatiotemporal differences in opening and closing restaurants before and after COVID19, but the temporal differences were not significant because the pandemic has persisted. Although these approaches are valuable spatiotemporal dynamic snapshots, they do not illustrate continuous changes at the city level.

3.3. Methods

To better understand the spatiotemporal dynamics in restaurant entrepreneurship at the city level, we employ three exploratory methods in a space-time framework: 1) spatial hot spot analysis, 2) trend analysis of clusters, and 3) spatiotemporal scan statistics with exponential models. Collectively, these analytic approaches facilitate accessibility of public open data since they are available as open source software, making the analysis relatively easy to replicate.

3.3.1. Spatial hot spot analysis

Methods to identify (dis)similarities in geographic events have been developed and widely applied in a variety of fields (Getis 2008). Global spatial autocorrelation measures the relationship of a variable across spatial units. Although global statistics are useful to evaluate the strength of spatial dependence between spatial units in a region, they are limited in identifying whether and where similarity or dissimilarity occurs. As an alternative, local statistics, such as the local indicator of spatial autocorrelation (LISA) (Anselin 1995) and G_i^* (Getis and Ord 1992), focus on interaction between neighboring units. Among many local autocorrelation indices, G_i^* is an effective measure/tool to identify statistically significant clusters. G_i^* calculates the local average in a neighborhood as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{X} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\frac{[n \sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \quad 3.1$$

where x_j refers the value in neighbor of target unit, i, $\bar{X}$ indicates the local mean of the value ($\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$), and $S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{X}^2}$. As Equation 3.1 facilitates the comparison of values in neighborhood i, it can detect either hot or cold spots by identifying areas with relatively high or low clustered values. Hot spots reflect a high number of clustered start-ups, and can be considered as booming or popular areas, whereas cold spots represent a declining area. In this study, we analyze the number of currently operating restaurants and the number of net start-ups by subtracting the number of closed restaurants from the number of opened restaurants at *dong* level. The neighborhood is defined as the area sharing edges of the target unit.

3.3.2. Trend analysis of clusters

Trend analysis of clusters traces temporal changes via the local spatial autocorrelation index, G_i^*. Although the index is useful for exploring spatial distributions of geographic events, it is limited in its ability to identify temporal changes in spatial patterns. The Mann-Kendall statistic enables a test of temporal relationships between different time steps fora spatial unit (Kendall 1948; Mann 1945). As a rank correlation analysis, the test determines statistically significant temporal trends by comparing values in a time sequence as follows:

$$S = \sum_{t}^{n-1} \sum_{p=1}^{n} sgn\left(x_{i,t+p} - x_{i,t}\right) \quad 3.2$$

where $x_{i,t} = \begin{cases} 1, z_{it} > z_{i,t-1} \\ 0, z_{it} = z_{i,t-1} \\ -1, z_{it} < z_{i,t-1} \end{cases}$, (t = 0, 1, … n).

If the current value of standardized G_i^*, $z_{i,t}$, is larger than the previous value, $z_{i,t-1}$, the result of the paired comparison is 1. On the other hand, a result of -1 occurs when $z_{i,t-1}$ is larger. The paired results are summed by units and compared to the null hypothesis that trends do not exist over time (S = 0). As a result, the analysis categorizes current clusters into one of eight types of hot and cold spots.

For this study, we aggregate restaurants into a space-time cube defined by *dong* and year, and calculate the number of net start-ups in each spatiotemporal bin by counting 1 when a shop opens and -1 when a shop closes. In other words, the value of each bin in the space-time cube represents the number of net start-ups in a year from January 1st to December 31st at the *dong* level. The spatial neighborhood is defined as the area sharing edges of the target unit and each bin is analyzed in comparison to the entire time period.

3.3.3. Spatiotemporal Scan Statistics

The spatial and space-time scan statistics suggested by Kulldorff (1 The spatial and space-time scan statistics suggested by Kulldorff (1997) have been widely used to detect clusters of geographic events such as disease outbreaks (Smith et al. 2015) and crimes (Leitner and Helbich 2013; Nakaya and Yano 2010). Initial statistics relied on the Bernoulli probability distribution of binary events, whether it happens or does not happen. It detects subareas in which events more (or less) frequently occur than they do in other areas by scanning the study region using a moving window. However, the Bernoulli assumption is restricted to spatiotemporal disparities of continuous variables, such as lifespan of diseases. Huang et al. (2007) propose using an exponential distribution for these scan statistics to find lower (or higher) survivability areas, and it has been widely used to identify geographic disparity of survivability in diverse diseases (Lin et al. 2015; Wan et al. 2012; Henry et al. 2009). Let θ_z represent the mean survival time for each individual inside a subarea, *Z*. The null hypothesis is that there is no difference in the mean survival time inside or outside the subarea (e.g., H_0: $\theta_z = \theta_{Z'}$). The likelihood function for the exponential case is:

$$L(Z, \theta_Z, \theta_{Z'}) = \frac{1}{(\theta_Z)^r} e^{-\sum_i \frac{t_i}{\theta_z}} \frac{1}{(\theta_{Z'})^{r'}} e^{-\sum_j \frac{t_j}{\theta_{z'}}} \quad 3.3$$

where *t* is the survival time of an individual ($i \in Z, j \in Z'$) and *r* the number of non-censored individuals ($R = r + r'$; $G = Z + Z'$). Under the alternative hypothesis (H_α: $\theta_z \neq \theta_{Z'}$), one is interested in the zone $\hat{Z}$ that maximizes likelihood function (3). This can then be derived as:

$$\lambda = \frac{max_{Z,\theta_Z \neq \theta_{Z'}} L(Z, \theta_Z, \theta_{Z'})}{max_{Z,\theta_Z = \theta_{Z'}} L(Z, \theta_Z, \theta_{Z'})} = \frac{L(\hat{Z})}{L_0} = \frac{max_Z (\frac{1}{\theta_Z})^r (\frac{1}{\theta_{Z'}})^{r'}}{(\frac{1}{\theta_G})^R} \quad 3.4$$

The statistical significance of λ is assessed using a p-value generated through Monte Carlo simulation. Therefore, a detected cluster indicates that individuals in the subarea are significantly shorter (or longer) in survival than outside of the subarea if the null hypothesis is rejected.

The spatiotemporal scan statistic under exponential conditions is applied to survival data for restaurants in order to assess not only spatial clusters but also which restaurants survive shorter (or longer) durations. If a restaurant closed before the end of 2018, it is regarded as non-censored, and its survival time is counted from its opening date to the closing date. On the other hand, for a restaurant still operating at the end of 2018, its survival time is counted from its opening date to 12/31/2018 by censoring data. The maximum size of each cluster is restricted to 50% of the total individuals, and the p-value of each cluster is derived from 999 random permutations.

3.4. Data

Local governments in South Korea, *si-gun-gu*, have the authority to approve new businesses. All businesses are required to report their closure to their local government. South Korea recently made this business data available, detailing business types, location, starting date, and closing date. The data covers 191 types of business, such as markets, residential services, and restaurants, and updates in real time are available through an open application programming interface. It has been used in a wide range of fields, including academia, public, and private sectors, because it is regarded as a population versus a sample

of businesses. With a focus on restaurants (general food services and drinking places), 1.7 million records were identified at the end of 2018. Seoul has 21% of total restaurants whereas the population and household represent 18.9% and 19.6% of the country, respectively. Due to uncertainty in old records, we only use currently operating restaurants and those that closed after 2000.

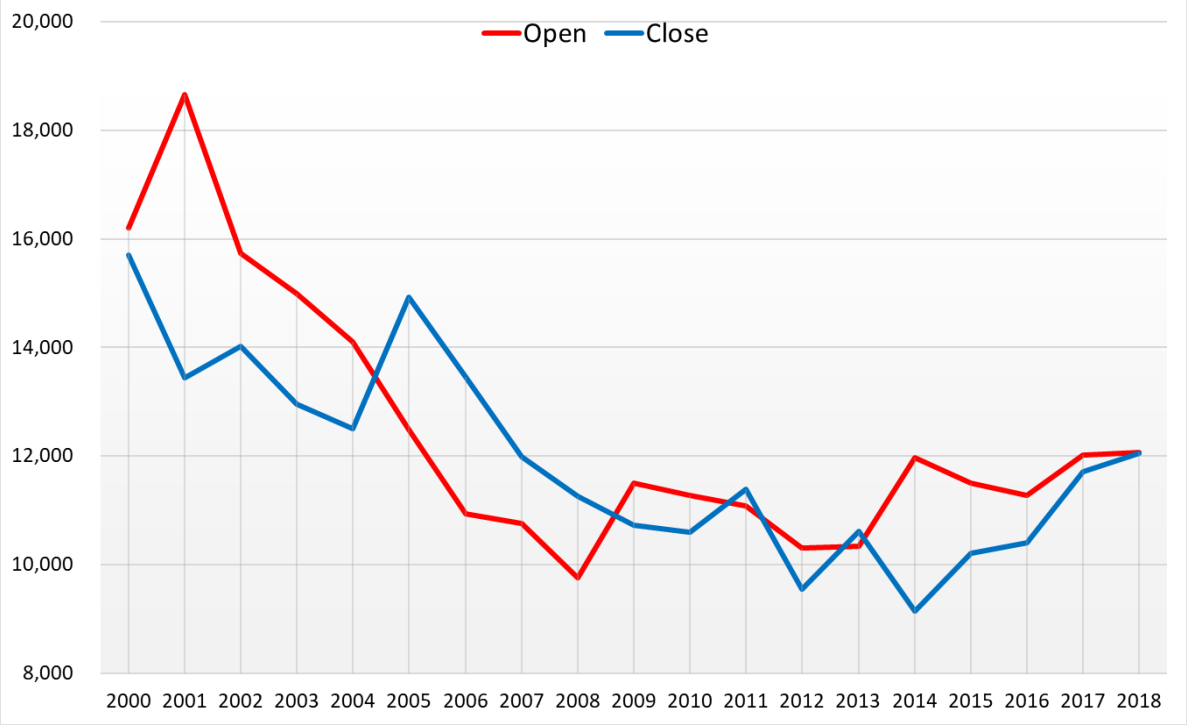


Figure 3.1 The number of opening and closing restaurants from 2000 to 2018.

Since 2000, 346,628 restaurants have opened and closed in Seoul. The number of start-ups peaked in 2001 with 18,659, but steadily declined until reaching a low in 2008. The number of closures outnumbered starts-up beginning in 2005 until 2008. The number decreased by 2014 but increased again, whereas the number of opening restaurants had been quite constant. As a consequence, the number of openings and closings in 2018 are almost even (Figure 3.1). Figure 3.2 suggests an exponential distribution with a long right tail associated with survival times of restaurants. The overall average of the life span is 8.24

years, and the median is 5.73 years. However, closed restaurants have survived for 6.98 years on average and 31.4% of businesses have failed in 3 years.

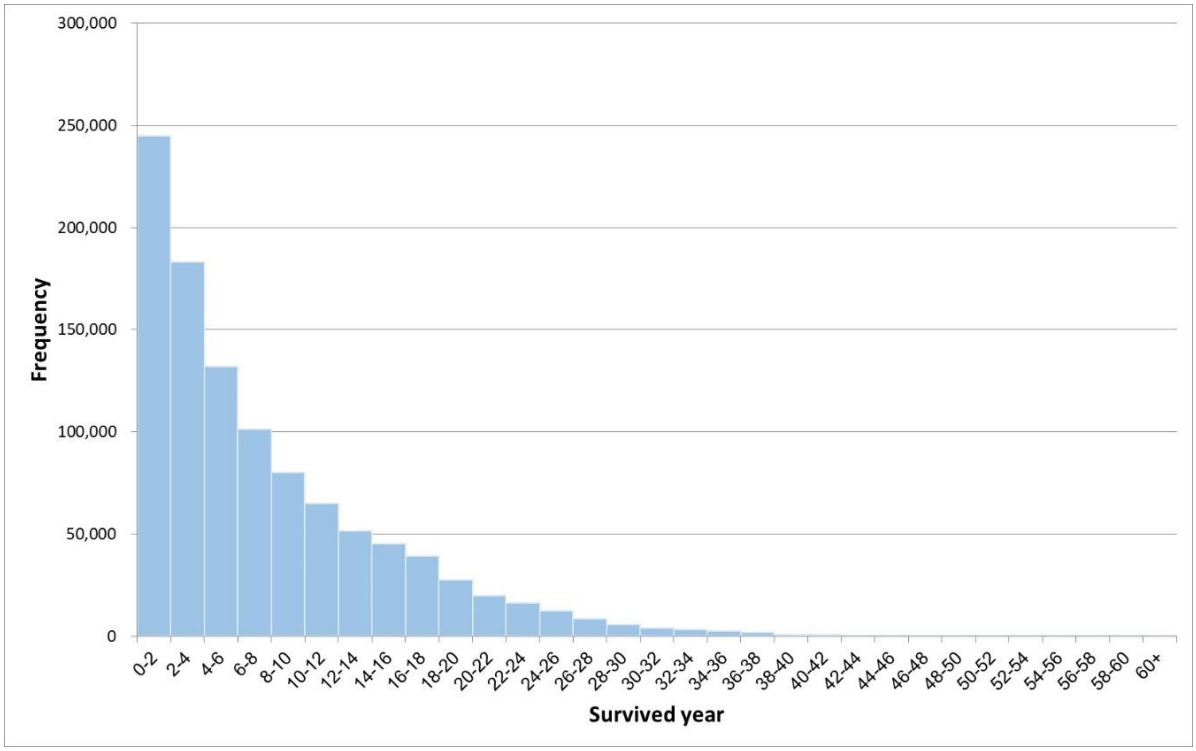
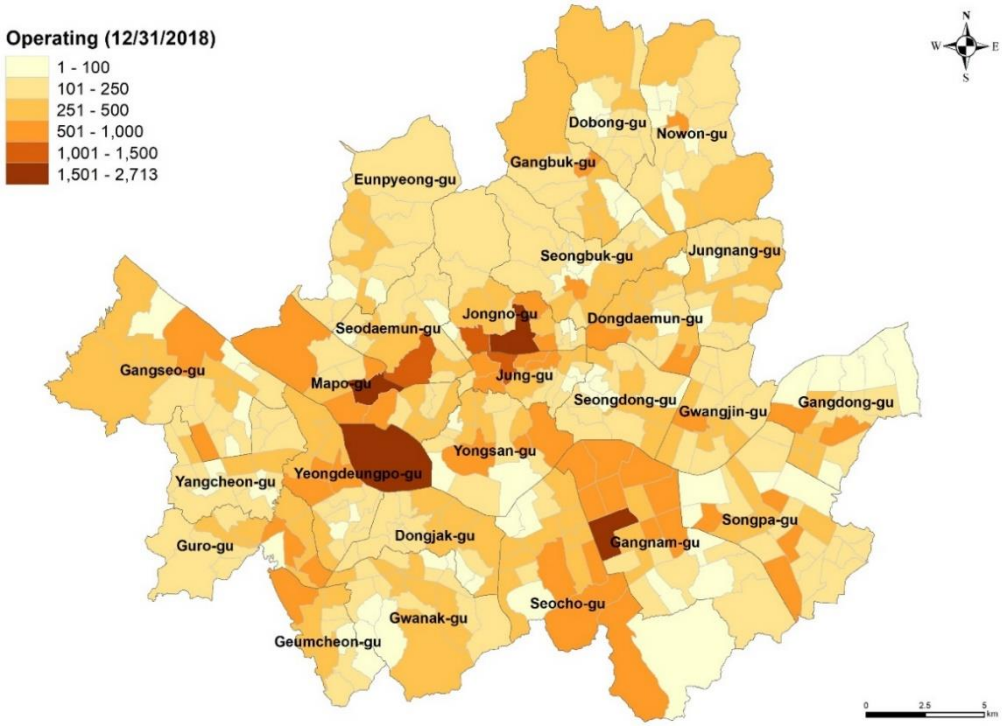


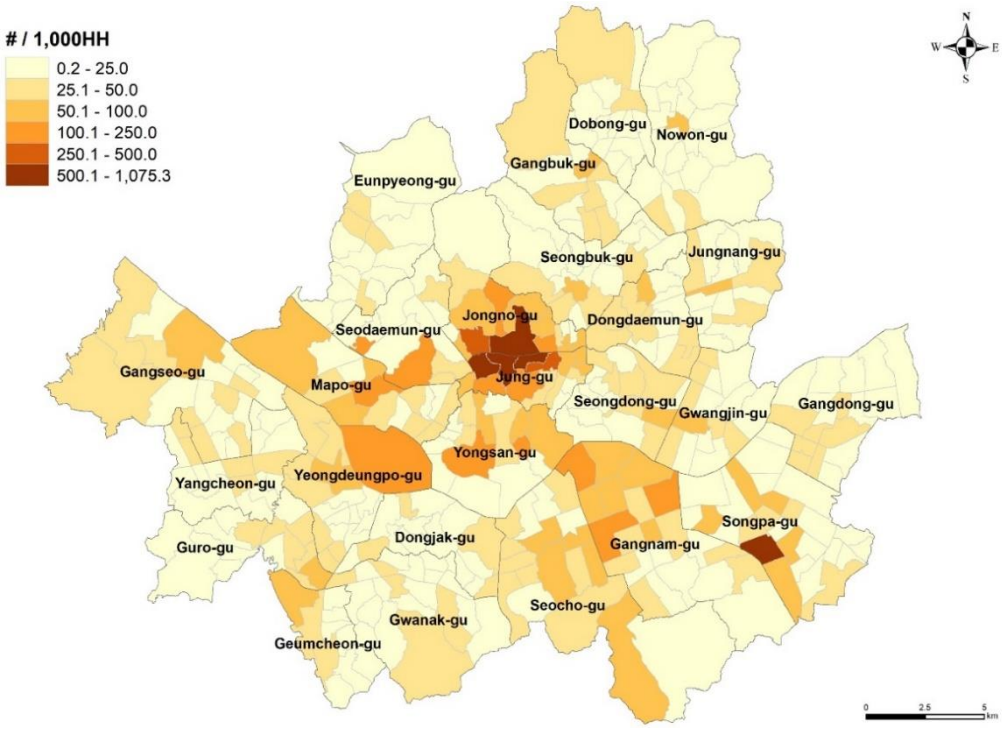
Figure 3.2 Distribution of survived years of restaurants

The number of currently operating restaurants at the end of 2018 was 120,011. That equates to 198.3 shops per km^2 and 28.28 shops per 1,000 households. Figure 3.3a illustrates the spatial distribution at the dong scale. 283.04 restaurants were operating in a unit on average, but four dongs, *Jongno-gu*, *Mapo-gu*, *Yeongdeungpo-gu*, and *Gangnam-gu*, showed significantly high numbers. With the exception of *Seogyo-dong* in *Mapo-gu*, which is a popular college campus town, the other areas are major business districts in Seoul. When the number is normalized by households, *Jongno-gu* and *Jung-gu* are stand out (Figure 3.3b). While those major commercial areas contained more start-ups than closures in 2018, most areas experienced more closures. Notably, decreases in *Seocho-gu*, which is considered a thriving area, were considerable (Figure 3.3c). Based on Figure 3.3d presenting the average

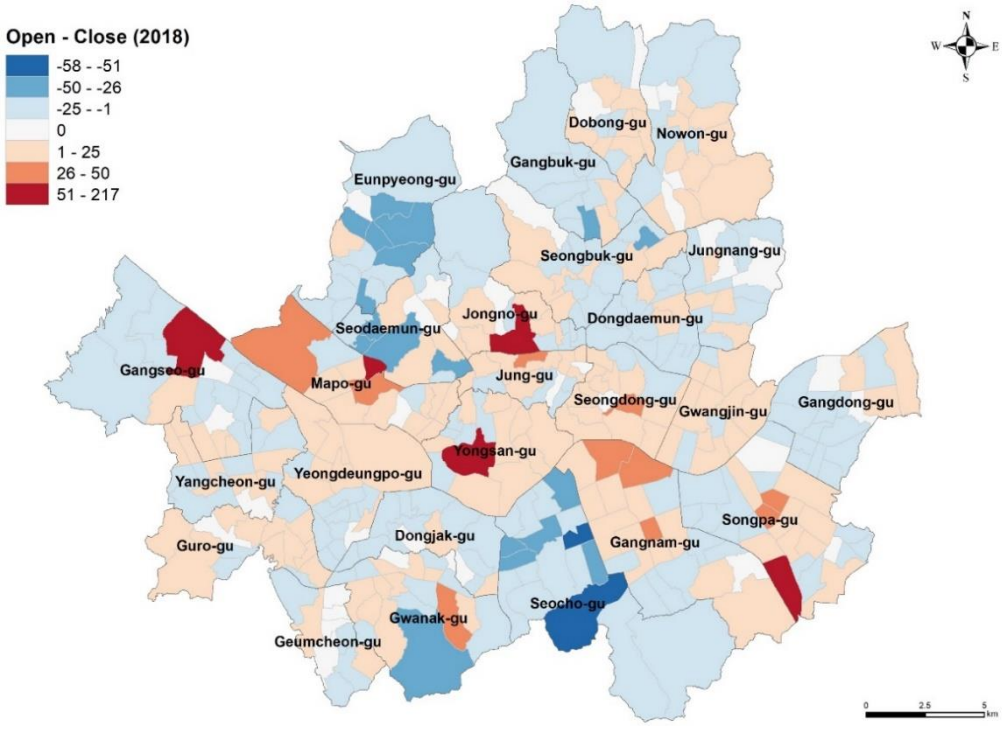
survival years for closed businesses, restaurants in the campus town of *Mapo-gu* survived less than the other major districts.



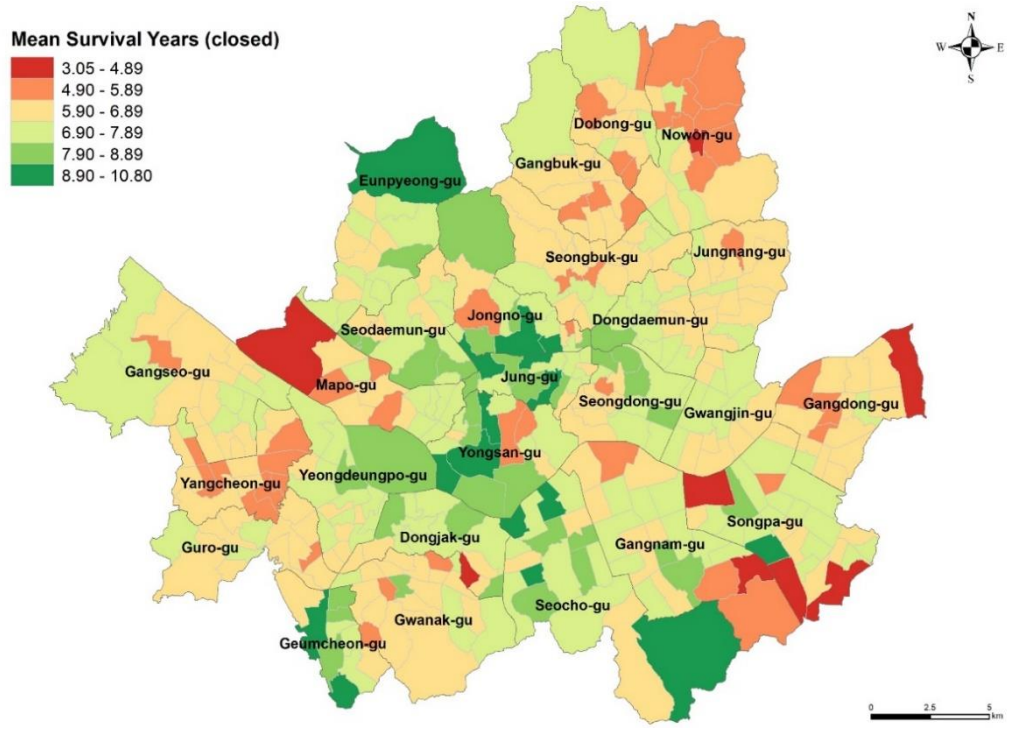
(a)


/ 1,000HH
0.2 - 25.0
25.1 - 50.0
50.1 - 100.0
100.1 - 250.0
250.1 - 500.0
500.1 - 1,075.3
N
W
E
S
Dobong-gu
Nowon-gu
Gangbuk-gu
Eunpyeong-gu
Seongbuk-gu
Jungnang-gu
Jongno-gu
Dongdaemun-gu
Seodaemun-gu
Gangseo-gu
Mapo-gu
Jung-gu
Seongdong-gu
Gwangjin-gu
Gangdong-gu
Yongsan-gu
Yangcheon-gu
Yeongdeungpo-gu
Songpa-gu
Guro-gu
Dongjak-gu
Gangnam-gu
Seocho-gu
Gwanak-gu
Geumcheon-gu
0
2.5
5
km

(b)


Open - Close (2018)
-58 - -51
-50 - -26
-25 - -1
0
1 - 25
26 - 50
51 - 217
N
W
E
S
Dobong-gu
Nowon-gu
Gangbuk-gu
Eunpyeong-gu
Seongbuk-gu
Jungnang-gu
Jongno-gu
Dongdaemun-gu
Seodaemun-gu
Gangseo-gu
Mapo-gu
Jung-gu
Seongdong-gu
Gwangjin-gu
Gangdong-gu
Yongsan-gu
Yangcheon-gu
Yeongdeungpo-gu
Songpa-gu
Guro-gu
Dongjak-gu
Gangnam-gu
Seocho-gu
Gwanak-gu
Geumcheon-gu
0
2.5
5
km

(c)



(d)

Figure 3.3 Spatial distribution of restaurant businesses: (a) the number of operating restaurants at the end of 2018; (b) the number of restaurants per 1,000 households; (c) the number of net openings in 2018; (d) the mean survival years of closed restaurants.

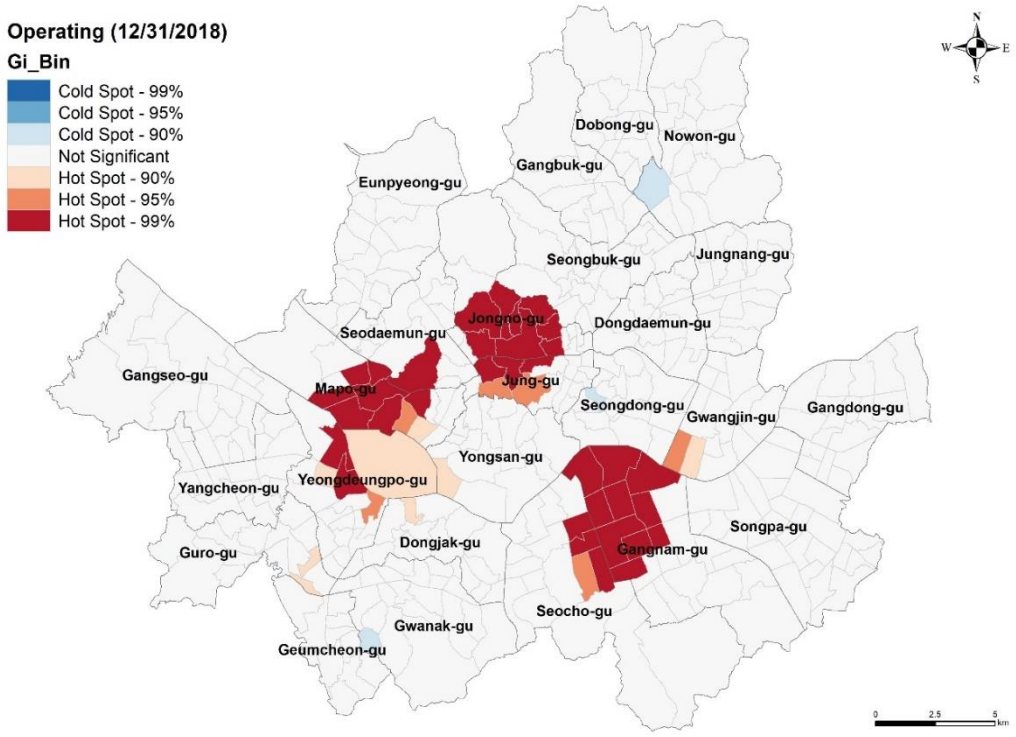
3.5. Results

Moving beyond the descriptive details offered in the previous section regarding restaurant openings and closings, exploratory analysis is not offered based on the application of the previously reviewed methods.

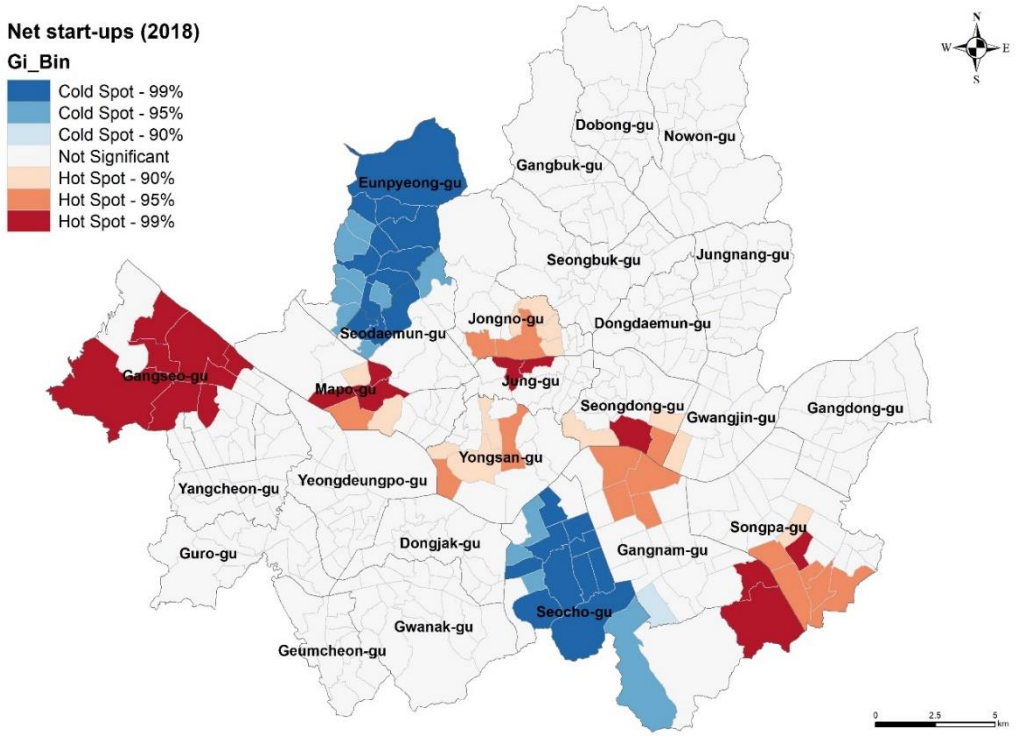
3.5.1. Spatial Clusters of Restaurants

To assess spatial patterns observed in Figure 3.3, G_i^* local spatial autocorrelation statistics were derived. Figure 3.4 displays detected hot and cold spots of current businesses and net start-ups in 2018. The first map shows three large hot spots with a significantly higher number of operating businesses than their neighbors (Figure 3.4a). These areas

include three major cores in Seoul: *Jongno-Jung-gu* (central business district), *Yeongdeunpo-Mapo-Seodaemun-gu* (Yeouido business district), and Gangnam-Seocho-gu (Gangnam business district) (Shin and Shin 2009). In *Jongno-Jung-gu*, as the traditional central business district, there are many restaurants for both tourists and office workers. *Gangnam-Seocho-gu* is also a well-known district and a socioeconomic center of Seoul which supports many popular shopping districts. On the other hand, *Mapo-gu* has mixed characteristics. The area has a business-oriented section, but it is well-known as a campus town with four prestigious universities. The general location patterns of restaurants have not much changed since 2000.



(a)



(b)

Figure 3.4 Spatial hot and cold spots of restaurant businesses: (a) operating restaurants at the end of 2018; (b) the number of net openings in 2018.

These clusters are still viable for starting new establishments (Figure 3.4b). The central business district area shows hot spots for the number of net start-ups in 2018, which indicates the new business outnumbered closed shops. Although the year is considered as an economic downturn period, significantly large numbers of businesses started in the three core areas. However, the area of hot spots is much smaller than the area of currently operating restaurants' hot spots (See Figure 3.4a). In the region, the most significant hot spots shrink from 11 to 2 dongs. Similarly, the hot spots in *Yeouido Business District* diminished to three dongs. *Seocho-gu* resulted in significant cold spots indicating that more restaurants closed than opened. Moreover, *Seongdong-gu*, rather than *Gangnam-gu*, is detected as a new hot spot for new restaurants. Additionally, the southern part of *Gangnam-*

gu, which does not contain hot spots of current businesses (See Figure 3.4a), became a hot spot because new towns had been developed (similarly to *Gangseo-gu*).

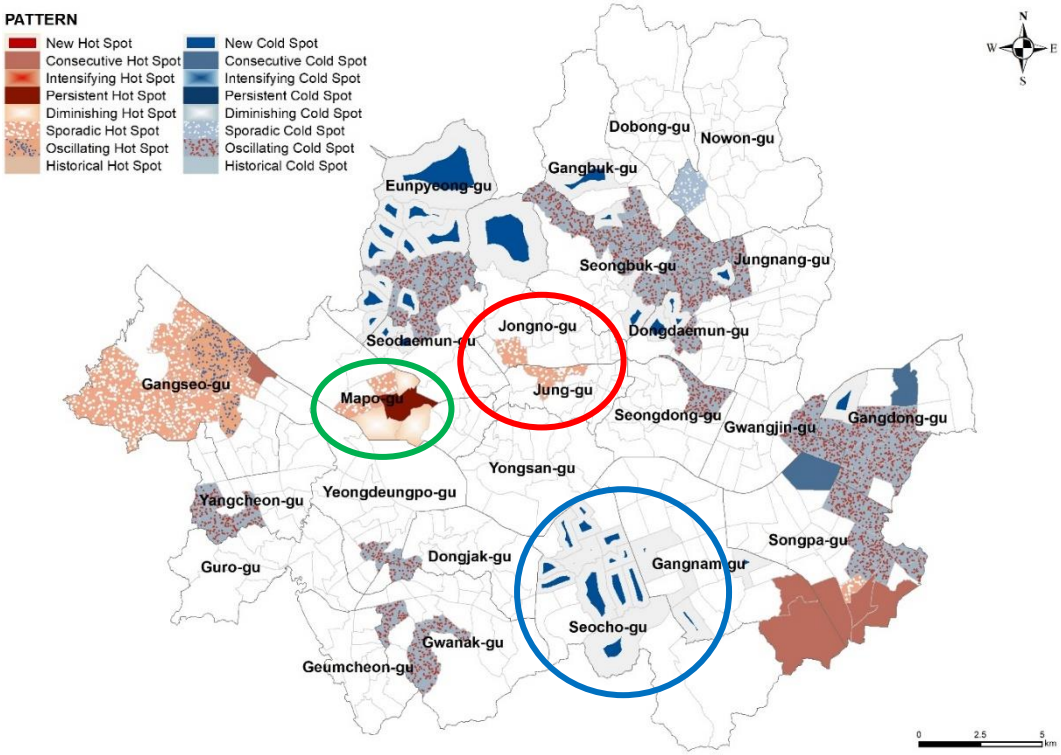
3.5.2. Temporal dynamics of spatial clusters

Based on a space-time cube with bins representing net openings in a dong by year, the results of emerging hot spot illustrate temporally categorized current clusters. Compared to Figure 3.4b, Figure 3.5a shows a large area of cold spots. The north-eastern area, including *Gangbuk, Seongbuk, Jungnang, Dongdaemun-gu*, have oscillating cold spots. This indicates areas that have a history of statistically significant hot spots for less than 90% of the total time period but become a cold spot at the final time step, 2018. Another notable pattern is the large area of new cold spots in *Eunpyeong-gu* and *Seocho-gu*. New cold spots represent areas which have never been a cold spot except in the final time step, 2018. Although these areas have different socioeconomic composition, current environments in both areas are not favorable for restaurants starting new businesses.

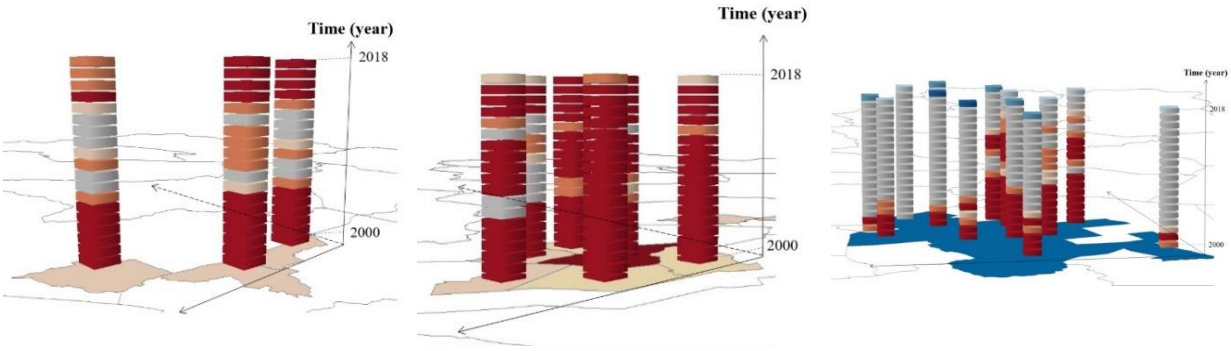
On the other hand, centric areas appear as hot spots, except Gangnam district. The central business district contains sporadic hot spots, which have never been cold spots. During most of the time period, the area has shown statistically significant hot spots especially from 2000 to 2006 and 2015 to 2018 (Figure 3.5b). Although the net opening restaurants from 2005 to 2006 were negative at the entire city level (see Figure 3.1), more restaurants opened than closed in the region. Also, *Yeouido* district has significant hot spots with persistent, diminishing, and sporadic areas. A persistent hot spot denotes that the areas have maintained the status of a hot spot for 90% of the time period. Diminishing hot spots are like persistent hot spots, but the intensity of clusters decreases. In the *Mapo-gu* area (Figure 3.5c), the two diminishing hot spots are detected in 2018 compared to the persistent

hot spots even though they have more significant clusters over time. Unlike the two centers, *Gangnam* district (Figure 3.5d) presents new cold spots referring the areas were favorable for restaurants to start at least by 2015 but it recently turned to be cold spots.

These temporal fluctuations in clusters implies that the preferable locations for new restaurants have changed even for cores of the city. While *Yeouido* district is still attractive for business investments and the central business district recovers its reputation, *Gangnam* district is experiencing a decline in popularity of its restaurants. Also, recent economic decreases are observed through many cold spots in local (or town) centers where it has never occurred before. Although some edge areas are booming in 2018 with development of new towns, the restaurant business generally has faced a downturn and thus, its concentration into city centers has intensified.



(a)



(b) (c) (d)

Figure 3.5 Temporal changes in spatial clusters of restaurant businesses: (a) the number of net openings from 2000 to 2018; (b) yearly bins of spatial cluster in Jongno and Jung-gu (area encircled in red in a); (c) yearly bins of spatial cluster in Mapo-gu (area encircled in green in a); (d) yearly bins of spatial cluster in Gangnam and Seocho-gu (area encircled in blue in a).

3.5.3. Spatiotemporal variations in survivability of restaurants

To identify spatiotemporal disparities in survivability of restaurants, we analyze survival time of restaurants with a spatiotemporal scan statistic (exponential model). We count only statistically significant clusters with greater than 0.05 p-values derived from 999 random permutations. Moreover, we define a risky cluster of restaurants when it has a significantly large number of observed closures than expected. On the other hand, a safe cluster is when the number of observed closures is significantly lower than expected. The number of expected closures is calculated under the hypothesis that survival times of all restaurants in the city follow an exponential distribution with the homogeneous mean over space and time. As a result, thirty-four clusters in Seoul over19 years are identified, with 23 clusters deemed risky. The detected clusters are ordered by λ denoting higher likelihood of being a statistically significant.

First, restaurants in the major districts survived longer than those in other areas (Figure 3.6). Based on relative survival time (RST), restaurants in cluster 1 in *Jongno-gu* last 73.8% longer than those outside of the area (RST: 1.738). In cluster 7 in *Gangnam* district and 9 in

Yeouido district, restaurants had shorter lifespans than ones in cluster 1 in the central business district; they show 26.0% and 23.3% longer survived time, respectively. In contrast, restaurants in cluster 8 ran their businesses for an average 47.3% shorter length of time than those outside of the area.

Figure 3.7a shows the ratio of observed closures compared to expectations. Generally, the three cores have higher survivability clusters than suburban areas. Most of the risky clusters are located in the north-eastern and south-western areas while the safest cluster with the lowest observation to expectation ratio (OE ratio: 0.582) is detected in the central business district (Figure 3.7b). This means that in that particular cluster, 41.8% more restaurants had survived than the expected number whereas the riskiest cluster (Figure 3.7d) located in the *Gangnam* district indicates 89.5% more restaurants failed in the cluster area. Another cluster across *Gangnam-gu* and *Seocho-gu* (Figure 3.7c) is identified as a safe cluster with relatively low OE ratio (0.797). Compared to *Gangnam*, a cluster in *Yeouido* district (Figure 3.7e) shows a high survivability with 0.814. Both indices, OE ratio and relative survival time, demonstrate that the cluster in *Gangnam* district (#8) is the riskiest area for restaurant businesses in this study area and time period.

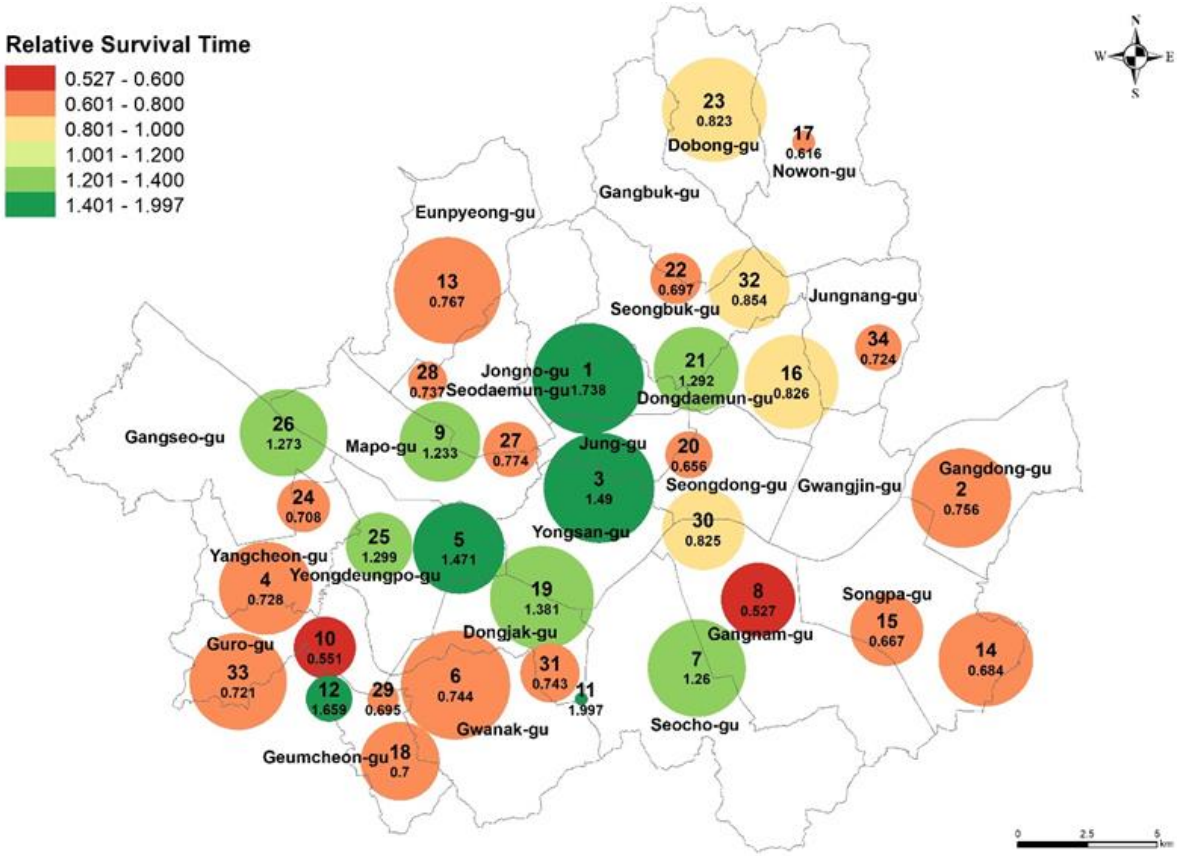
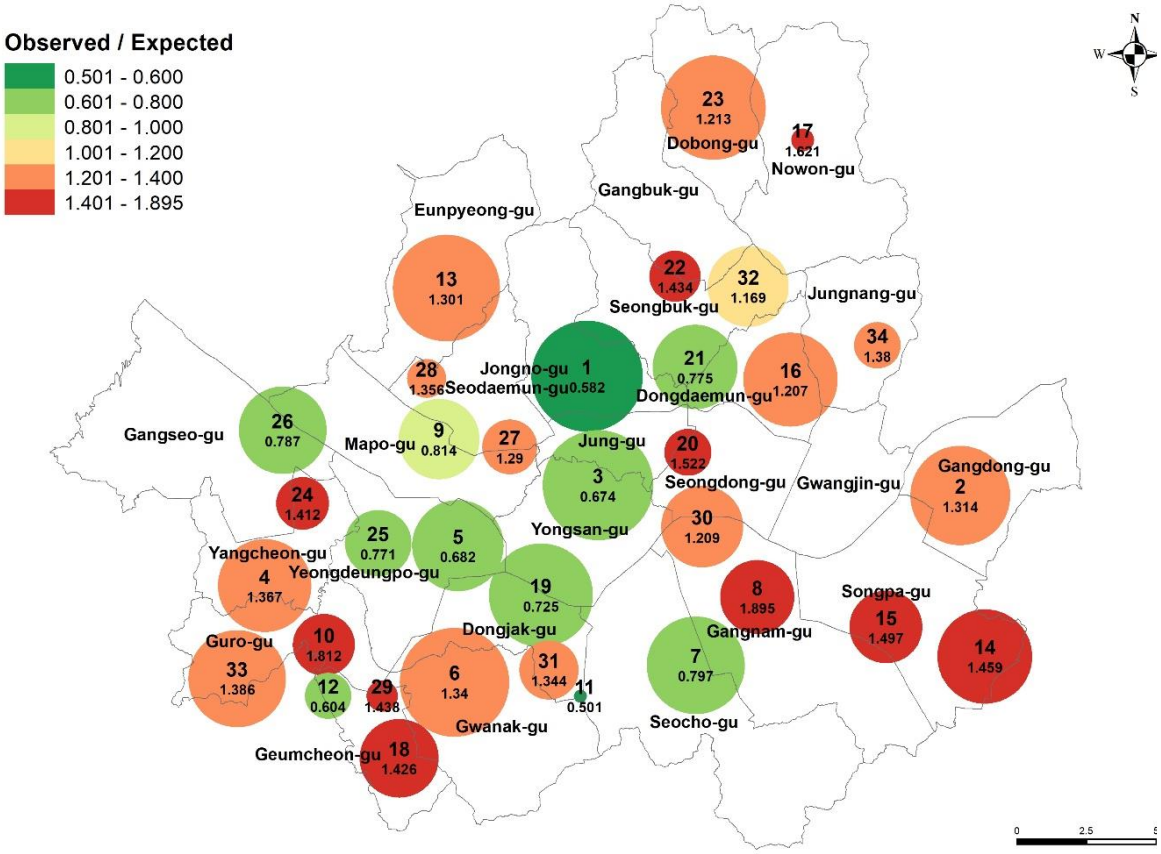
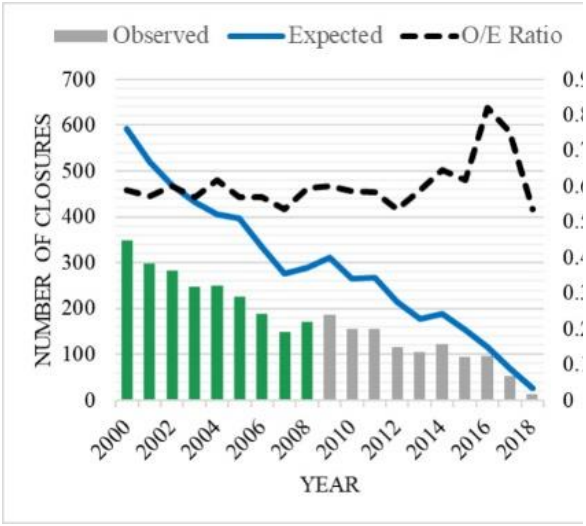


Figure 3.6 Spatial distribution of relative survival time.


Observed / Expected
0.501 - 0.600
0.601 - 0.800
0.801 - 1.000
1.001 - 1.200
1.201 - 1.400
1.401 - 1.895
N
W
S
23 1.213 Dobong-gu
17 1.621 Nowon-gu
Gangbuk-gu
Eunpyeong-gu
13 1.301
22 1.434 Seongbuk-gu
32 1.169
Jungnang-gu
34 1.38
28 1.356
Jongno-gu 1 0.582 Seodaemun-gu
21 0.775 Dongdaemun-gu
16 1.207
Gangseo-gu
26 0.787
Mapo-gu 9 0.814
27 1.29
Jung-gu
20 1.522 Seongdong-gu
Gwangjin-gu
Gangdong-gu 2 1.314
3 0.674
24 1.412
Yongsan-gu
30 1.209
25 0.771
5 0.682
Yangcheon-gu
4 1.367
Yeongdeungpo-gu
19 0.725
8 1.895
Songpa-gu
15 1.497
Dongjak-gu
Gangnam-gu
Guro-gu
10 1.812
14 1.459
33 1.386
12 0.604
29 1.438
6 1.34
31 1.344
11 0.501
7 0.797
Gwanak-gu
Seocho-gu
Geumcheon-gu 18 1.426
0
2.5

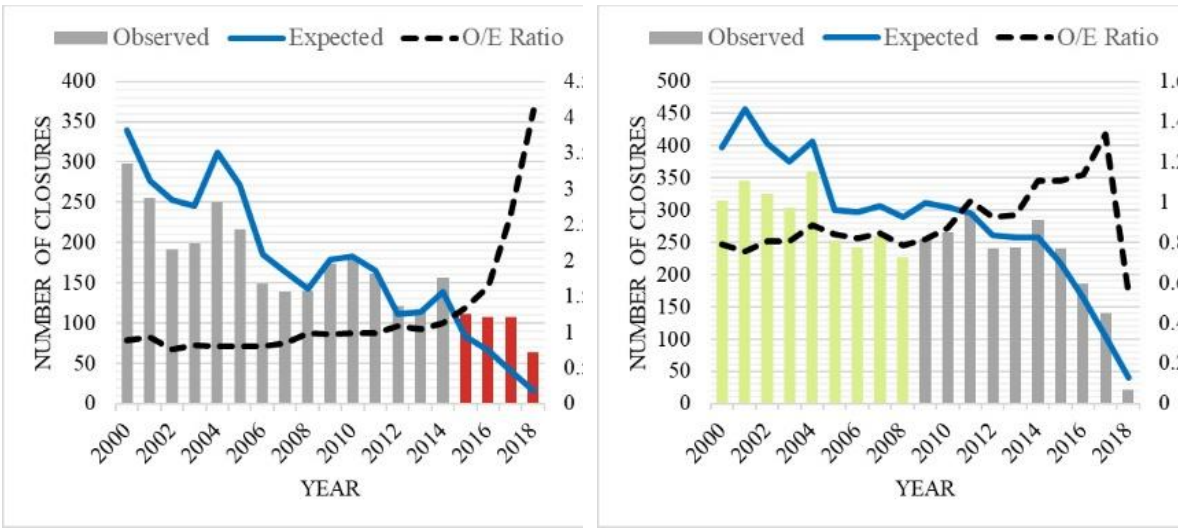
(a)


Observed
Expected
O/E Ratio
NUMBER OF CLOSURES
YEAR
0
100
200
300
400
500
600
700
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

(b)


Observed
Expected
O/E Ratio
NUMBER OF CLOSURES
YEAR
0
50
100
150
200
250
300
350
400
450
500
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

(c)



(d) (e)

Figure 3.7 Spatiotemporal distribution of observation to expectation ratio: (a) spatial distribution of risky clusters; (b) the number of observed and expected closures in cluster #1; (c) the number of observed and expected closures in cluster #7; (d) the number of observed and expected closures in cluster #8; (e) the number of observed and expected closures in cluster #9.

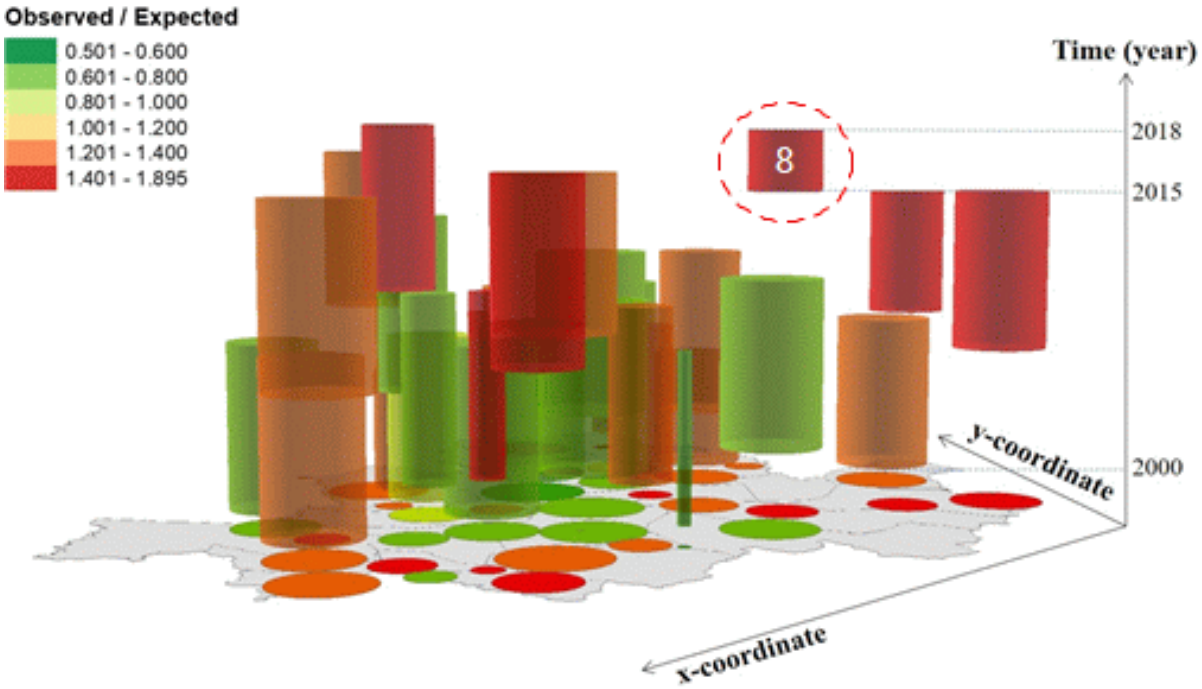


Figure 3.8 Spatiotemporal distribution of observation to expectation ratio in 3D view.

From a 3D perspective, Figure 3.8 illustrates the temporal gaps between clusters. Red cylinders represent risky clusters with high OE ratio and occur in relatively recent years compared to safe clusters. For example, the most likely and safest cluster (#1) was from 2000 to 2008. This result corresponds to the patterns of net opening clusters in Figure 4(B). Although these two clusters do not share the same time period and spatial extent, the results demonstrate that the central business district was a favorable environment for restaurants in the early 2000s. During the period, restaurants in the area could survive longer than they could in other areas and more restaurants opened rather than closed. Likewise, *Yeouido* district also experienced a longer survivability of restaurants in the early 2000s. The area still has more openings than closures (see Figure 5(C)), but the restaurant's lifespans shortened from 2014 to 2017 compared to its history. Another safe cluster across *Gangnam-gu* and *Seocho-gu* (#7) appeared from 2003 to 2011 while the riskiest cluster (#8) started in 2015. Since 2015, both areas experienced a higher number of closures than expected even though the ratio in cluster #7 was not statistically significant. In 2018, the OE ratio in cluster 8 soared to about 4.0, indicating that the number of closed restaurants is four times higher than the expected failures. Similar to the result of spatiotemporal cluster analysis on net openings, the results strongly support the posture that *Gangnam* district has recently undergone a decline in restaurant businesses.

3.6. Discussion

Within three core areas in Seoul, including the central business district, *Yeouido,* and *Gangnam*, they have been the most favorable for restaurants since 2000. These patterns are based on advantages of agglomeration. These business districts are the most profitable areas with the highest number of lucrative companies, including headquarters of global conglomerates. A large number of workers in these areas has generated a great amount of demand for restaurants. Not just commuters, but also travelers, contribute to the growth of restaurant businesses in the areas because of their unique vibes and media impacts. International travelers, in particular, have been major consumers in the central business district. However, these advantages are not fully observed in all areas through time. Reflective of this is the downturn in *Gangnam* compared to central business district and *Yeouido*, which have been revitalized and remain favorable environments. There are two potential reasons for the observed declines. Firstly, the overall regional economy has slowed due to a recent nationwide downturn. Although *Gangnam* is the most affluent area, it can be impacted by national scale economic changes. Another possible reason is that the environment for restaurant businesses in the district is no longer favorable. As a huge commercial center in Seoul, rents have been increasing in the area, but restaurants are less likely to be able to afford rising rents compared to other services, such as shopping or other leisure services.

Failures in smaller markets can be more critical than those in the core areas. In the context of Korean economic and labor structure, a considerable number of small restaurants are launched by the early-retired, who have low capital and little experience in the restaurant industry. They have few choices except opening a new restaurant in a small market with

limited capital. This likely causes saturation of the local market, leading to massive failure during economic downturns. As small businesses in a local market consume labor, including the early-retired as well as low-skilled workers, their failures have a great impact on the local and national economy. This aggravates an economic downturn. Evaluating the level of saturation based on risky cluster detection suggested in this research is helpful to manage stability of local markets by alerting government agencies to potential risk in opening new businesses in certain areas. Based on the knowledge of risky clusters for new businesses, individuals can re-consider start-ups and county-level local governments can require stricter standards to open new businesses in the risky areas.

Although this exploratory approach is noteworthy, it has a few limitations. For instance, it is challenging to explain reasons underlying spatiotemporal patterns of clusters. Although two potential causes of declines in *Gangnam* were highlighted, more formal modeling with additional covariates, including relationships with hotels, shopping centers, etc., would be important as a secondary assessment. Secondly, details about different types of restaurants could present diverse patterns of urban dynamics. More information about ownership and food types would be particularly valuable. As many studies have pointed out, type of food and/or ownership can determine a phase of development in a region (Zukin et al. 2009; Hyde 2014; Minner and Shi 2017; Ryu and Park 2019; Widaningrum et al. 2020). This would facilitate the evaluation of risky local markets. Finally, impacts of failures on local markets should be closely investigated to determine whether governmental interventions are helpful. Further research extension along these lines to address such issues would be of great interest.

3.7. Conclusion

This chapter explored the spatiotemporal dynamics of restaurant entrepreneurship in Seoul, South Korea based on the availability of public open data using three exploratory methods within a space-time framework. Spatial hot spot analysis identified core areas in Seoul that remain favorable for restaurants. The individual records of restaurants facilitated delineating a more precise extent of restaurant hot spots, generally not corresponding to traditional administrative units. Moreover, the individual records on opening and closing date allowed us to examine temporal changes in restaurant businesses. Trend analysis revealed intensifying or diminishing cluster patterns, finding that *Gangnam* district and many other areas had recently become less favorable for restaurants in contrast to other core areas. Spatiotemporal scan statistics examined risky areas, revealing that lifespans of restaurants were significantly shorter than other areas.

Based on the findings, we conclude that the general downturn in restaurant businesses in Seoul started after 2010, but Gangnam district experienced significant decreases in restaurant businesses beginning in 2015. The applied spatiotemporal exploratory approaches illustrate dynamic changes in restaurant businesses, with the results highlighting that the concentration of restaurants in popular areas has intensified in Seoul. Despite limitations of exploratory approaches, this study suggests a methodological framework for investigating spatiotemporal changes at micro scales within a city featuring a series of analyses verifying the changes from multiple perspectives. This research provides fundamental knowledge of urban dynamics by demonstrating that locational advantages are not permanent, but rather change continuously, and even dramatically, over time. This knowledge enables the private

and public sectors to make better decisions such as avoiding high-risk areas to open new businesses and imposing stricter requirements for new start-ups within riskier areas.

4. Explaining Urban Dynamics with Human Mobility through GeoAI

4.1. Introduction

Human mobility encompasses diverse human activities such as physical travels of people and objects, imaginative travels through texts and images, virtual travels transcending geographical and social distance, and communicative travels via media (Urry, 2002). Human mobility that occurred by unevenly distributed resources in cities is a key to examine urban dynamics (Shaw & Sui, 2018). For example, goods and services have to move to meet peoples' demands because the best location for businesses is not evenly distributed, but spatially and temporally limited. Meanwhile, whenever, and wherever people move, travel costs are inevitably incurred regarding either time or money. Since the travel costs limit the range of services, the location is critical to sustain one's business, and competitions for better locations determine businesses' successes (Church & Murray, 2009). Also, urban structures are spatiotemporally dynamic through the location competitions (Jin & Murray, 2021). In other words, human mobility has not only shaped urban structures but has also been determined by unevenly distributed resources in cities. Understanding the complicated relationship between spatial structure and spatial interaction has been an enduring research topic in spatial science (Lo, 1991).

Traditional location theories and models, such as central place theory and gravity models, developed many years ago still provide considerable insights on interpreting complex spatiotemporal patterns of human and urban dynamics (Jin & Murray, 2021). Along with a surge of individual-level human mobility data, artificial intelligence (AI) techniques, in particular deep neural networks, have been used to analyze a large amount of

data and predict patterns of human mobility and urban dynamics (Li et al., 2021; Hagenauer & Helbich, 2022). For a deeper understanding of complexities and non-linearity in geographic processes, GeoAI, or geospatial artificial intelligence, is suggested by combining AI, geospatial large data, and high-performance computing (Li, 2020). Rather than a simple analytic tool, GeoAI encompasses diverse issues related to the entire research process as a research agenda (Janowicz et al. 2020).

Many studies have built deep learning models to understand geospatial objectives and phenomena from remote sensing images (Dorji et al., 2019; Snyder et al., 2019); nevertheless, it is still challenging to explain the relationships between inputs and outputs in the deep learning models. As deep neural networks have complicated architectures to enhance predictivity of a model by increasing the number of hidden layers and neurons, it is difficult for human to understand the functions and meaning of each input and layer in the process of model training. Due to this lack of transparency, deep neural networks are often regarded as "black-box" models (Gilpin et al., 2018). To justify the outcome of deep neural network models and support decisions in practice, the modeling process and outcome need to be more clearly and fully explained from humans' perspectives.

Explainable AI (XAI), which aims to increase the transparency and the interpretability of deep learning models, is a relatively new research agenda and has been extensively discussed in many disciplines (Samek et al., 2018). Van der Velden et al. (2022) categorized XAI based on three criteria including model base (model-based versus post-hoc), model agnostic (model-agnostic versus model-specific) and the scope of the explanations (global versus local). Model-based explanations, as a traditional approach, require an understandable size of input variables enabling human to comprehend the entire decision-

making process in a model. Post-hoc explanations, however, are more feasible to analyze a trained model such as a neural network. Rather than enforcing model to be explained, the use of post-hoc explanations attempts to examine the model's behaviors (e.g., how the model is trained) (Murdoch et al., 2019). Therefore, post-hoc approaches are beneficial to achieving insights from complex and nonlinear interactions between inputs and outputs without making the relationships simple. On the other hand, model-agnostic approaches, as a general model evaluation method, are more generally applicable to diverse types of neural networks than mode-specific explanations because perturbing is employed to test the impacts of inputs after models are built, regardless of the model types or architectures. Inherently, model-agnostic explanation is post-hoc (van der Velden et al., 2022). Furthermore, explanations can be conducted at both the global and local levels. While global explanations provide general relationships learned by the model with the entire dataset, the local level explanations utilize a limited number of subset data for understanding a specific case.

In geographic studies, model based XAI approaches have been developed to enhance our understanding of geospatial processes with deep learning. Hagenauer and Helbich (2022) developed neural networks imbedded geographic weighted regression models to identify nonlinear relationships under geographic phenomena such as house prices through nonlinear architectures of neural networks. Yudistira et al. (2021) have focused on explainable deep learning models to understand the impacts of between socioeconomic and environmental factors on spatiotemporal disparities in COVID19. Cheng et al. (2020) suggest a more general explainable model that can handle spatiotemporal changes in human activities within a space-time cube. However, these approaches use predefined spatial structures between

spatial units in the models rather than examine how spatial structures are potentially constructed through other factors such as human mobility. It is limited to understand the interactions between human activities and spatial structures.

In this study, we identify the relationships between urban dynamics and human mobility with recurrent neural networks based on the survival analysis framework. As neural networks generate a number of hidden variables (neurons), it is possible to predict well with only a small amount of given information by capturing unmeasurable patterns from peoples' observation. To enhance explainability of the neural network model, this study proposes a local model-agnostic approach by adapting Local Interpretable Model-agnostic Explanation (LIME). It evaluates local fitness of a model, to geographic local samples to analyze spatial variation in the predictive power of the model. Moreover, we apply sensitivity analysis to the geographic local samples that explains spatial variations in importance of input variables. For this study, we analyze population flow data and businesses license data in Seoul, Korea. Population flow data is an origin-destination matrix estimating hourly de facto population at a spatial unit based on mobile tracking information. Due to privacy issues, the spatial unit of origins is aggregated to a larger scale than the unit of destination. The hourly flow data is comprised of 25 origins and 425 destinations collected from May 1st, 2018, to December 31st, 2020 (976 days). Regarding businesses license data, it is individual-level information managed by county-level local governments in Korea, including location coordinates and opening and closing dates (Jin & Murray, 2021).

4.2. Recurrent Neural Networks for Survival Analysis (RNN-Surv)

Survival analysis is a statistical approach to measure the impacts of events such as diseases or on survivorship and lifespan. As survival analysis manages two types of data,

binary and continuous data, within a model, it has been widely applied to predict probability of events in a certain period time from medical studies to social sciences including finances and economics (Parsa et al., 2011). Originally, it requires lifespan data, which is the length of an event occurring for an entity, such as a patient, but it is not fully observed in many cases due to limited resources. In that case, we treat the observation as right censored data. As we do not know the exact time that the event occurs (e.g., a death of entity), the lifespan is censored at the right end on a timeline, and it is treated as censoring time C_i. To manage the censored data, life span of an event defined at an observed time is $Y_i = min\ (T_i, C_i)$, where T_i is the time when the event happens for an observation i. To predict survivability over time, a survival function is widely used. For our problem at hand, the survival function of an observation i, S_i, is the probability that an event is not happening yet at the time, t.

$$S_i(t) = \Pr\ (T_i > t)\ , S_i(t_0) = 1 \tag{4.1}$$

Based on the survival function, survival probability of a cohort is estimated by Kaplan-Meier estimator (Kaplan et al., 1958).

$$S(t) = \mathrm{S}(\mathrm{t} - 1) * (1 - \frac{d_i}{n_i}) \tag{4.2}$$

where d_i is the number of events at time t and n_i is the number of observations alive just before the time t. As a time series model, machine learning approaches to the survival analysis has been developed with diverse convergence models such as random survival forest and dependent logistic regression and neural networks (Ishwaran et al., 2008; Yu et al., 2011). Deep learning techniques such as convolutional networks are also developed (Zhu et al., 2016), but recurrent neural networks (RNN) are more eligible to manage the sequential nature of the problem in survival analysis (Giunchigilia et al., 2018). The benefit of RNN in survival analysis is to identify the time-variant effects of given covariates. A

fully recurrent neural network generates a hidden layer that recurrently influence the status of the next steps (Williams et al., 1986). With this recurrent structure, a network is able to model cumulatively changing and threshold changing events.

In this study, a recurrent neural network structure is employed for handling hourly flow populations in 31 months based on an assumption that populations visiting an area are potential customers of restaurants and which may affect on the survivability of restaurants. However, the changes do not impact on restaurants' entrepreneurship immediately, For example, it rarely happens in the real world that a restaurant closes when it faced decreases in the number of customers in the last month. Rather, cumulative effects are more critical to restaurant businesses. This assumption fits well to the basic idea of the recurrent neural network structure.

For the model evaluation, root mean square errors (RMSE), mean absolute errors (MAE), and coincidence index are used. Concordance index or C-index indicates the Area Under the ROC (receiver operating characteristic) Curve (or AUC) of censored data. It measures a model's discrimination power whether the model correctly provide a reliable order of the survival times based on the individual risk scores. It can be computed with the following formula (Uno et al., 2011):

$$\mathrm{C} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \qquad 4.3$$

where η_i is the risk score of a unit, i.

4.3. Explainability of Neural Networks

4.3.1. Sensitivity Analysis (SA)

Sensitivity analysis is used to measure the sensitivity of an input on the output of the model. This can be used to calculate the variations in the output by minimizing or vanishing certain gradients (Patil et al., 2019).

$$R_d = (\frac{\partial f}{\partial w_d}(x))^2 \quad 4.4$$

Relevance score, R_d, is a variance of the output result, ∂f, when a feature or input variable changes. The score provides information about the more sensitive parameters in a network. A large score means that the performance of output highly changes when an input, x_d, vanishes.

4.3.2. Geographic Local Interpretable Model-Agnostic Explanations (GLIME)

Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm which provides an explanation for predictions of a machine learning model, *f(x)*, through approximating it with a local model, *g(x)*, at a data point *x* and local samples around *x* (Ribeiro et al., 2016).

$$\xi(x) = argmin\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad 4.5$$

where ξ is the lime explanation, $\mathcal{L}$ is the fidelity function, π_x is the proximity measure defining locally around a data point *x*, and Ω is the complexity measure (Patil, et al., 2019). To ensure interpretability and local fidelity, minimizing $\mathcal{L}(f, g, \pi_x)$ is required.

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \quad 4.6$$

where *z* is a perturbed data point in original data space, and $\pi_x(z)$ are the local weights on data points *z* around data point *x*. In other words, LIME is an approach of explaining a point through a localized model minimizing differences to the global model with a certain number of data points neighboring the point to be explained.

In the LIME approach, the neighboring points are in not a physical space, but a conceptual variable space. However, in geographic analysis, spatial characteristics such as spatial dependence and heterogeneous do critically determine results. Therefore, defining the optimal range of neighborhoods and spatial weights has been tackled in geographic studies (Hagenauer & Helbich, 2022). As geographic events vary by space, a localized approach enhances interpretability of a neural network model in terms of local structures. When a specific geographic range is fixed, Equation 4.6 turns to Geographically Localized Interpretable Model-Agnostic Explanations (GLIME).

$$\mathcal{L}_{\pi_x}(f, g) = (f(z) - g(z'))^2 \quad 4.7$$

GLIME enhances spatial interpretability of neural network models by measuring local fidelities based on comparison to the global model at well-known geographic units such as states and counties. However, it does not explain locally various impacts of input variables on the local samples. To evaluate the local impacts of input variables, we calculate local relevance scores by combining sensitive analysis with the GLIME as follows:

$$R_{\pi_x,d} = \sqrt{\left(\frac{\partial f}{\partial w_d}(z) - \frac{\partial g}{\partial w_d}(z)\right)^2} \quad 4.8$$

The localized relevance score informs more critical variables in the local areas. A higher relevance score of an input variable *A* in an area reveals that the input variable *A* has greater impacts on the area than others.

4.4. Data

For this study, two types of public open data in Seoul, Korea, is used: population flow data and business licenses data. First, population flow data is an hourly estimated de-facto population data based on the largest telecommunication company's user information such as addresses and call detailed record. The original dataset is built on 50 m by 50 m grids with hourly population by ages and sexes, but due to privacy issues, it is only accessible with aggregated spatial units. Origin information is based on users' addresses at the county-level local autonomy unit, named *gu*, while the population at a destination is estimated through Call Detailed Records (CDR) at smaller administrative unit, named *dong*. As Seoul has 424 dongs and 25 gus, the given origin-destination matrix is the shape of 25 x 424.The temporal scale of the data is a monthly average hourly population collected from May 1st, 2018, to December 31st, 2020 (976 days). For example, if there are 100 people in a dong from a gu at noon in May 2018, 100 indicates the average of population between 12:00:00 pm and 12:59:59 pm from May 1st to 31st. Because hourly trends are too narrow to understand dynamics in more than two years changes, additional temporal aggregation is conducted. Hourly population is summed up to 4-time periods; early morning (0 to 5 am), morning (6 am to 11 am), afternoon (12 pm to 5 pm), night (6 pm to 11 pm). With these preprocesses, the shape of temporal origin-destination matrix is 25 (origin) x 424 (destination) x 4 (time periods) x 31 (months).

The second dataset involves the business license data in Seoul, Korea. As local governments (gu) in South Korea have the authority to approve new businesses, all businesses are required to report their closure to their local government (Jin & Murray, 2021). South Korea recently made this business data available, detailing business types, location, starting date, and closing date of 191 types of business, such as groceries, residential services, and restaurants. With a focus on restaurants, among more than 1.7 million records, about 150,000 restaurants are selected to match the time period of the input dataset by removing restaurants closed before May 2018. Based on opening and closing dates, each restaurant's life span is calculated in months, and the life span is going to be the predictive output value. If a restaurant is still operating at the end of 2020, the life span is calculated as the time between its start date and the end of 2020.

As the dataset includes individual characteristics such as location information and types of food they sell, they are used for other set of input variables. The location information is about coordinates and an administrative unit, *gu*, belonged to. In particular, as gu is nominal data, it is transformed to one-hot encode with 25 columns of binary variables. Similarly, types of food are encoded as the same way. As restaurants are categorized by 13 food types, restaurant's types, representing one's own merits, are encoded 13 binary variables (Table 4.1).

Table 4.1 Definition of variables.

	Variables	Dimension
Y (output)	Length of life span of restaurants in months	1
X (input)	Population flow	25 (origin) * 4 (time period)
	Location	2 (coordinates) + 25 (gus)
	Types of restaurants	13 (types)*

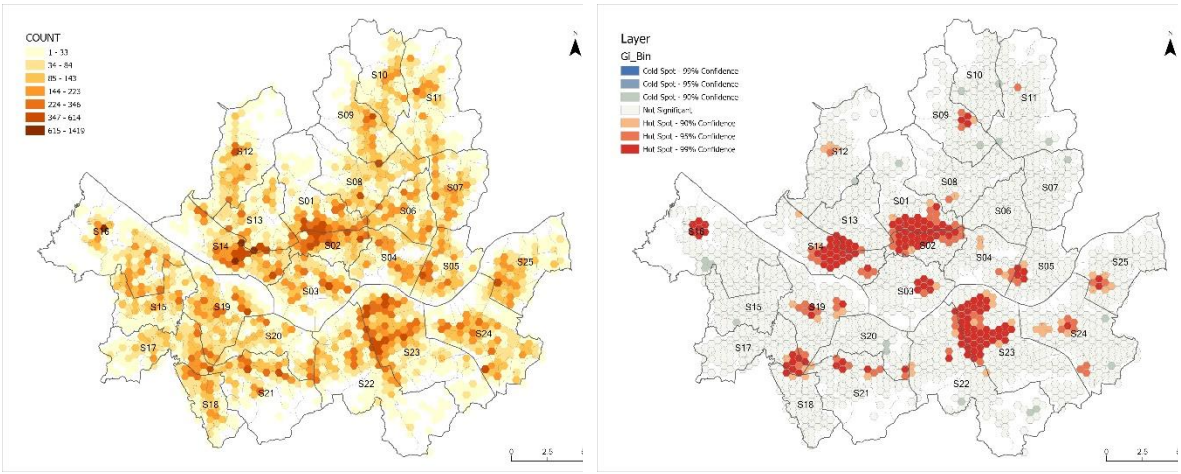
* 13 types are bars, buffet, café, chickens, Chinese, ethnic, family, fast food, Japanese, Korean, snacks, Western and others.

4.5. Results

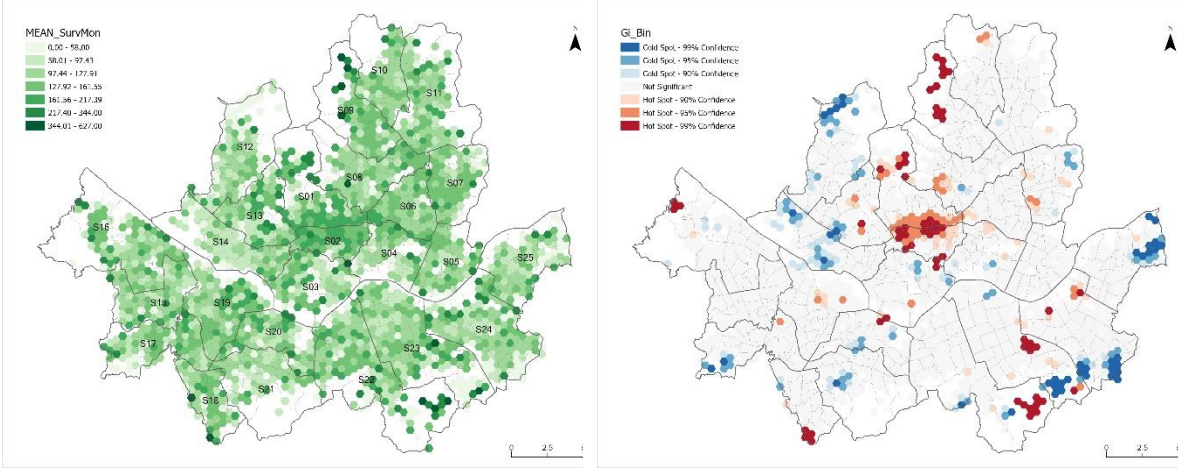
4.5.1. Data description

Figure 4.1(a) shows the number of restaurants in the study period between May 2018 and December 2020. Within a 500 m hexagon, the largest number of restaurants is in S14, a popular place for young adults with four major universities. Based on Getis-Ord Gi*, three major hot spots are detected: across S1 and S2 (central business district: CBD), S14 (Hongdae district), and S23 (Gangnam business district: GBD).

Regarding mean life spans of restaurants, these three major districts show different patterns. As the original central business district, S1 and S2 areas has a larger number of long-standing restaurants, the S14 area shows relatively short life spans as a trendy place. Moreover, local hot spots around suburb areas in Figure 4.1(b) also show relatively short life spans. It indicates that business environments for restaurants are less friendly to new restaurants in minor districts.



(a) (b)

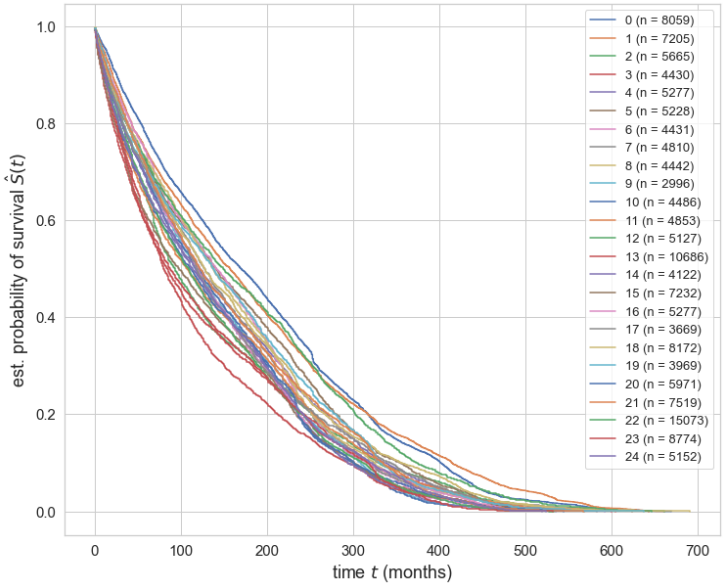


(c) (d)

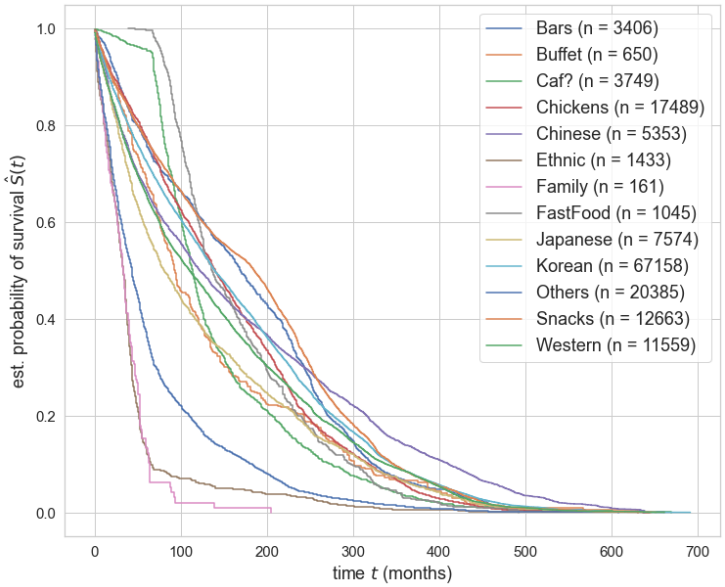
Figure 4.1. Spatial distribution of restaurants in Seoul, Korea: (a) number of restaurants in a 500m hexagons; (b) Hot-cold spots of number of restaurants; (c) mean life spans of restaurants in a 500m hexagons (in months); (d) hot-cold spots of life spans of restaurants.

Survival probabilities through the Kaplan-Meier estimator varies by locations and types of food. Figure 4.2(a) presents locally different survival probability. Similar to Figure 4.1(a), the survival probability of restaurants in S14 area (red line) more steeply decreases than S1 area (blue line). When they are compared at the point of 0.5 survival probability, restaurants in S14 survive 50% within 80 months, while more than half restaurants in S1 are able to survive over 150 months.

Survival probability by food types shows more dynamic patterns than locational variances (Figure 4.2b). It is because the number of samples for each type is unbalanced, and it reflects that the trend in dining quickly changes. For example, family restaurants, such as an Outback Steak House, boomed in 2000's, but they are now facing down turns due to other trendy foods such as original western food or ethnic food. Ethnic food also decreases steeply indicating short median life spans, but this pattern comes from a recent boom in ethnic food.


0 (n = 8059)
1 (n = 7205)
2 (n = 5665)
3 (n = 4430)
4 (n = 5277)
5 (n = 5228)
6 (n = 4431)
7 (n = 4810)
8 (n = 4442)
9 (n = 2996)
10 (n = 4486)
11 (n = 4853)
12 (n = 5127)
13 (n = 10686)
14 (n = 4122)
15 (n = 7232)
16 (n = 5277)
17 (n = 3669)
18 (n = 8172)
19 (n = 3969)
20 (n = 5971)
21 (n = 7519)
22 (n = 15073)
23 (n = 8774)
24 (n = 5152)
est. probability of survival $\hat{S}(t)$
time t (months)

(a)


Bars (n = 3406)
Buffet (n = 650)
Caf? (n = 3749)
Chickens (n = 17489)
Chinese (n = 5353)
Ethnic (n = 1433)
Family (n = 161)
FastFood (n = 1045)
Japanese (n = 7574)
Korean (n = 67158)
Others (n = 20385)
Snacks (n = 12663)
Western (n = 11559)
est. probability of survival $\hat{S}(t)$
time t (months)

(b)

Figure 4.2 Kaplan-Meier estimation of survivability of restaurants: (a) by locations (25 gus); (b) by types of food (13 types).

Table 4.2 Architecture of recurrent neural network (RNN) model

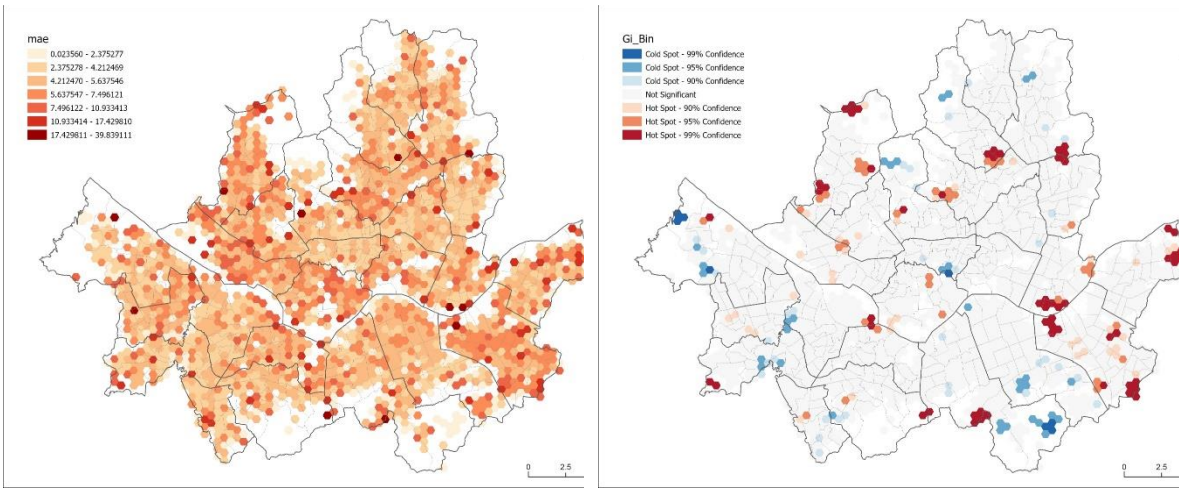
Layer	Output shape	# of parameters	Connected to
Input 1	None x 31 x 100	0	-
Masking	None x 31 x 100	0	Input
RNN	None x 31 x 1	101	Masking
Flatten	None x 31	0	RNN
Input 2	None x 40	0	-
Concatenate	None x 71	0	Flatten & Input 2
Dense	None x 1	71	Concatenate
Total parameters: 172 Trainable parameters: 172 Non-trainable parameters: 0			

4.5.2. RNN-Surv model for survivability of restaurants

For the recurrent neural network layer (RNN in Table 4.3), 3,100 (31 months x 25 origins x 4 time periods) flow population variables are used as inputs. The recurrent layer generates a single parameter per each time step (31 months), indicating monthly features of human mobility. By combining these monthly features and restaurants' non-temporal features including locations and characteristics, the second layer (Dense in Table 4.2) predicts life span of each restaurant in months through ReLU activation function because it is activated only if input values overcome a threshold. It is similar to the process of making restaurant entrepreneurship decisions. Although a restaurant is facing loss at a single month, it may not decide to close the restaurant right after because it could be an exceptional month, or it may have some money to operate for a while. However, when the sum of losses exceeds expected revenues, it will consider the business closed. In this model, we use a small number of hidden features to make the model interpretable, but its insufficient information is backed up by a number of iterations (the number of epochs is 30) with highly accurate predictive results.

In terms of mean absolute error (MAE), the model shows 4.011 in general, indicating only 4- months differences between predicted and actual life spans of restaurants. The Root Mean Squared Error (RMSE) value is slightly higher than MAE, 7.32, because there are some outliers who have survived more than 10 years. However, the RMSE value also supports the model's predictive power. Regarding coincidence index, it is 0.9768, which demonstrates that 97.68% of pairs randomly chosen from the data are correctly predicted in terms of their orders. For example, Restaurant A survives 5 months while Restaurant B survives 10 months. In this case, if the model predicts A as 10 months and B as 5 months, MAE and RMSE are the same that A as 10 and B as 15. However, the second prediction is more reliable because Restaurant B survives longer than Restaurant A. With low errors and high proportion of matching orders, the model's results prove that limited information works well in predictions via neural networks.

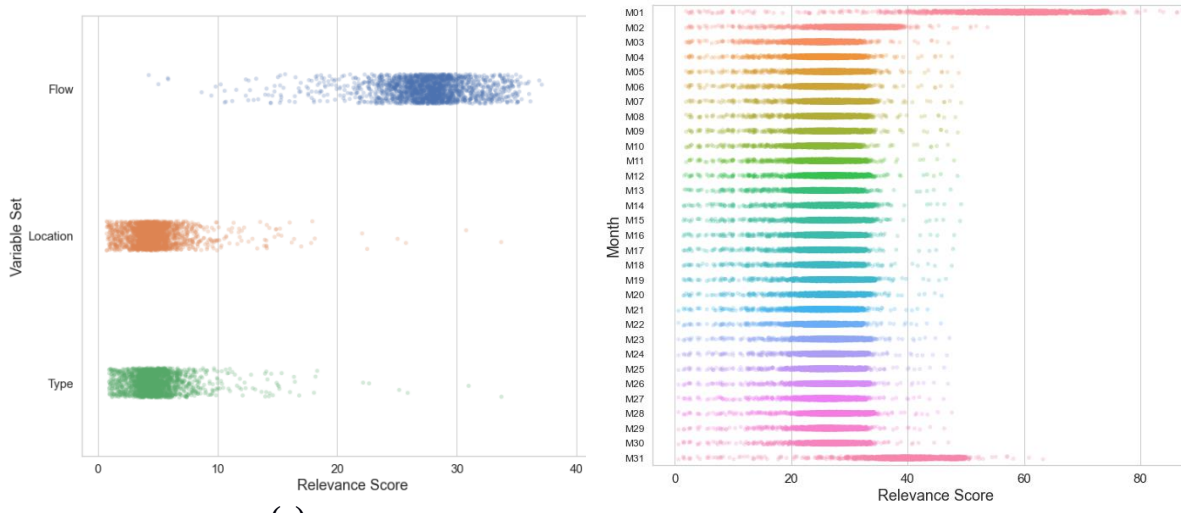
To visualize spatial patterns of errors, we aggregate absolute errors into 500 m hexagons (Figure 4.3), but the model does not show severe spatial dependence, except some small outliers at the edge of the cities. For instance, S1, S2, S14 and S23 has smaller MAE values than the global mean value. Because the majority of restaurants is located in the three major districts (see Figure 4.1), the model focuses on minimizing errors in the areas.



(a) (b)

Figure 4.3 Spatial distributions of mean absolute errors: (a) local mean absolute error; (b) hot spots of local mean absolute errors.

To understand the impact of various variable sets, sensitivity analysis is conducted by sets of variables rather than each variable because of high correlations between variables. By vanishing variable sets, relevance scores are computed (Figure 4.4). Each dot on the graph represents a relevance score at a data point. In Figure 4.4a, the set of population flow variables has more sensitive impacts on the model than the other two sets of variables. Generally, without the population flow variables, the model works poorly except under some extraordinary cases. Among population flow variables, the first and last month information critically affects the model's performance (Figure 4.4b). It mainly comes from a high correlation between human mobility patterns. As the data based on call details records and daily movements, the monthly patterns are not much different to each other. Therefore, when the first-month mobility patterns are used as inputs, it significantly improves the model's performance, whereas others do not have the similar impacts because it is not much different to the first month patterns. However, the last month pattern is also important. Even though the daily patterns are stabilized in a short time period (31 months), some changes in daily mobility occurs in the period, and they enhance predictive powers of the model.



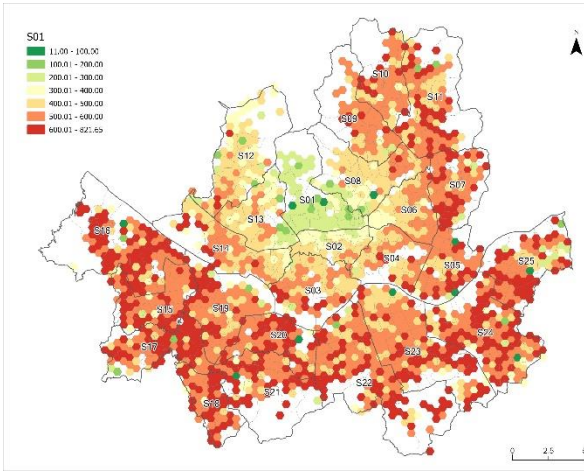
(a) (b)

Figure 4.4 Relevance scores of input variables: (a) by set of variables; (b) by flow inputs

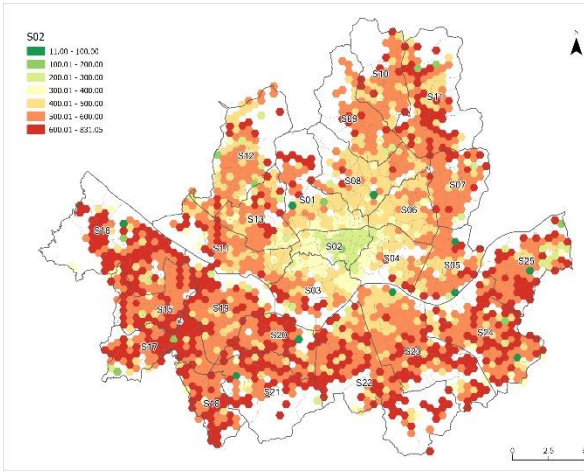
4.5.3. Geographic Local Interpretable Model-Agnostic Explanations (GLIME)

To identify local impacts of human mobility patterns on survivability of restaurants in Seoul, Korea, Geographic Local Interpretable Model-Agnostic Explanations (GLIME) is applied to the inputs of population flow information. By vanishing all other flows except a target area, the local models are computed. For example, to evaluate the influence of population from S1 area, we input only the population from S1 to the model and calculate local MAE. Therefore, each map in Figure 4.5 shows local MAEs with input of population flows from S1 to S25. More greenish areas have smaller MAEs, indicating the local model works well. In most areas, inner mobility is significantly important than others. Population flows from the S1 (central business district) make local models in S1 and neighboring areas including S2, S4, S12, and S13 predict well, but does not significantly improve local models in other areas. On the other hand, population flows from S23 (Gangnam business district) have broader impacts than those from S1. It clearly explains the survivability of restaurants

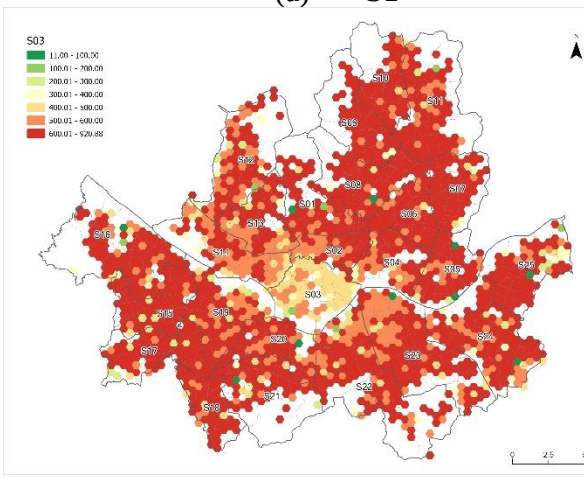
in the area (S23) with small gaps between limited local models and the full global model in terms of MAE. With high purchasing power, populations from S23, S24, and S25 have global and significant impacts on the survivability of restaurants in Seoul.



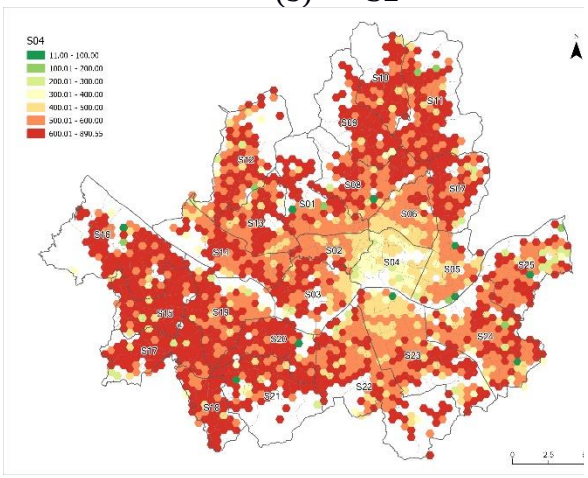
(a) S1



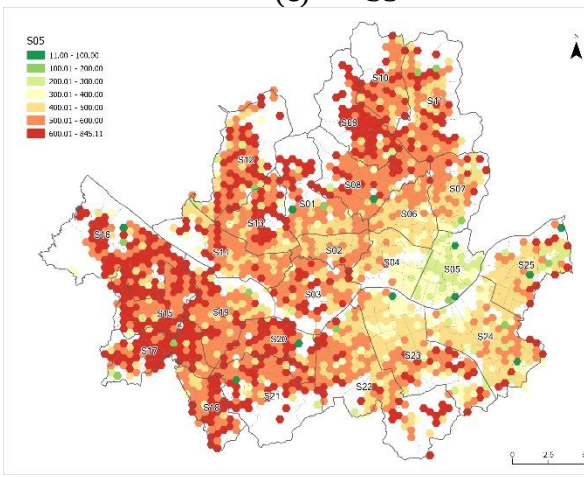
(b) S2

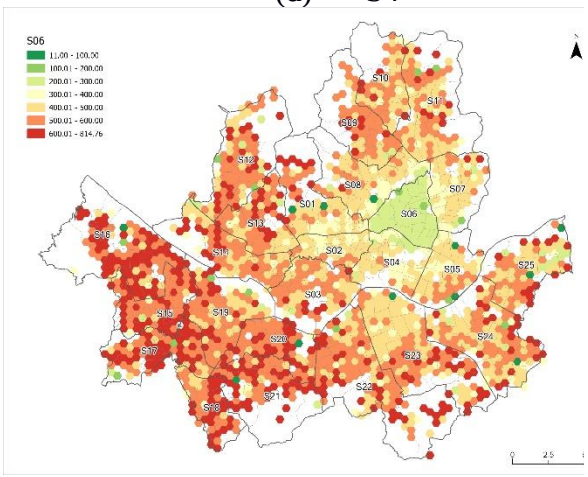


(c) S3

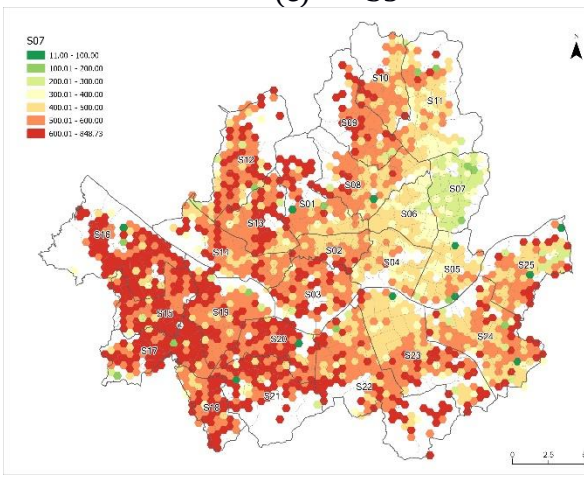


(d) S4

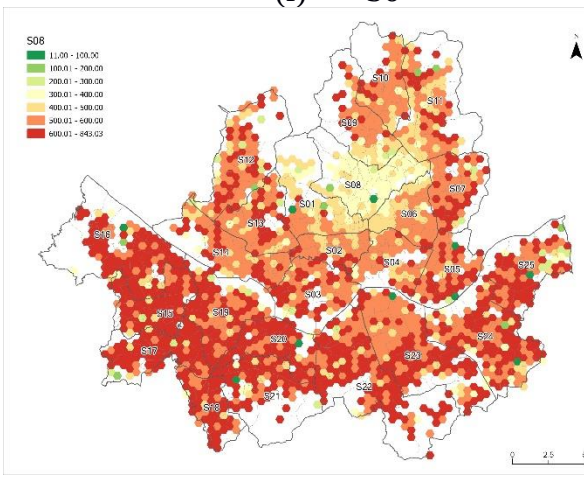




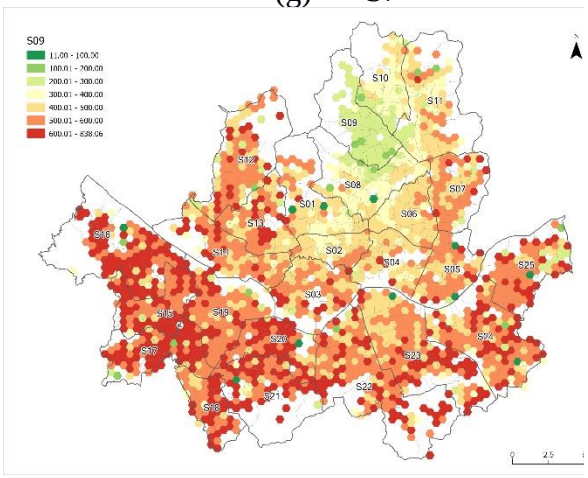
(e) S5



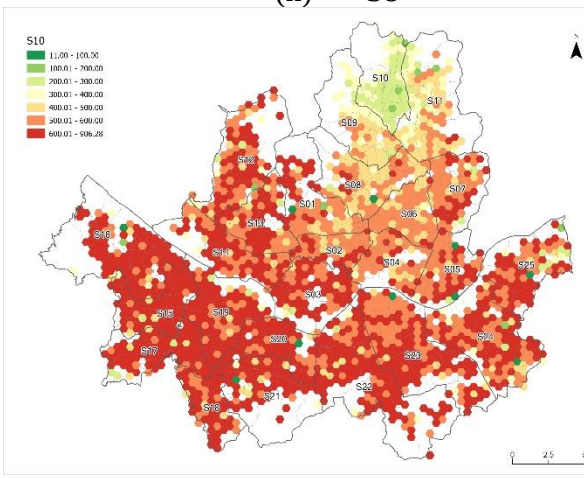
(f) S6



(g) S7

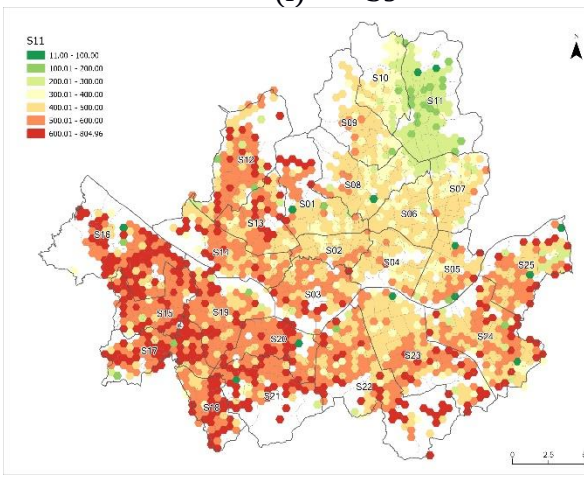


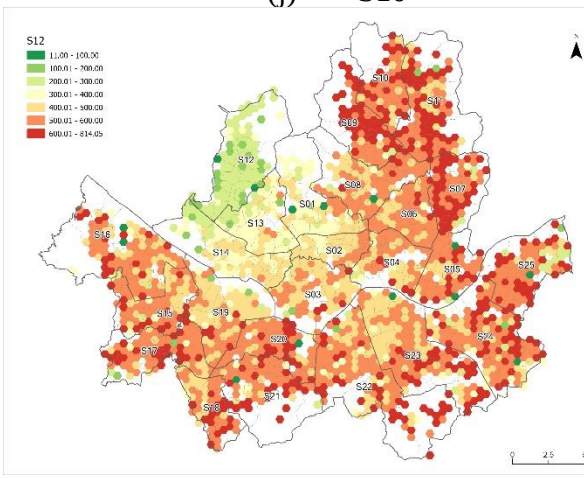
(h) S8



(i) S9

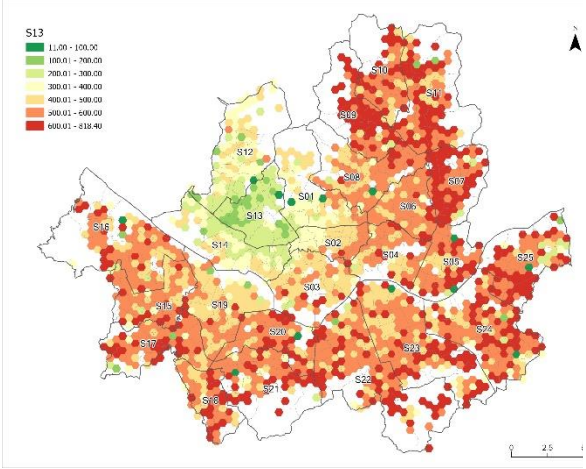
(j) S10



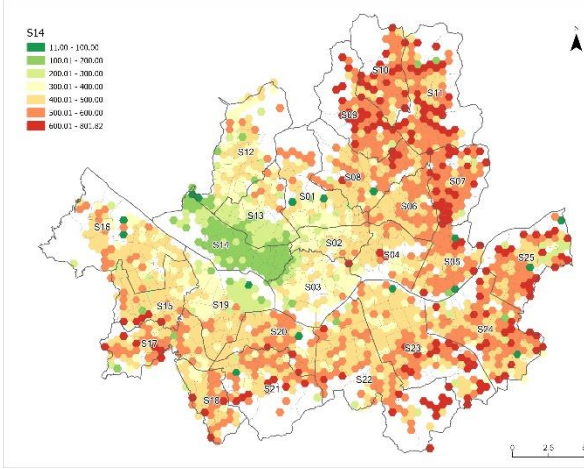


(k) S11

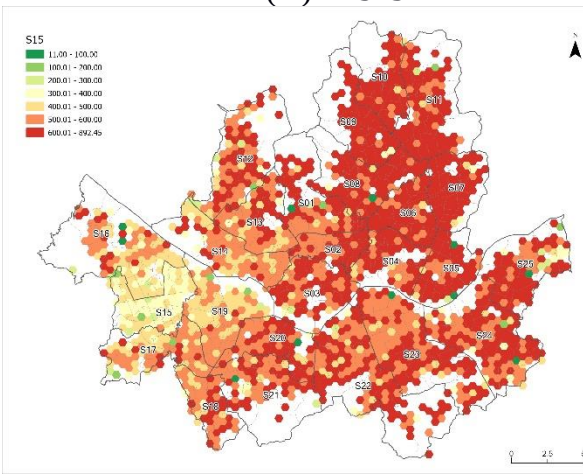
(l) S12



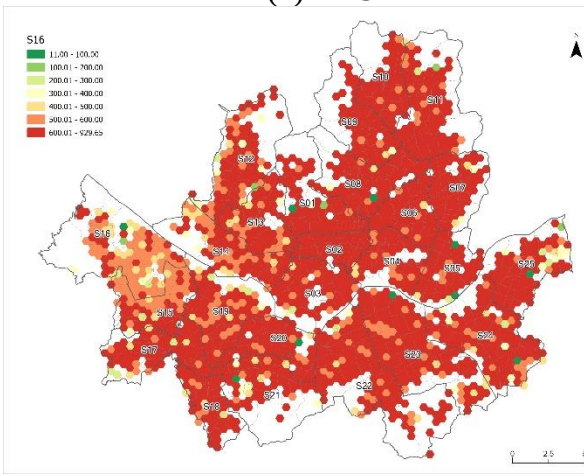
(m) S13



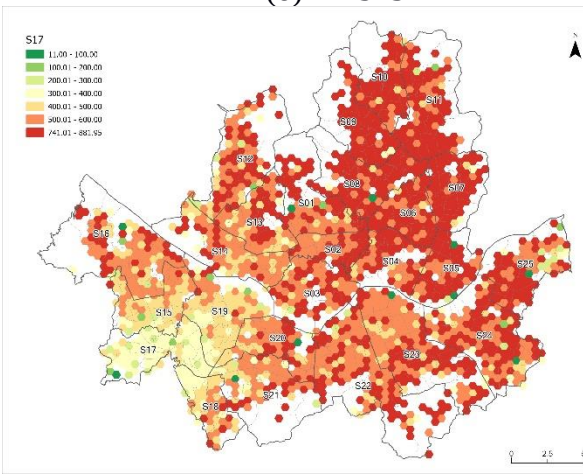
(n) S14



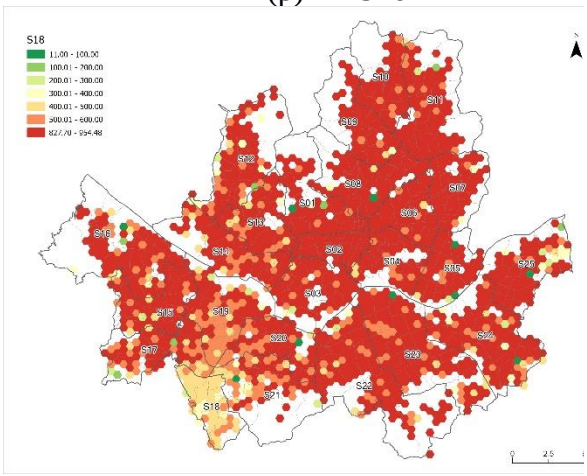
(o) S15



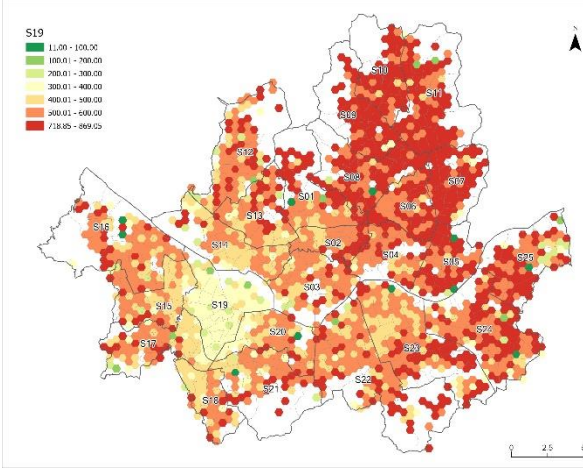
(p) S16



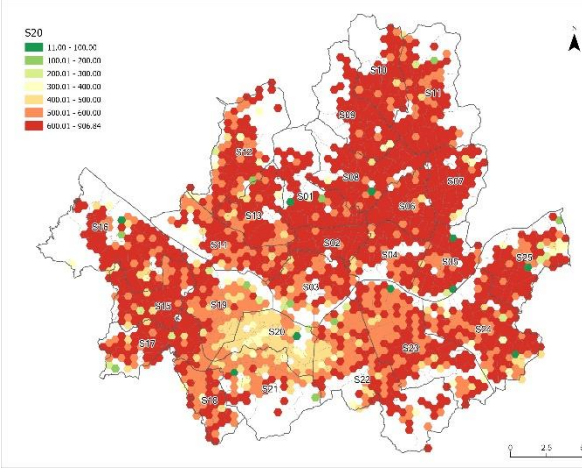
(q) S17



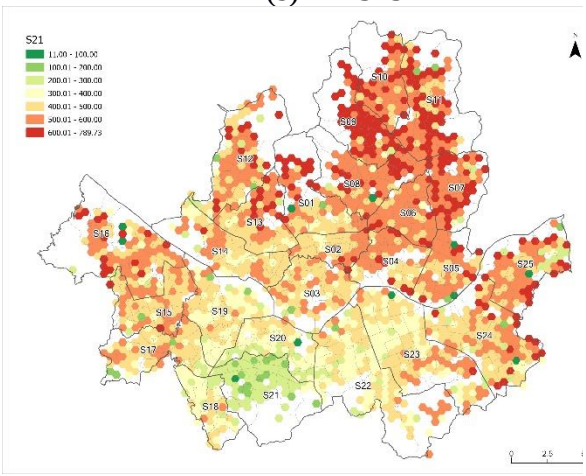
(r) S18



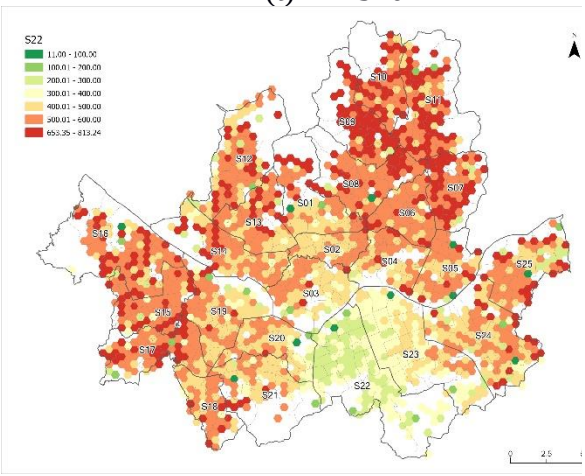
(s) S19



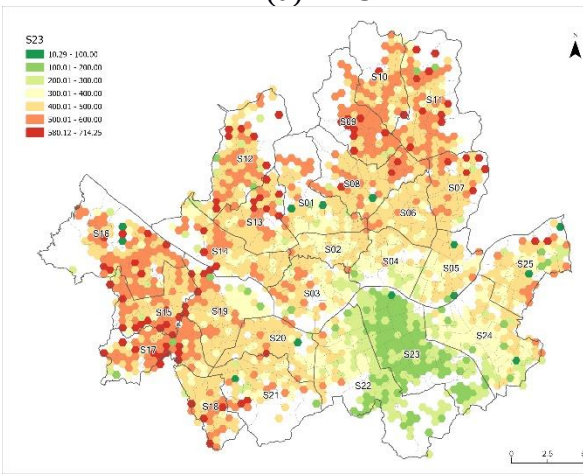
(t) S20



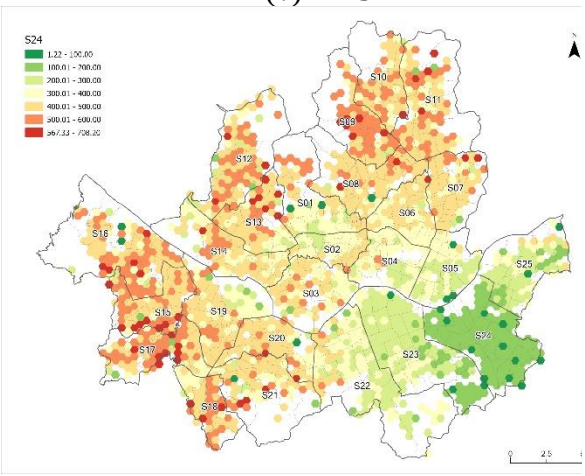
(u) S21



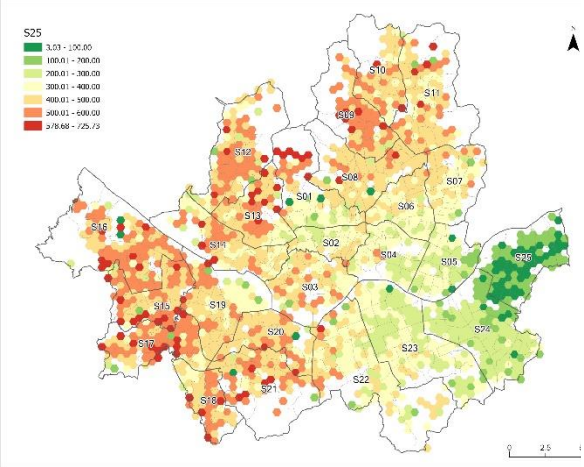
(v) S22



(w) S23



(x) S24



(y) S25

Figure 4.5 Local relevance scores by inflow population of 25 *gu*s.

4.6. Conclusion

This research proposes an explainable deep learning approach with consideration of geographic context to understand nonlinear relationships between human mobility and urban dynamics. We developed one-layered RNN model to predict survivorship of restaurants in a city and attempted to interpret the deep neural networks from geographical perspectives by applying the GLIME. The RNN model generates the interpretable number of hidden parameters with GLIME, and local relevance scores enable spatially explicit interpretations and the generation of new hypotheses for future study. Our findings especially highlight the importance of purchasing powers of human mobility instead of the amount of flow, which has been emphasized by the rich literature on business analytics (Frank & Dana, 1994; Kim et al., 2018; Shebl et al., 2021).

We acknowledge that the performance of the model is not as accurate as other deep learning models demonstrated in other applications such as image detection and natural language processing. The results with relatively poor predictions can be explained by the nature of complexity in human mobility and urban dynamics compared to the natural

sciences or computer sciences. For example, although image classification problems handle a great amount of data with more complicated structures of neural networks, at least, the patterns are quite predictable as people can catch them immediately. However, socioeconomic patterns cannot be clearly observed and predicted. There are large uncertainties and irregularities in human decision-making that result in unexpected patterns and outliers, which hinders the deep neural networks from predicting the trajectory accurately. Moreover, the relationships between human mobility and urban dynamics are still investigated to identify their precedence relation. In other words, it is still arguable whether urban structures pull visitors travels or people's willingness to travel build urban structures. Despite the veiled relationships, this study proves that human mobility patterns strongly affect survivability of restaurants, which arouse regional ups and downs regarding economic developments. It also explains the impact of human mobility that depends on mobility origins representing purchasing power.

The explainable deep neural networks suggested in this study demonstrate for human mobility and urban dynamics improve prediction as well as shed light on complex mechanisms underlying their relationships. In particular, the study suggests that we need to delve into the impact of path dependency and inter-dependencies among people's travels. Most importantly, the findings emphasize the importance of spatial heterogeneities in drivers of urban dynamics across the city.

5. General Conclusion

Through this dissertation, I suggest methodological frameworks and models that enhance our understanding of human mobility, urban dynamics, and their interactions with micro-scale geospatial big data. Three individual studies are conducted to (1) evaluate human mobility patterns imprinted in various data sources including social media data and public data; (2) understand micro-scale urban dynamics through publicly available individual-scale data; and (3) explain urban dynamics with human mobility through highly accurate and interpretable neural network models.

The first study in this dissertation addresses a challenge of understanding diverse characteristics of human mobility patterns across geographic scales by utilizing fine-scale data extracted from diverse sources. Due to discordance of spatiotemporal resolution between data sources, aggregation at a certain level is required. An origin–destination (OD) matrix provides mobility patterns among spatial units within a given temporal scale). A new measure, Spatial Weighted SSIM (SpSSIM), is developed to solve the sensitivity problems of SSIM. The applicability of SpSSIM is evaluated with OD matrices generated from his three data sources in San Diego County, CA: Census-based longitudinal employer-household dynamics origin-destination employment statistics, Twitter, and Instagram. This case study shows that SpSSIM can capture the similarity of migration patterns between different datasets by distance. The results in this study enhance our understanding of diversity and complexity in human mobility and broaden opportunities of using social media data in human mobility studies.

The second study in this dissertation investigates the value of micro-scale public open data for better understanding of urban dynamics. This study uses license data in Seoul,

Korea at the individual-level to discover patterns of spatiotemporal changes in restaurant locations because the restaurant industry is one of several industries that contributes significantly to the urban economy, and spatiotemporal differences in survivability are important indicators of urban dynamics. By applying three exploratory analytics, including hot spot analysis, trends analysis of spatial clusters, and space-time scan statistics, this study identifies the continuous temporal changes in spatial clusters of restaurants. Spatiotemporal ups and downs in restaurant businesses and expected shorter lifespans of restaurants in suburb areas verify deepening regional inequalities and economic disparities. The perspective offered by this study can be used to assess market conditions using spatiotemporally fine scale data, which can assist private and governmental decision-making processes regarding economic development and growth.

The third study in this dissertation identifies the relationships between urban dynamics and human mobility with recurrent neural networks based on the survival analysis framework. To enhance explainability of the neural network model, this study proposes Geographically Localized Interpretable Model-agnostic Explanation (GLIME) by extending Local Interpretable Model-agnostic Explanation (LIME) within geographic context. The geographically explainable deep neural networks demonstrate nonlinear relationships between human mobility and urban dynamics with improved prediction as well as shed light on complex mechanisms underlying human mobility and urban dynamics. Through the case study of restaurant survivability in Seoul, Korea with flow population estimated by cell-phone uses, origins of population flow are key factors determining entrepreneurships of restaurant businesses. Moreover, through the proposed localized explanatory approach, it is verified that the effect varies by origins regarding the strength and range of impacts. The

findings remind importance of spatial heterogeneities in complexity of human mobility and urban dynamics.

This work takes a new perspective on long-standing research questions related to cities by integrating new forms of spatial information with geospatial big data analytics. By entering new territory, however, this research faces numerous issues. Though these concerns are not severe enough to invalidate the research results, it is nonetheless important to fully disclose these caveats. Since data-integration is a core component of this research, it should be noted that aligning multiple data sources can be frequently problematic. Data-alignment is a concern throughout this dissertation regarding time and space. The Modifiable Area Unit Problem (MAUP) and the Modifiable Temporal Unit Problem (MTUP) are the key challenges and should be examined further to identify the sensitivities of suggested methodological approaches.

Additionally, although the new forms of data used here offer many advantages in terms of spatiotemporal resolution, the data must be systematically updated to maintain this benefit. Restaurants, for instance, can go out of business or change ownership, and such changes can produce different results. Further, since this data is gathered through publicly available application programming interfaces (APIs), these methodologies are vulnerable to changes in the companies' data-sharing policies. Indeed, any data restrictions may prevent future analysis. The recent availability of user-contributed, fine-grained spatial data enables this research, but it is critical when working with such data to be mindful of privacy concerns. As Goodchild (2007) has discussed, volunteered geographic information from services such as Twitter or Yelp provides researchers new advantages in terms of the quantity and resolution of spatial data. However, researchers must be mindful to safeguard

this personal information against potential abuses (Harvey, 2010). Though this study has taken precautions against revealing personal information, any future applications of this work must be similarly concerned with protecting individuals' information.

Despite these fundamental difficulties, this dissertation provides a comprehensive framework for studying human mobility and urban dynamics, particularly from geographic perspectives. By incorporating detailed spatiotemporal data into big data analytic models, the framework allows researchers and policymakers to understand human mobility, urban dynamics, and their interactions at a finer scale.

First, I highlight the importance of spatial configurations underlying data in the study on human mobility and urban dynamics. By adding spatial components on existing methodology, aspatial analytic tools are successfully applied to reveal spatial patters in diverse types of datasets. In Chapter 2, for example, applying a spatial weight matrix on the image analytic tool, SSIM, is proposed to consider spatial relationships, such as geographic closeness, in the arrangement of OD matrices because similarities in human mobility vary by distance. Understanding the differences between datasets that vary in space provides fundamental knowledge for making the best use of data integration. Moreover, in Chapter 4, Geographic boundaries are also proposed as an explanatory tool of neural networks for interpreting a "black box" model with geographic context. Based on the idea of spatial heterogeneity, spatial administrative units have distinctive characteristics in terms of demographic and socioeconomic status although the boundaries are artificially and arbitrary defined. By extending local explanations geographically, I propose a method for interpreting deep learning models in urban studies, where geographic context in a phenomenon is essential, with better predictions.

Second, I broaden the applicability of deep learning models in the social sciences with a focus on explainability to provide comprehensive explanations of complex social phenomena for a better society. GLIME, in Chapter 4, is the tool to explain a deep learning model with geographic perspectives that combines the ideas of Explainable AI (XAI) and GeoAI. As this study is an initial study, more cases and models will be investigated with motivating questions on the complex relationships between human mobility and urban dynamics.

Finally, I maximize the utility of publicly accessible data. Although we are in the era of Big Data, data accessibility is still challenging. While big data analytics can play a significant role in business, potential benefits can be limited to large companies and corporations due to high cost in data access, data management, and data analyses. Therefore, a data analytic framework using free publicly accessible data is important. In Chapter 3, I offer a methodological framework that utilizes public data to evaluate local market situations. Furthermore, in Chapter 4, I suggest a model that predicts survivability of businesses with the limited information. Rather than analyzing comprehensive information including location characteristics and potential consumers' demographic and socioeconomic backgrounds, the model provides accurate predictions with only the origins of population flows. The results can be less accurate than extensive big data analytic models operated by data companies; however, without paying for expensive data, it provides reasonable and feasible solutions that are beneficial for individuals and governments to make better decisions in business application from locating their new businesses to monitoring market saturation. These efforts will contribute to fostering economic growth and development.

References

Alonso, W. (1960). A theory of the urban land market. *Papers in Regional Science, 6*(1), 149-157.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis, 27*(2), 93–115.

Arribas-bel, D. (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography, 49*, 45–53.

Austin, S. B., Melly, S. J., Sanchez, B. N., Patel, A., Buka, S., & Gortmaker, S. L. (2005) Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments. *American Journal of Public Health, 95*(9), 1575–1581.

Batty, M. (2007). Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. The MIT press.

Batty, M. (2012). A Generic Framework for Computational Spatial Modelling. In *Agent-Based models of Geographical Systems*, eds. Heppenstall, A.J., Crooks, A.T., See, L.M., & Batty, M. 19-50, Springer.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics, 518*, 481–518.

Batty, M., Xie, Y., & Sun, Z. (1999). Modeling urban dynamics through GIS-based cellular automata. *Computers, environment, and urban systems, 23*(3), 205-233.

Behara, K. N., Bhaskar, A., & Chung, A. B. E. (2017). Insights into geographical window based SSIM for comparison of OD matrices. In *Australasian Transport Research Forum (ATRF)*, 39th, 2017, Auckland, New Zealand.

Behara, K. N., Bhaskar, A., & Chung, A. B. E. (2018). Levenshtein distance for the structural comparison of OD matrices. In *Australasian Transport Research Forum (ATRF)*, 40th, 2018, Darwin, Northern Territory, Australia.

Boulos, M. N. K., Peng, G., & VoPham, T. (2019). An overview of GeoAI applications in health and healthcare. *International journal of health geographics, 18*(1), 1-9.

Brunet, D., Vrscay, E. R., & Wang, Z. (2012). On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing, 21*(4), 1488–1495.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(3), 431-443.

California Department of Transportation. (2019). *California Highways with 70 MPH Speed Limits*. Retrieved from http://www.dot.ca.gov/hq/roadinfo/70mph.htm

Church, R. L., & Murray, A. T. (2009). *Business Site Selection, Location Analysis, and GIS*. John Wiley & Sons.

Clayton, C. (1977). The structure of interstate and interregional migration: 1965-1970. *The Annals of Regional Science, 11*(1), 109–122.

Cresswell, T. (2012). Geographic Thought: A Critical Introduction. John Wiley & Sons.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., & Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science, 29*(5), 720–741.

Djukic, T. (2014). *Dynamic OD Demand Estimation and Prediction for Dynamic Traffic Management* (doctoral dissertation). Delft University of Technology, Delft, Netherlands.

Dock, J. P., Song, W., & Lu, J. (2015). Evaluation of dine-in restaurant location and competitiveness: Applications of gravity modeling in Jefferson County, Kentucky. *Applied Geography, 60*, 204-209.

Dodge, S., Weibel, R., Ahearn, S. C., Buchin, M., & Miller, J. A. (2016). Analysis of movement data. *International Journal of Geographical Information Science, 30*(5), 825–834.

Dorji, U. J., Plangprasopchok, A., Surasvadi, N., & Siripanpornchana, C. (2019, November). A machine learning approach to estimate median income levels of sub-districts in Thailand using satellite and geospatial data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 11-14).

Drezner, T. (2014). A review of competitive facility location in the plane. *Logistics Research, 7*(1), 114.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Flake, G. W. (1999). The Computational Beauty of Nature. MIT Press.

Forrester, J. (1969). *Urban dynamics*. MIT Press.

Fotheringham, A. S. (1985). Spatial competition and agglomeration in urban modelling. *Environment and planning A: Economy and Space, 17*(2), 213-230.

Frank, D., & Dana, F. (1994). Purchasing power: Consumer organizing, gender, and the Seattle labor movement, 1919-1929. Cambridge University Press.

Gao, S. (2015). Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation, 15*(2), 86-114.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J. A., McKenzie, G., Ju, Y., Gong, L., Adams, B., & Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science, 31*(6), 1245–1271.

Gao, Y., Li, T., Wang, S., Jeong, M. H., & Soltani, K. (2018). A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science, 32*(7), 1304–1325.

García-palomares, J.C., Salas-olmedo, M.H., Moya-gómez, B., Condeço-melhorado, A., & Gutiérrez, J. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities, 72*, 310–319.

Garrison, W. L., & Marble, D. F. (1964). Factor-analytic study of the connectivity of a transportation network. *Papers in Regional Science, 12*(1), 231–238.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences, 114*(50), 13108-13113.

Getis, A. (2008) A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis, 40*(3), 297–309.

Getis, A. & Ord, J. K. (1992). The Analysis of Spatial Association by Distance Statistics. *Geographical Analysis, 27*(4), 286–306.

Ghosh, A. & Craig, C.S. (1983). Formulating retail location strategy in a changing environment. *Journal of Marketing, 47*(3), 56-68.

Giannotti, F., & Pedreschi, D. (2008). Mobility, Data Mining and Privacy: Geographic knowledge discovery. Springer.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics*. 80-89.

Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social media update 2016*. Pew Research Center: http://www.pewinternet.org/2016/11/11/social-media-update-2016

Hagenauer, J., & Helbich, M. (2022). A geographically weighted artificial neural network. *International Journal of Geographical Information Science, 36*(2), 215-235.

Han, S. Y., Tsou, M. H., & Clarke, K. C. (2018). Revisiting the death of geography in the era of Big Data: The friction of distance in cyberspace and real space. *International Journal of Digital Earth, 11*(5), 451-469.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271.

Henry, K.A., Niu, X., & Boscoe, F.P. (2009) Geographic disparities in colorectal cancer survival. *International Journal of Health Geographics, 8*(1), 48.

Horner, M. W., & Schleith, D. (2012). Analyzing temporal changes in land-use-transportation relationships: A LEHD-based approach. *Applied Geography, 35*(1–2), 491–498.

Hotelling, H. (1929). Stability in competition. *Economic Journal, 39*(153), 41–57.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics, 63*(1), 109–118.

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status, and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science, 30*(9), 1873–1898.

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

Hurst, M. E. (1972). A Geography of Economic Behavior. North Scituate, Massachusetts: Duxbury Press.

Hyde, Z. (2014). Omnivorous gentrification: Restaurant reviews and neighborhood change in the downtown eastside of Vancouver. *City and Community, 13*(4), 341–359.

ImageNet, 2017, *ImageNet Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)*. https://image-net.org/challenges/LSVRC/2017/index.php

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri B. (2020). GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, *International Journal of Geographical Information Science, 34*(4), 625-636.

Jeong, D. G. & Yoon, H. Y. (2017). Survival analysis of food business establishments in a major retail district and its extended area. *Journal of The Architectural Institute of Korea Planning & Design, 33*(3), 57–68.

Jin, C. & Murray, A. T. (2021). Exploring Public Open Data: Spatiotemporal Dynamics of Restaurant Entrepreneurships in Seoul, Korea, *International Journal of Geospatial and Environmental Research, 8*(3). 5.

Jin, C., Nara, A., Yang, J. A., & Tsou, M. H. (2020). Similarity Measurement on Human Mobility Data with Spatially Weighted Structural Similarity Index (SpSSIM). *Transactions in GIS, 24*(1), 104-122.

Joassart-Marcelli, P., Bosco, F. J., & Delgado, E. (2014). *Southeastern San Diego's Food Landscape: Challenges and Opportunities*. Policy Report. San Diego, CA: Department of Geography, San Diego State University and Project New Village.

Jung, S. & Jang, S. (2019). To cluster or not to cluster?: Understanding geographic clustering by restaurant segment. *International Journal of Hospitality Management, 77*, 448–457.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering, 29*(10), 2318-2331.

Kedron, P., Frazier, A.E., Trgovac, A. B. and Fotheringham, A. S. (2021). Reproducibility and replicability in geographical analysis. *Geographical Analysis, 53*(1), 135-147.

Kendall, M. G. (1948). *Rank Correlation Methods*. Oxford, UK: Griffin.

Kim, H. & Lee, S. (2019) A study on the factors affecting the revenue in Seoul's side street trade areas. *Seoul Studies, 20*(1), 117–134.

Kim, H., Lee, K., Lee, Y. & Song Y. (2021). Restaurants' survival in the era of COVID-19: A case study of Seoul. *Journal of the Korean Geographical Society, 56*(1), 35-51.

Kim, W., Yim, J. & Song, A. (2018). Spatio-temporal changes of the agglomerated marketplace by use of the pedestrian flow data. *Journal of Korean Cartographic Association, 18*(1), 49–63.

Krogstad, J. M. (2015). *Social media preferences vary by race and ethnicity*. Pew Research Center: http://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods, 26*(6), 1481–1496.

Lansley, G., Smith, M. De, Goodchild, M. & Longley, P. (2018). Big data and geospatial analysis. *Big Data and Research*, 547–570

Larsen, J., Urry, J., & Axhusen, K. (2006). *Mobilities, Networks, Geographies*. Ashgate.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lee, K., Park, S. & Shin, H. (2020). Online information retrieval and changes in the restaurant location: The case study of Seoul. *Journal of the Economic Geographical Society of Korea, 23*(1), 56-70.

Leitner, M., & Helbich, M. (2013). The impact of hurricanes on crime: A spatio-temporal analysis in the city of Houston, Texas. *Cartography and Geographic Information Science, 38*(2), 213–221.

Li, H., Liu, J., & Zhou, X. (2018). Intelligent map reader: A framework for topographic map understanding with deep learning and gazetteer. *IEEE Access, 6*, 25363-25376.

Li, M., Gao, S., Lu, F., Liu, K., Zhang, H., & Tu, W. (2021). Prediction of human activity intensity using the interactions in physical and social spaces through graph convolutional networks. *International Journal of Geographical Information Science, 35*(12), 2489–2516.

Li, W. (2020). GeoAI: Where machine learning and big data converge in GIScience. *Journal of Spatial Information Science, 2020*(20), 71-77.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography, 32*, 591–600.

Li, Y., Steiner, M., Wang, L., Zhang, Z. L., & Bao, J. (2013). Exploring venue popularity in Foursquare. *Proceedings IEEE INFOCOM*, 3357–3362.

Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science, 45*(2), 97–110.

Lin, Y., Schootman, M., & Zhan, B.F. (2015). Racial/ethnic, area socioeconomic, and geographic disparities of cervical cancer survival in Texas. *Applied Geography, 56*, 21–28.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems, 14*(4), 463–483.

Lo, L. (1991). Substitutability, spatial structure, and spatial interaction. *Geographical Analysis, 23*(2), 132-146.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of twitter usage. *Environment and Planning A, 47*(2), 465–484.

Mann, H. B. (1945). Nonparametric Tests Agianst Trend. *Econometrica, 13*(3), 245–259.

Manson, S. M. (2001). Simplifying complexity: a review of complexity theory. *Geoforum, 32*(3), 405-414.

Martín, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLoS ONE, 12*(7): e0181701.

Mazur, M. & Manley, E. (2016). Exploratory models in a time of Big Data. *Interdisciplinary Science Reviews, 41*(4), 366–382.

Miller, H. J. (2010). The Data Avalanche is Here. Shouldn't We Be Digging. *Journal of Regional Science, 50*(1), 181–201.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80*(4), 449-461.

Miller, H. J., & Han, J. (2009). Geographic Data Mining and Knowledge Discovery. CRC Press.

Miller, H. J., & Shaw, S.-L. (2015). Geographic Information Systems for Transportation in the 21st Century. *Geography Compass, 9*(4), 180–189.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging. *Journal of Regional Science, 50*(1), 181–201.

Minner, J. S. & Shi, X. (2017). Churn and change along commercial strips: Spatial analysis of patterns in remodeling activity and landscapes of local business. *Urban Studies, 54*(16), 3655–3680.

Mulligan, G. F. (1984) Agglomeration and central place theory: A Review of the Literature. *International Regional Science Review, 9*(1), 1–42.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071-22080.

Murray, A. T., Koschinsky, J., Liu, Y., Rey, S. J. & Brown, L. A. (2013) Are foreclosures contagious? An exploratory space-time analysis of franklin county, Ohio, 2001-2008. *International Journal of Applied Geospatial Research, 4*(4), 19–36.

Nakaya, T. & Yano, K. (2010). Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. Transactions in GIS, 14(3), 223–239.

Nara, A., Yang, X., Machiani, S. G., & Tsou, M.-H. (2017). An integrated evacuation decision support system framework with social perception analysis and dynamic population estimation. *International Journal of Disaster Risk Reduction, 25*, 190–201.

Newman, K. (2010). Go public! Journal of the American Planning Association, 76(2), 160–171.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE, 7*(5): e37027.

Nystuen, J. D., & Dacey, M. F. (1961). A graph theory interpretation of nodal regions. *Papers of the Regional Science Association, 7*(1), 29–42.

O'Sullivan, D. (2004). Complexity science and human geography. *Transactions of the Institute of British geographers, 29*(3), 282-295.

Openshaw, S. & Openshaw, C. (1997). *Artificial intelligence in geography*. John Wiley & Sons.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System, 1*(4), 335-358.

Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z., & Colizza, V. (2017). Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. *Royal Society Open Science, 4*: 160950.

Papadakis, E., Adams, B., Gao, S., Martins, B., Baryannis, G., & Ristea, A. (2022). Explainable artificial intelligence in the spatial domain (X-GeoAI). *Transactions in GIS, 26*(6), 2413-2414.

Pearce, D. G. (1996). Domestic Tourist Travel in Sweden: A Regional Analysis. *Geografiska Annaler. Series B, Human Geography, 78*(2), 71–84.

Peng, B., Liu, X., Meng, Z., & Huang, Q. (2019, November). Urban flood mapping with residual patch similarity learning. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 40-47.

Pollard, T., Taylor, N., & Vuren, T. Van. (2013). Comparing the Quality of OD Matrices: in Time and Between Data Sources, In *Proceedings of the European Transport Conference*. Frankfurt, Germany: AET.

Poon, J. & Pandit, K. (1996). The geographic structure of cross-national trade flows and region states. *Regional Studies, 30*(3), 273–285.

Potamias, M., Patroumpas, K., & Sellis, T. (2006). Sampling trajectory streams with spatiotemporal criteria. In *Proceedings of the International Conference on Scientific and Statistical Database Management*, 275–284.

Prayag, G., Landre, M., & Ryan, C. (2012). Restaurant location in Hamilton, New Zealand: Clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management, 24*(3), 430–450.

Reilly, W. J. (1929). *Method for the study of retail trade relationships*. Research Monograph No. 4. Austin, Texas: University of Texas Press.

Ryu, H. Y. & Park, J. (2019) A study on the variation process of commercial gentrification phase in residential area in Seoul. *Journal of Korea Planning Association, 54*(1), 40–51.

Salvador, S. & Chan, P. (2004). Determining the Number of Clusters / Segments in Hierarchical Clustering / Segmentation Algorithms, In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, 576-584.

Samek, W., Wiegand, T., & Müller, K. R. (2018). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *ITU Journal ICT Discoveries, 1*(1), 39-48.

Shebl, S., Abd Elhady, D., & Refaat, A. (2021). Economic and Social Factors Affecting the Purchasing Power of Customers in Fast Food Restaurants (Applied in Marsa Matrouh City). *Journal of Tourism, Hotels and Heritage, 2*(1), 1-14.

Self, J. T., Jones, M. F., & Botieff, M. (2015). Where restaurants fail: A longitudinal study of micro locations. *Journal of Foodservice Business Research, 18*(4), 328–340.

Shaw, S. L., & Sui, D. (2018). Human dynamics research in smart and connected communities. Springer

Shaw, S. L., Tsou, M. H. & Ye, X. (2016). Editorial: Human dynamics in the mobile and big data era. *International Journal of Geographical Information Science, 30*(9), 1687–1693.

Shin, W. J. & Shin, W.H. (2009). Spatial patterns of retail stores in Seoul, Korea. *Korea Real Estate Review, 19*(2), 279–296

Singleton, A. & Arribas-bel, D. (2021). Geographic data science. *Geographical Analysis, 53*(1), 61-75.

Smith, C. M., Le Comber, S. C., Fry, H., Bull, M., Leach, S. & Hayward, A. (2015). Spatial methods for infectious disease outbreak investigations: Systematic literature review. *Eurosurveillance, 20*(39), 1-21.

Smith, H. T. R. (1970). Concepts and Methods in Commodity Flow Analysis. *Economic Geography, 46*, 404–416.

Smith, S. (1985). Location patterns of urban restaurants. *Annals of Tourism Research, 12*(4), 581–602.

Snyder, L. S., Karimzadeh, M., Chen, R., & Ebert, D. S. (2019, November). City-level geolocation of tweets for real-time visual analytics. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 85-88.

Steiger, E., de Albuquerque, J.P. & Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS, 19*(6), 809–834.

Sun, Y. & Paule, J. D. G. (2017). Spatial analysis of users-generated ratings of yelp venues. Open Geospatial Data, *Software and Standards, 2*(1), 5.

Sun, Y., Fan, H., Li, M., & Zipf, A. (2015). Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B: Planning and Design, 43*(3), 480–498.

Thrift, N. (1999). The place of complexity. *Theory, Culture & Society, 16*(3), 31-69.

Tien Bui, D., Shahabi, H., Shirzadi, A., Chapi, K., Hoang, N. D., Pham, B. T., ... & Saro, L. (2018). A novel integrated approach of relevance vector machine optimized by imperialist competitive algorithm for spatial modeling of shallow landslides. *Remote Sensing, 10*(10), 1538.

Tsou, M. H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science, 42*(sup1), 70–74.

Tu, W., Cao, J., Yue, Y., Shaw, S.L., Zhou, M., Wang, Z., Chang, X., Xu, Y. & Li, Q. (2017) Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science, 31*(12), 2331–2358.

Tukey, J. W. (1962). The Future of Data Analysis. *Annals of the Institute of Statistical Mathematics, 33*(1), 1–67.

Urry, J. (2002). Mobility and proximity. *Sociology, 36*(2), 255-274.

Vandell, K. & Carter, C. (1994). Retail store location and market analysis: A review of the research. *Journal of Real Estate Literature, 2*(2), 13-45.

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 102470.

VoPham, T., Hart, J. E., Laden, F., & Chiang, Y. Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environmental Health, 17*(1), 40.

Wan, N., Zhan, F.B., Lu, Y. & Tiefenbacher, J.P. (2012). Access to healthcare and disparities in colorectal cancer survival in Texas. *Health & Place, 18*(2), 321–329.

Wang, Q. & Taylor, J. E. (2014). Quantifying human mobility perturbation and resilience in hurricane sandy. *PLoS ONE, 9*(11): e112608.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simmoncelli, E. P. (2004). Image quality assessment: form error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen K., & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences, 112*(38), 11887–11892.

Westfall, P. H., & Young, S. S. (1989). P value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association, 84*(407), 780–786.

Widaningrum, D. L., Surjandari, I. & Sudiana, D. (2020). Discovering spatial patterns of fast-food restaurants in Jakarta, Indonesia. *Journal of Industrial and Production Engineering*, 37(8), 403-421.

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS ONE, 9*(5): e97010.

Xia, Y., Wang, G. Y., Zhang, X., Kim, G. B., & Bae, H. Y. (2011). Spatio-temporal Similarity Measure for Network Constrained Trajectory Data. *International Journal of Computational Intelligence Systems, 4*(5), 1070–1079.

Xing, T., Gu, Y., Song, Z., Wang, Z., Meng, Y., Ma, N., ... & Chai, H. (2019, November). A traffic sign discovery driven system for traffic rule updating. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 52-55.

Xiong, X., Ozbay, K., Jin, L., & Feng, C. (2020). Dynamic origin–destination matrix prediction with line graph neural networks and Kalman filter. *Transportation Research Record, 2674*(8), 491-503.

Xu, M., Li, Z., Shi, Y., Zhang, X., & Jiang, S. (2015). Evolution of regional inequality in the global shipping network. *Journal of Transport Geography, 44*, 1–12.

Yin, Y., Sunderrajan, A., Huang, X., Varadarajan, J., Wang, G., Sahrawat, D., ... & Ng, S. K. (2019, November). Multi-scale graph convolutional network for intersection detection from gps trajectories. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 36-39.

Yin, Z., Xiong, H., Zhou, X., Goldberg, D. W., Bennett, D., & Zhang, C. (2019, November). A deep learning based illegal parking detection platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 32-35.

Yu, K. & Lee, S. (2017). An analysis of factors affecting the agglomeration of food industry in Seoul using geographically weighted regression model. *Journal of The Korean Regional Development Association, 29*(2), 189–210

Yuan, M. & Nara, A. (2015). Space-time Analytics of Tracks for the Understanding of Patterns of Life. In *Space-Time Integration in Geography and GIScience*, 373–398. Springer.

Yuan, Y. & Raubal, M. (2014). Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. *International Journal of Geographical Information Science, 28*(3), 496–520.

Zhai, S., Xu, X., Yang, L., Zhou, M. & Zhang, L. (2015). Mapping the popularity of urban restaurants using social media data. *Applied Geography, 63*, 113–120.

Zhang, H., Zhou, X., Tang, G., Xiong, L. & Dong, K. (2021). Mining spatial patterns of food culture in China using restaurant POI data. *Transactions in GIS, 25*(2), 579-601.

Zheng, Y. & Zhou, X., (2011). *Computing with spatial trajectories*. Springer.

Zukin, S. (2009). Naked City: The Death and Life of Authentic Urban Places. Oxford University Press

Zukin, S., Lindeman, S. & Hurson, L. (2017). The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *Journal of Consumer Culture, 17*(3), 459–479.

Zukin, S., Trujillo, V., Frase, P., Jackson, D., Recuber, T. & Walker, A. (2009). New retail capital and neighborhood change: Boutiques and gentrification in New York City. *City & Community, 8*(1), 47–64.