EXPLORING HUMAN ACTIVITIES IN CITIES THROUGH DATA AND MODEL

by

Yang Zhou
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Social Science

Committee:

______________________________	Dr. Andrew Crooks, Committee Chair
______________________________	Dr. William G. Kennedy, Committee Member
______________________________	Dr. Arie Croitoru, Committee Member
______________________________	Dr. Andreas Züfle, Committee Member
______________________________	Dr. Jason Kinser, Department Chairperson
______________________________	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
______________________________	Dr. Fernando Miralles-Wilhelm, Dean, College of Science
Date: _________________________	Spring Semester 2021 George Mason University Fairfax, VA

Exploring Human Activities in Cities Through Data and Model

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Yang Zhou
Master of Science
Cornell University, 2015
Bachelor of Science
New York University, 2013

Director: Andrew Crooks, Professor
Department of Computational and Data Sciences

Spring Semester 2021
George Mason University
Fairfax, VA

Acknowledgement

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

First and foremost, I would like to thank my advisor and dissertation chair, Dr. Andrew Crooks, for guiding me through the dissertation writing process. He has provided insightful feedback that pushed me to sharpen my thinking and brought my work to a higher level. During the four years that I worked as a graduate research assistant under his supervision, I have gained so much experience in developing models and doing research. I must also thank my committee members, Dr. William Kennedy, Dr. Andreas Züfle and Dr. Arie Croitoru for their great advice. They have provided feedback that challenged me to think deeper and further my research.

In addition, I would like to thank my parents, my spouse Xiaotong, my best friend form school Salwa, and my cat "Pancake" for their mental support and encouragement, I could not have completed this dissertation without them.

Table of Contents

Page

List of Tables

Table Page

List of Figures

List of Abbreviations

Computational Social Science	CSS
World Wide Web	WWW
Natural Language Processing	NLP
Latent Dirichlet Allocation	LDA
Term Frequency–Inverse Document Frequency	TF-IDF
Agent Based Model	ABM
Points of Interest	POI
Internet of Things	IoT
User-generated Content	UGC
Coordinated Universal Time	UTC
Eastern Standard Time	EST
International Federation of Association Football	FIFA

Abstract

EXPLORING HUMAN ACTIVITIES IN CITIES USING DATA AND MODEL

Yang Zhou, PhD

George Mason University, 2021

Dissertation Director: Andrew Crooks

Human activities in cities, such as eating foods in restaurants and attending special events, are traditionally measured using surveys, questionnaires, and interviews. While such methods have provided valuable insights, there are issues, such as cost when exploring large geographical areas during long periods of time. With the development of social media platforms and open data initiatives, a large amount of data that contains information on people's interests and opinions, geolocation, and timestamps are becoming available to the general public. The question is, how can we utilize these datasets with three dimensions (i.e., textual, geographical, and temporal) to explore questions related to human activities in cities? How do we use the results discovered from the data to better study and plan our cities? This dissertation research used a multi-disciplinary Computational Social Science approach to explore these questions. The research specifically demonstrates how to explore food related discussions, special events and their impact on traffic flow using open sources data. It also shows the potential and

methods to explore human activities in cities using social media data together with other data sources, and through agent-based simulation and analysis, the results can help planners and engineers to study, plan, and manage our cities.

Chapter 1 : Introduction

1.1 Motivation of the Dissertation

Cities provide homes to more than half the world's population (UN, 2018), and the percentage is expected to grow in coming years. Understanding cities is a complex challenge, because they are composed of a lot of different people interacting with each other. Their activities and interactions have significant impact on the economy, public policy, public health, social trends, etc. Due to population change, cities are going to experience increasing pressures on their infrastructure, such as housing, transportation, communication, energy and water. The problem is, how can we study cities and the people living in cities using the new wealth of information to provide more effective methods compared to the traditionally surveys, questionnaires, and interviews? While the traditional methods have provided valuable insights, they are also expensive when exploring large geographical areas during long periods of time. Furthermore, many of these methods fail to capture the interactions of people. With the advancement of web technology, there is a huge volume of data generated on the web by its users every day. Nowadays, more and more people exchange their ideas and opinions on online platforms. Social media platforms allow people from all over the world to express their ideas, discuss topics, express their interests in things, and talk about what's going on around them. According to Twitter's report, there are 192 million daily active users on Twitter;

55 million daily active users are in the United States, which makes up 28.6% of Twitter's user base, and the number of international users has a 28 percent yearly increase (Twitter, 2021). The top hashtag used by the most people in 2020 was #COVID19 as the virus spread around the world, and it was tweeted nearly 400 million times (McGraw, 2020). Meanwhile, #StayHome was the third most popular hashtag of 2020, and we could see the change in people's habits through Twitter.

In turn, the data published on the web reflects their opinions, interests, and activities that are happening. What’s more, as social media messages are usually time-stamped, and users are allowed to tag their locations or simply check into Point of Interest (POI), some of the data has spatiotemporal information together with the text message, that provides the opportunity to study not only what people are doing and thinking about, but also when and where that happened. Social Media platforms, such as Twitter and Instagram have APIs (application programming interface) that make the data accessible for researchers. This provides a new opportunity for researchers interested in studying social science problems using computational methods, especially those related to people’s behaviors and activities.

Moreover, other sources of data are also becoming available on the internet. The government's open data, for example, the Metropolitan Transportation Authority of New York has published the Metro turnstile data online. This dataset manages both space and time information; it provides the count of entries and exits at each Metro station. Google Places provides information on places all over the world, using their API we can access the locations, coordinates, descriptions, and even ratings of over 200 million places.

OpenStreetMap provides a free editable map of the world built by volunteers and released with an open-content license. With the development of the internet, more and more data sources are emerging.

These data sources provide new opportunities to explore questions in social sciences using computational methods. This requires methods from Computational social science (CSS). The definition of CSS is “the interdisciplinary investigation of the social universe on many scales, ranging from individual actors to the largest groupings, through the medium of computation” (Cioffi-Revilla, 2017). In other words, it is an interdisciplinary field using computational methods to investigate questions in social science. In this research, I attempt to use CSS methods to explore human activities. Specifically, I will use data analytics methods and computational simulation to analyze social media data with spatiotemporal information to study food related discussion, events and their influence on traffic. The research methods and study results presented in this dissertation will contribute to the fields of computational social sciences, geography, and urban planning.

1.2 Previous Research

Spatiotemporal data are data that relate to both space and time, and they are analyzed to discover patterns and knowledge. Yang et al. (2020) introduced and explained the concept of "Big Spatiotemporal Data", which refers to Big Data with space and time stamps. This kind of data is obtained from various methods, such as social media, mobile devices, in-situ sensing, the Internet of Things (IoT), mobile phone navigation systems, etc. The field of Big Spatiotemporal Data Analytics has been rapidly growing over time. For example, spatiotemporal data has been used to study the evolution of cities, weather patterns, and global warming trends (Han et al., 2012). Many other problems related to the patterns in human activities are explored using bid spatiotemporal data, such as human mobility and accessibility (e.g. Pappalardo and Simini, 2018), hourly dynamics of urban population (e.g. Liu et al., 2018), segregation based on environment, gender, racial, ethnic and socioeconomic aspects (e.g. Park and Kwan, 2017).

The large amount of data obtained from human sensing devices provide a lot of useful information for research, such as trajectory data and other human mobility data reconstructed from mobile phone records. While some of these data are published online by government authorities for the general public for research or study purposes, most of them are not accessible to the general public. With the development of the World Wide Web (WWW) and social media platforms, the concept of information contribution and dissemination has been drastically altered, and the general public are able to publish and distribute user-generated content online (Wayant et al., 2012). Social media feeds may also be viewed as sensors of humans acting, to detect events and activities in which they participate, or commenting on others that are somehow affecting them, or catching their attention. Normally, social media feeds have geolocation information associated with them, such as coordinates or city names (Croitoru et al., 2012). Moreover, social media feeds, such as tweets, usually provide textual data, which we can use to understand the opinions of people and the activities that they participate in. The textual data from social media allows for topics and events detection. Scholars have utilized tweets with spatiotemporal textual information to detect earthquakes (e.g. Sakaki et al., 2010; Crooks et al., 2013; Poblete, Barbara, et al., 2018) and fire emergency (e.g. Wang et al., 2016; Noori and Mehra, 2020). Yuan et al. (2020) used geo-textual data from Twitter and TripAdvisor and topic modeling to explore relationships between places, and found the dominant topics of the communities in the network.

Social Media data, such as Twitter data have been used by many scholars to detect events, and the various methods have been summarized in papers (Weng et al., 2011; Hasan et al., 2019). For example, Weng & Lee (2011) developed a statistical model to filter away the trivial words in tweets and detect real time events, and used topic modeling to find topics in the events, such as discussions relating to elections. Walther and Kaisser (2013) implemented event detection method using another approach - by clustering tweets based on their geographical and temporal features, and they produced maps visualizing geospatial events. Liu et al. (2016) proposed a method to discover the core semantics of event from short texts, and it was able to extract core semantics accurately and efficiently. Twitter data has been used to track discourse about the COVID-19 pandemic (e.g. Chen et al., 2020; Medford et al., 2020), identify drug use (e.g. Tassone et. Al, 2020), predict a riot (e.g. Alsaedi et al., 2017), etc. Social media data with spatial and temporal information has been used to do explore problems related to locations. For example, Crooks et al. (2015) used information harvested from social media to study urban form and function, such as land use function, traffic situations, and topics in different areas. Social media content is also used to assess the impact area of a natural disaster (Panteras et al., 2015). In another study, Jenkins et al. (2016) developed a topic model to categorize geolocated tweets and used them to study the characteristics of places. Besides this work, Stefanidis et al. (2017) studied the spatiotemporal patterns of tweets during the Zika outbreak to explore people's participation in Twitter during a disease outbreak, while Koren et al. (2021) studied the effects of food insecurity and water insecurity through geolocated tweets.

Many scholars are using Point of Interest (POI) data obtained from map and navigation applications to study cities. For example, Hu and Han (2019) used POI data from AMap to classify urban functional areas in Guangzhou, China. While Wang et al. (2018) used POI data to study land use intensities, in another study, Zhang et al. (2020) examined the food culture of China using millions of items of restaurant point of interest data. While these studies have focused on land use and urban functions within cities, they did not study the actual activities of people in cities.

These studies show the potential to use spatial-temporal data collected from social media to study problems related to space and time, such as human activities. However, there are still some potential topics and research questions that could be explored using similar methods. Furthermore, they have not combined social media data with other data sources or computational simulation to examine the impact of human activities on other objects.

1.3 Research Questions

The main theme of this dissertation research is to study human activities in cities and their effects. This is done by processing data obtained from social media and other online sources, and analyze them using computational social science (CSS) methodologies and tools such as machine learning techniques and agent-based modeling (ABM). Human activities are the various actions done by people for interest, pleasure, or living. For example, it includes transportation, recreation, consuming foods, and exercise. Specifically, this research explores two kinds of human activities: (1) food related discussion (i.e., where, and when they like to talk about foods) and (2) the impact of events on traffic flow. Through this research, I attempt to explore these two types of human activities and demonstrate the potential to explore human activities by using computational social science methodologies to analyze data accessible to the general public on the internet. Figure 1.1 describes the questions I explore in this research. To begin with, social media data with spatiotemporal data can be used to explore a few questions. The textual messages can be used to study what topics are people talking about. The geotags provide geological location information, and can be used to study where are people talking about these topics. Timestamps of the social media messages can be used to answer when are people talking about these topics.

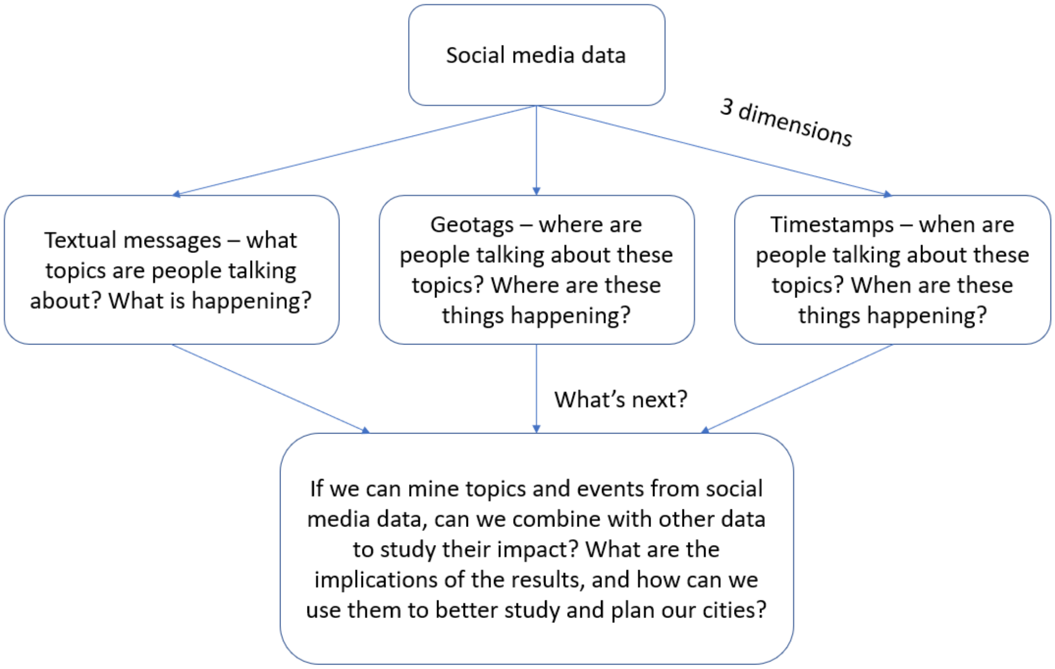

Social media data
3 dimensions
Textual messages – what topics are people talking about? What is happening?
Geotags – where are people talking about these topics? Where are these things happening?
Timestamps – when are people talking about these topics? When are these things happening?
What's next?
If we can mine topics and events from social media data, can we combine with other data to study their impact? What are the implications of the results, and how can we use them to better study and plan our cities?

Figure 1.1 High level research questions explored in this dissertation

Therefore, my first research question is: *How can we explore food related discussions using social media data with spatiotemporal information?* Food related questions are traditionally studied using surveys, questionnaires, and interviews. While such methods have provided valuable insights into food related topics, there are issues, such as cost when exploring large geographical areas during long periods of time. For example, the 2010 Census costs $42.11 per person (Goldenkoff, 2011). In the second chapter, I explore the potential of using social media data to gain insights into the food related discussions in New York City. Specifically, I investigated how people's interests in food change over meals and identified hot spots of popular foods. Expected outcomes of the research include discovering time frames for three meals (breakfast, lunch, dinner), popular foods during each meal, hot spots and cold spots of popular foods, and diagrams showing foods that are often discussed together. As such this chapter shows how social media analysis provides a new method to explore some food related questions using social media data.

Furthermore, if we can mine topics and events from social media data, how can we link them with other data sources to understand their impact? As a result, my second research question is: *How can we examine the impact of events on traffic flow by linking social media data and traffic data? What are the implications?* The development of social media and the increasing amount of open data sources provided by organizations and government authorities have provided a lot of potentials to study people living in cities. The third chapter combines Twitter data and traffic data published by the government authorities to study the impact of events on traffic flow in New York City. Tweets can be used to identify events that happened in the city, and the traffic data are used to explore how traffic flows change during those events. The research provides insight on patterns in the traffic data, different categories of events, and their impacts on traffic flows.

Last but not least, the question is what are the implications of the results? How can we use them to better study and plan our cities? Therefore, my last research question is: *How can we use these results to forecast the impact of different types of events on traffic flow?* For this goal, I write the fourth chapter based on the results of the previous chapter and build an agent-based traffic model to simulate the impact of special events on traffic flow under different scenarios.

1.4 Dissertation Outline

This dissertation is organized into 5 chapters. Chapter 1 introduces the motivation, background, and research questions of the dissertation. Chapter 2, 3, and 4 are three research studies that answer the three research questions presented in this chapter. Chapter 5 summarizes my results and discusses the limitations and identifies areas of future work.

Chapter 2 Exploring the Food Related Discussions in New York City through Social Media

This chapter explores the potential of using new sources of data generated through Web 2.0 technology, specifically that of social media data to gain insights into the food related discussions of the population using New York City as a case study. Specifically, I investigated how people's preferences with respect to foods change over the three daily meals, and where are the hot spots and cold spots of certain foods. Word frequency analysis and hot spot analysis are utilized to obtain the results, which were then visualized using maps, word-clouds, and concept diagrams. The research is able to identify time frames for three meals (breakfast, lunch, dinner), discover popular foods during different meals, hot spots and cold spots of popular foods, and foods that are often discussed together. As such this chapter shows how social media analysis provides a new method to explore food related discussions in cities.

2.1 Introduction

This chapter studies food related discussion through social media data. Food related discussions can reveal what, when, and where people eat. This information is important for public health since food is associated with many health problems, such as obesity and diabetes (e.g. Stunkard, 1959; van Dam et al., 2002). It is also valuable information for restaurants, in the sense that restaurants could use such information to meet the desires and needs of their potential customers. Traditionally, food related data, e.g., data on dietary habits has been collected by in-person and phone interviews (Moshfegh et al., 2009) and through surveys (Ku & Lee, 2000). Such collection methods covering large areas and over time tend to cost a large amount of time and money.

Nowadays, data available from social media services may provide a better solution to the problems. Social media services are currently defined as Web 2.0 Internet-based applications, and user-generated content (UGC) is a significant component of social media (Obar & Wildman, 2015). There is a growing population that uses social media, which represents a revolutionary new trend that is of interest to not only researchers but also business executives (Kaplan & Haenlein, 2010), as will be further described in the next section. One of the most popular social networking sites in the United States is Twitter, a micro-blogging and social networking service with approximately 7 million members (Golbeck et al., 2010). The tremendous amount of data created on social media has been used to investigate a wide range of phenomena, as further discussed in the next section.

This chapter explores the potential of using social media data, specifically that of Twitter to gain insights into the food related discussions of New Yorkers. Specifically, this chapter seeks to answer the following questions: What are the time frames during which New Yorkers have their three daily meals, namely breakfast, lunch, and dinner? Secondly, how do their interest in different foods change over different meals (breakfast, lunch, and dinner)? Thirdly, are there hot spots and cold spots for food related discussions in New York City? Finally, what foods do people like to discuss at the same time? The remainder of the chapter is organized as follows. Section 2.2 introduces the background of this study and talks about previous literature related to this research. Section 2.3 describes the datasets used for this chapter, and the methods used to address the questions above. In Section 2.4, the results of the study are shown. Section 2.5 discusses how the results address the four questions, what are their implications, and how the study can be enhanced by further work.

2.2 Related Work

Many studies have used data from multiple social media platforms to address food-related problems. Luca (2016) studied how consumer reviews on Yelp.com affected the revenue of both individual and chain restaurants, and its influence on their potential consumers. While in another study, Luca & Zervas (2016) presented several possible reasons for restaurants to commit review frauds on Yelp.com by utilizing its filtering algorithm to identify suspicious reviews. By assessing pictures posted on Instagram, Mejova et al. (2015) revealed the relationship between fast food, chain restaurants, and obesity. Grinberg et al. (2013) developed a model to identify verbal expression in social media data that refers to categories of activity such as nightlife, food or shopping, and found patterns in such activities. In another study, Guidry et al. (2015) revealed that Instagram has a new emerging crisis information form for food companies by analyzing the negative content of the 10 largest fast-food companies. Geotagged tweets were also used by Koren et al. (2021) to study food and water insecurity.

Social media data with spatial and temporal information has been used to do explore problems related to locations. For example, Crooks et al. (2015) used information harvested from social media to study urban form and function, such as land use function, traffic situation, and topics in different areas. Social media (Microblog) data with geolocation information has also been used to explore emotions on various topics and in different cities in the United States (Sikder and Züfle, 2019). While Panteras et al. (2015) utilized social media content to assess the impact area of a natural disaster, Jenkins et al. (2016) developed a topic model to categorize geolocated tweets and used them to study the characteristics of places. In another study, Stefanidis et al. (2017) studied the spatiotemporal patterns of tweets during the Zika outbreak to explore people's participation in Twitter during a disease outbreak. Moreover, to explore concept-related events, Stefanidis et al. (2017) extracted the co-occurrence of concepts (named entities) within single tweets regarding abortion. They created concept diagrams to capture the terms most frequently encountered in conjunction with the word abortion in the tweets. Furthermore, Shao et al. (2020) used geotagged and times-tamped tweets to develop a framework to predict massively unreported travel mode choices of Twitter users by measuring the similarity between a user without reported mode choice and the users with known travel modes. These studies show the potential to use spatial-temporal data collected from social media to study problems related to space and time.

Twitter data specifically, has been used to study many public health problems from tobacco use to influenza tracking. Paul & Dredze (2011) introduced theories and methods for analyzing user messages in social media to measure public health measures and suggested that Twitter has broad applicability for public health research. In another study, Myslín et al. (2013) performed sentiment analysis on tobacco-related tweets and built classifiers to detect tobacco-relevant tweets and their sentiment. Lampos et al.

(2010) developed a method to search tweets daily for symptom-related statements (e.g. ‘fever’, ‘temperature’, ‘sore throat’), which were then converted into flu scores to track the influenza epidemics. Meanwhile, Aramaki et al. (2011) also studied the influenza epidemic by first filtering related tweets and then developing a classifier to extract actual influenza patients. By simply tracking keywords and their combinations on Twitter (e.g. ‘flu’ and ‘shot’), Culotta (2013) estimated influenza rates and alcohol sales volume with high accuracy. Tweets were also used to infer the health status of people, considering their social status, exposure to pollution, interpersonal interactions, and other important lifestyle factors (Sadilek & Kautz, 2013). Abbar et al. (2015) predicted county-wide obesity and diabetes statistics based on a combination of demographic variables and food names mentioned on Twitter. More recently, Chen et al. (2020) tracked tweets to study Twitter responses and reacts to COVID-19-related events and found the trends related to the disease. These studies show the potential to study the opinions and activities of people by analyzing Twitter data, such as tracking keywords and combinations of keywords. However, Twitter data has not been used to specifically study food related discussions in cities, and this chapter fills this gap.

In this chapter, I analyzed Twitter data to explore food related discussions in New York City. Firstly, I explored the popular foods during each meal by analyzing word frequency. High frequency of food-related words implies the popularity of these foods during this time; it also implies people are eating these foods, or at least are showing interest in these foods, since studies have shown that people tweet about what they eat, and Twitter data be used to extract such information to predict health-related statistics, for example, obesity and diabetes rate (Abbar et al., 2015).

One way to explore popular foods using Twitter data is through word frequency analysis. It has been shown that word frequency has a significant meaning in Twitter, and it has been used as one of the measures to detect trending topics (Benhardus & Kalita, 2013) such as the FIFA (International Federation of Association Football) World Cup. By tracking keywords in tweets, Culotta (2013) was able to estimate influenza rates and alcohol sales volume with high accuracy. Word frequency during different time periods will help us understand how popularity of foods change over time.

Furthermore, to explore the spatial characteristics of this data, hot spot maps of popular foods were created. The hot spot maps show statistically significant hot spots of food related tweets. The objective of creating the hot spot maps is to show areas where people are talking about specific foods. Since clusters of tweets in a place about a specific topic can then be used to infer the characteristics of this place (Jenkins et al., 2016), in this case, hot spots represent places of food-related activities, for example, eating or advertising foods. This analysis is done using Hot Spot Analysis which calculates the Getis-Ord statistic (Getis & Ord, 2010) to identify statistically significant spatial clusters.

2.3 Data

The case study area is that of New York City. New York City is the most populated city in the United States, with over 8 million population according to the Census Bureau (2019). It is also a city well known for its cultural, social, and ethnic diversity. One can find food from all over the world within the city. Therefore, it is an ideal location to perform research on food related discussions. The dataset used for this analysis contains 7,091,103 precisely geo-located tweets collected using the GeoSocial Gauge system (Croitoru et al., 2013) with a geographic bounding box from -74.2721 W to -73.6262 W, and 40.4830 N to 40.9325 N. The bounding box includes the five boroughs (Manhattan, Brooklyn, Queens, Bronx, and Staten Island) of New York City and its surrounding areas. The GeoSocial Gauge integrates heterogeneous social media feeds, parses the information from the tweets and stores them into the database.

As all tweets are time-stamped using the Coordinated Universal Time (UTC), the first step was to convert all tweets to Eastern Standard Time (EST). Figure 2.1 shows a heatmap of all the tweets in the study area during 2015, where the light-yellow color represents a higher density of tweets. Downtown and midtown Manhattan have the highest density of tweets, and areas near Manhattan, such as Queens and Brooklyn have a higher density of tweets than other areas. Figure 2.2 shows the number of tweets over a 12-month study period. Less data collected after April since Twitter changed their API, therefore, the numbers of tweets are not even over the months, which is an important element to take into consideration when analyzing change over time.

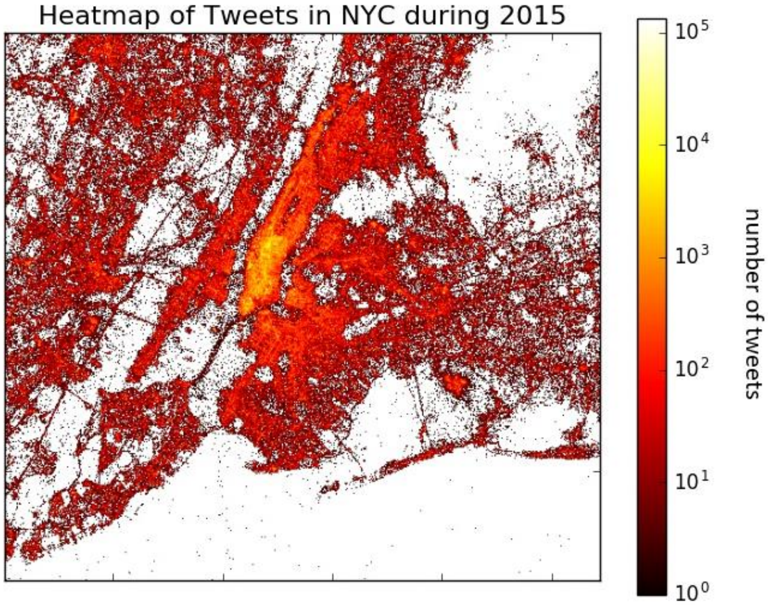


Figure 2.1 Heatmap of tweets in New York City during 2015

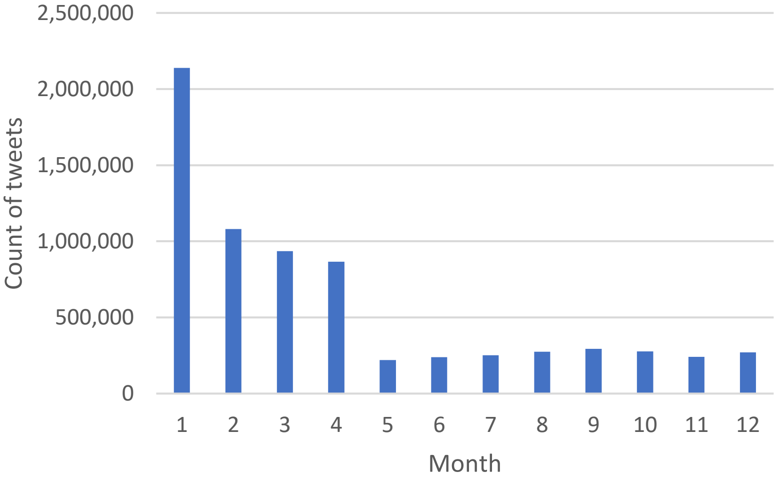


Figure 2.2 Count of tweets over the 12-month study period
(Less geolocated data collected after April since Twitter changed their API)

2.4 Methodology

In order to answer the research questions, several steps are taken, which are shown in Figure 2.3, and they will be explained in detail in this section. Firstly, I selected data of interest (i.e., food -related tweets) using a two-step process. Since I am interested how people's interest in foods change over time, my objective was to extract food related tweets for the three daily meals: breakfast, lunch, and dinner. In the first step, tweets were filtered by these food-related hashtags: "#breakfast", "#lunch", and "#dinner" to find out the time frames for these three daily meals. As specified by Twitter: "A hashtag—written with a # symbol—is used to index keywords or topics on Twitter". Fried et al. (2014) used similar food-related hashtags (such as #dinner, #lunch, #breakfast, #snack) to collect tweets related to meals and used the data to predict latent population characteristics. Filtering using hashtags only was not sufficient for the analysis presented in this research, since on average only 8% of the tweets contain hashtags (Kywe et al., 2012). In my dataset, only 5.46% of the food related data are labeled using "#breakfast", "#lunch", or "#dinner". That means a lot of tweets mentioning foods and meals can't be selected using the hashtags alone. Therefore, in the second step, I used the hashtags to infer the time frames for the three meals. In the second step, I selected a three-hour period of time for each meal when the hashtag was most frequently used, and then I filtered all tweets that mentioned at least one food and are within these time frames. A three-hour window was selected to capture the most frequent hours when people take meals. This was done under the assumption that if someone is talking about food during the breakfast time frame, they are probably tweeting about breakfast food. The list of food-related words was obtained from two online sources (Rodriguez, 2017; BBC Food List, 2017). The first source contained 101 food names that are popular on social media (such as sushi and chicken wings) and the second source contains a more comprehensive set of food names (such as cabbage and hummus). By using these two sources, I created a list of 1166 food names, and that allows for a larger data corpus of food-related tweets to be collected. The list of food names is available on Github (https://github.com/YangZhouCSS/cssphd).

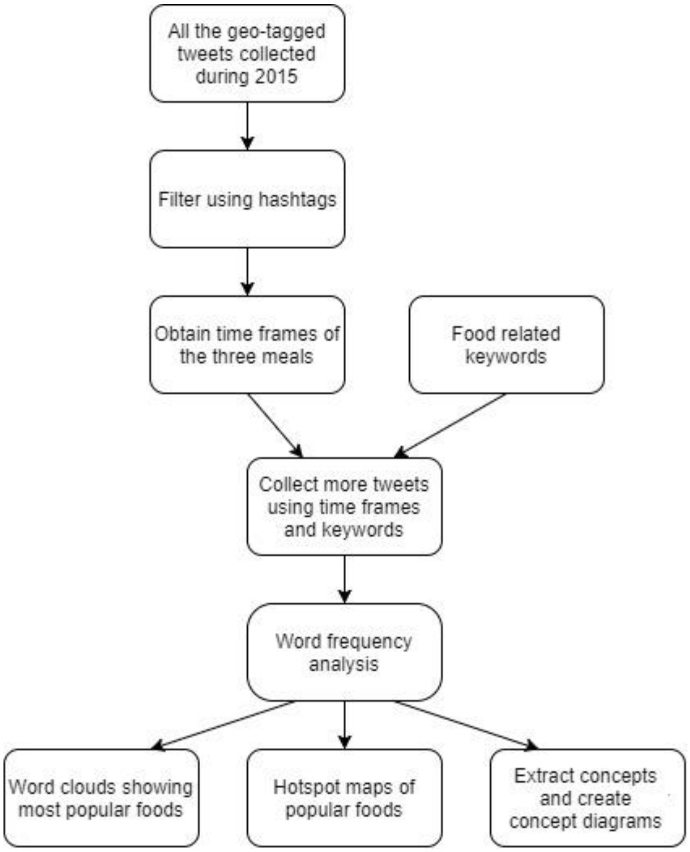

All the geo-tagged tweets collected during 2015
Filter using hashtags
Obtain time frames of the three meals
Food related keywords
Collect more tweets using time frames and keywords
Word frequency analysis
Word clouds showing most popular foods
Hotspot maps of popular foods
Extract concepts and create concept diagrams

Figure 2.3 Flowchart of the procedures

Secondly, I created word clouds to show the frequency of foods in tweets during each mealtime period to understand what people mentioned during breakfast, lunch, and dinner. This is done by analyzing the frequency of foods in tweets of each meal period and visualizing them using word clouds where the size of the words is proportional to their frequency in the corpus. Within this analysis, singular and plural forms of food-related words (i.e., orange and oranges) were treated as the same when counting word frequency. Also, if a word was repeated in a tweet, it was only counted once. Both summary tables and word clouds are provided to show the results.

Thirdly, frequency maps and hot spot maps of popular foods were created to explore the spatial characteristics of this data. The hot spot maps show statistically significant hot spots of food related tweets. This analysis is done using the Spatial Join and Hot Spot Analysis tool in ArcGIS. Spatial Join counts the number of tweets in each census tract. Hot Spot Analysis calculates the Getis-Ord (1992) statistic to identify significant spatial clusters. The Getis-Ord statistic is the z-score produced for each study area by looking at its neighbors given their distance, to determine spatial clusters of high values and low values. Getis-Ord statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} \omega_{i,j} x_j - \frac{\sum_{j=1}^{n} x_j}{n} \sum_{j=1}^{n} \omega_{i,j}}{S \sqrt{\frac{[n \sum_{j=1}^{n} \omega_{i,j}^2 - \left(\sum_{j=1}^{n} \omega_{i,j}\right)]^2}{n-1}}}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\frac{\sum_{j=1}^{n} x_j}{n}\right)^2}$$

x_j is the value for feature j, $\omega_{i,j}$ is the spatial weight between feature i and j, n is the total number of features. For my analysis, x_j is the number of related tweets in the census tracts, and equal weights are assigned to the neighbors. The Getis-Ord statistic returned for each feature in the dataset is a z-score. Hot spots are census tracts with z-scores significantly higher than the mean, while cold spots are census tracts with z-scores significantly lower than the mean. Since there are unequal numbers of tweets in different areas as shown in Figure 2.1, some local hot spots may be neglected since the mean is raised by areas in Manhattan. Therefore, hot spot analysis was performed separately on each borough, but the results were visualized on the same map. For example, if a neighborhood in Staten Island is classified as a hot spot while a neighborhood in Manhattan is classified as a cold spot, it does not necessarily mean that the one in Staten Island has a higher z-score since Manhattan has a higher mean. Maps of word frequency were also created to be compared to the hot spot maps. The frequency maps are also at Census Tract level, and the color of each tract represents the frequency of the keyword.

To provide a more detailed story-telling diagram, concept diagrams were created by extracting the food-related concepts in tweets containing keywords. These diagrams show the food of interest, and four key information about the tweets that mentioned this keyword: location (by boroughs), time (by meals as defined in the first step of this research), other foods that are often mentioned together, and restaurants that were mentioned together. The location information was obtained by the geolocation of the tweets. I identified the borough that each tweet belongs to using their coordinates. Further, organizations (usually restaurants, in this case) were identified by extracting organization names after the @ symbol. Information about meals were obtained using the hashtags (#breakfast, #brunch, #lunch, and #dinner) and added to the diagrams. The diagrams also show the co-occurrence of other foods and the food of interest. These concepts are drawn in bubbles (their sizes are proportional to the frequency they occur in the corpus), and the bubbles are linked to the food of interest. By combining the information of co-occurring foods, locations, time (implied by meals), and organizations (restaurants), it provides more information to understand the food related disscussions in New York City.

2.5 Results

Using the three hashtags about meals in Table 2.1, a total of 8,674 tweets were collected. Although the number of tweets collected is not sufficient to find the frequent food-related words, they provide sufficient information on the time frames of the three daily meals.

Table 2.1 Hashtags used to collect tweets and the number of tweets collected

hashtag	number of tweets
#breakfast	2345
#lunch	3117
#dinner	3212

Figures 2.4 to 2.6 show the numbers of tweets for #breakfast, #lunch, and #dinner. For each meal, a three-hour interval containing the highest number of tweets was picked to represent the time frame when most people discuss this meal. As a result, breakfast is from 8 am to 11 am, lunch is from 12 pm to 3 pm, and dinner is from 7 pm to 10 pm.

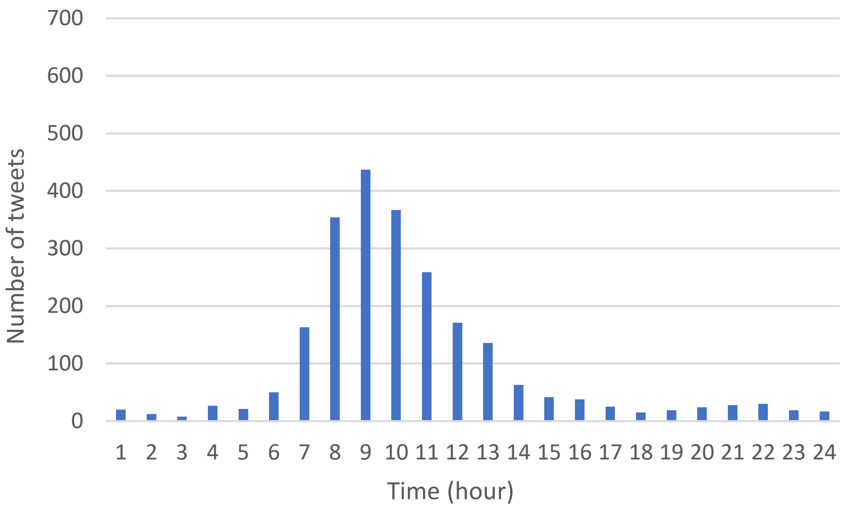


Figure 2.4 Number of tweets containing the hashtag #breakfast collected over a day

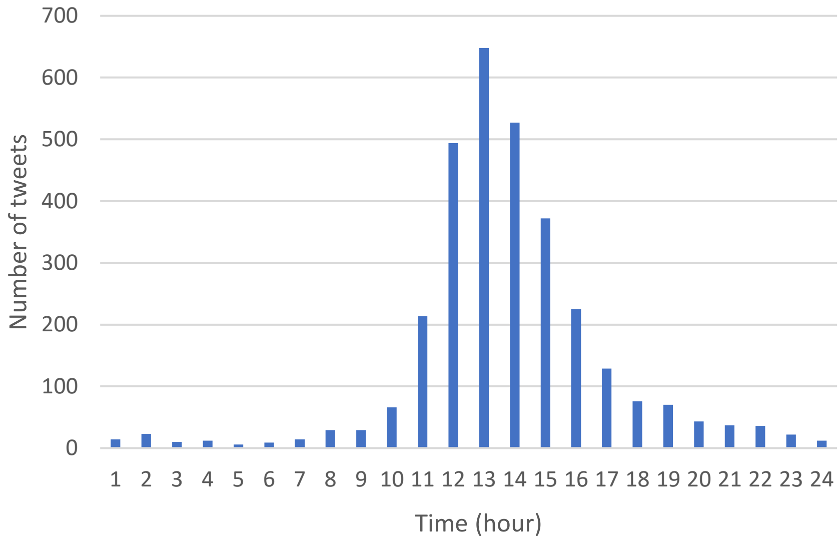


Figure 2.5 Number of tweets containing the hashtag #lunch collected over a day

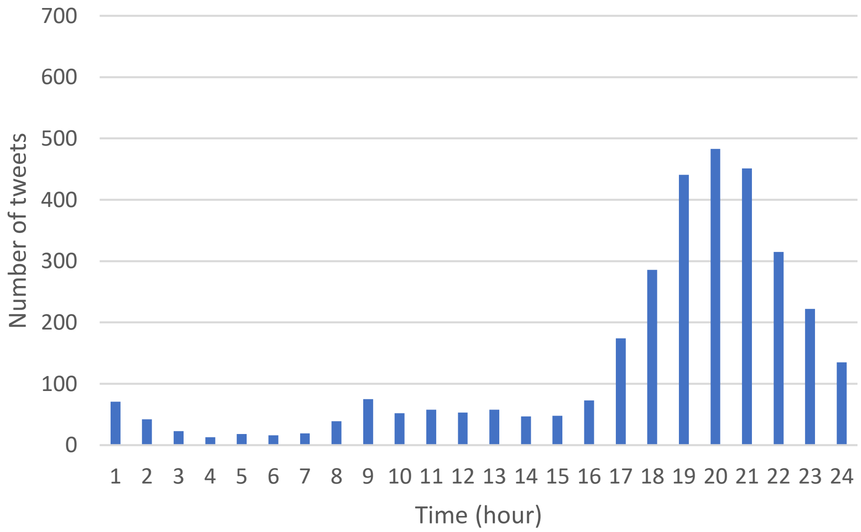


Figure 2.6 Number of tweets containing the hashtag #dinner collected over a day

In order to obtain a larger data corpus about tweets that discussed food, after the time intervals were obtained for the three daily meals, they were used to filter out tweets from the dataset of all the tweets that contained food names. A total of 334,254 tweets about food were obtained using these sources. Based on the time frames, this resulted in 34,878 tweets for breakfast, 56,843 tweets for lunch, and 67,120 tweets for dinner. Figure 2.7 shows the numbers of tweets collected using the time frames. By carrying out this two-step process, 18 times more food related tweets were extracted.

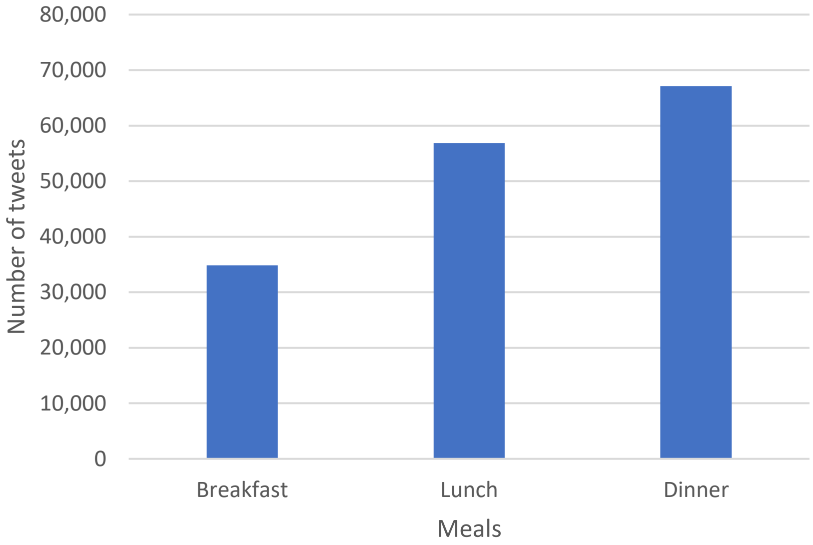


Figure 2.7 Numbers of tweets collected using time frames

By exploring the frequency of food names in the tweets collected, I am able to find popular food words during breakfast, lunch, and dinner. Table 2.2 shows the top 10 most frequent food related words during breakfast, lunch, and dinner time periods. To visualize the results, word clouds where the sizes of words are proportional to their frequency were produced and shown in Figures 2.8 to 2.10). During breakfast, coffee (5632 counts) and tea (1403) were the most popular drinks; chocolate (950), bagels (931), and donuts (727) were the most popular foods mentioned in tweets. During lunch, coffee (4382), beer (2303), tea (1391), and wine (1347) were the most popular drinks; while pizza (2835), chicken (2074), chocolate (1509), and burger (1346) were the most popular foods. During dinner, beer (4024), wine (3073), coffee (1668) were the most popular drinks; while pizza (3759), chicken (2360), chocolate (1741), sushi (1669) were the most popular foods. Strangely, cheese as an ingredient was found popular in all meal periods. The words "apple" and "orange" also occurred in the list of the top words, however, they may refer to other concepts frequently used on social media, as Apple is a company, and orange is a color. By manually reading samples of the tweets, it was discovered that many of the tweets that contained “apple” or “orange” were irrelevant to food. Therefore, these words were removed before carrying out any further analysis.

Table 2.2 Top 10 most frequent words during breakfast, lunch, and dinner time frames

Breakfast	Count	Lunch	Count	Dinner	Count
coffee	5632	coffee	4382	beer	4024
apple	1403	pizza	2835	pizza	3759
orange	1131	beer	2303	wine	3073
cheese	1035	apple	2074	chicken	2360
chocolate	950	chicken	1973	cream	1941
bagel	931	cheese	1751	apple	1880
tea	896	chocolate	1509	cheese	1757
pizza	838	tea	1391	chocolate	1741
beer	734	wine	1347	sushi	1669
donuts	727	burger	1346	coffee	1668

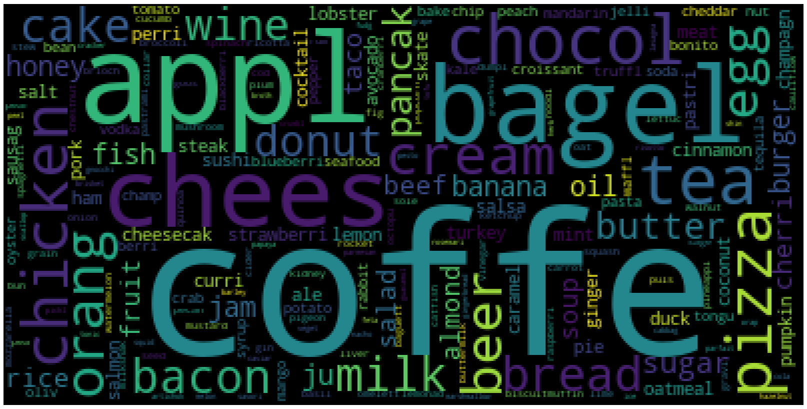


Figure 2.8 Word-cloud of most frequent foods during breakfast

(sizes of words are proportional to their frequency)

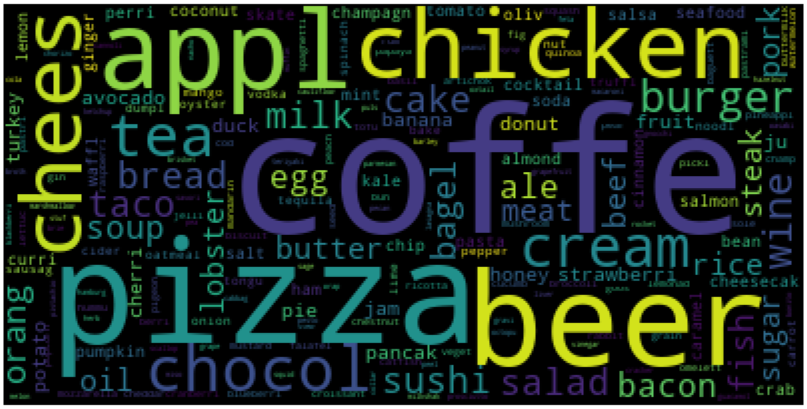


Figure 2.9 Word-cloud of most frequent foods during lunch

(sizes of words are proportional to their frequency)

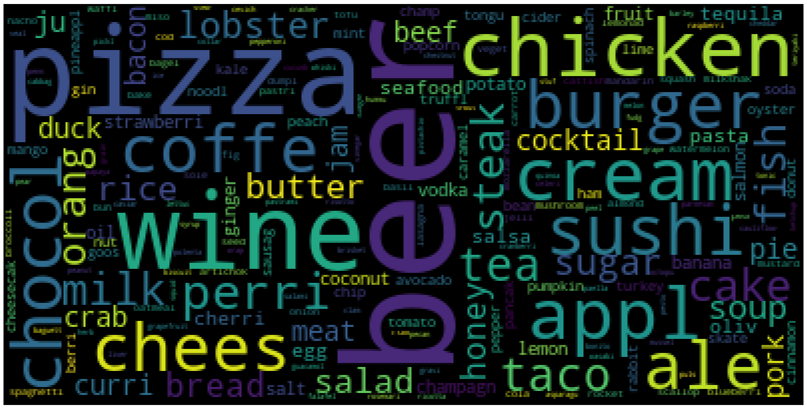


Figure 2.10 Word-cloud of most frequent foods during dinner

(sizes of words are proportional to their frequency)

To investigate the spatial distribution of food-related tweets, frequency maps and hot spot maps were created for two of the most popular words: coffee and pizza. The hot spots maps show where a specific food or drink was most discussed. Figures 2.11 and

2.12, 2.13 and 2.14 show the frequency of tweets and the hot spot maps of coffee and pizza respectively in different areas over a course of a year.

The frequency maps show how often the foods were mentioned in different areas. On the other hand, the hot spot maps were able to show clusters of food-related activities, and infer places where people like to eat, or at least discuss foods. The hot spots analysis was done separately for each borough to uncover local places where foods were popular. Despite the fact that some areas have more tweets in general than others, the maps were able to show hot spots in different areas, even where the density of tweets was comparably low. For example, an area in Staten Island near Midland Beach is a hot spot of coffee with a 99% confidence level as shown in Figure 2.12. John F. Kennedy International Airport is one of the hot spots where coffee and pizza were frequently desired. In the hot spot maps, blue areas represent locations where clusters were not found.

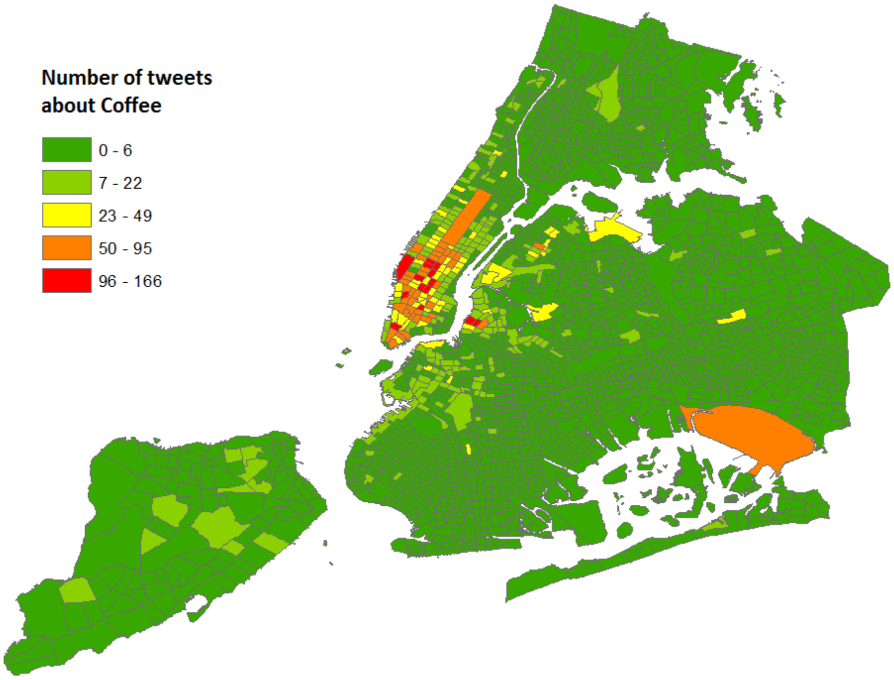


Figure 2.11 Frequency map of coffee tweets
(red = high frequency, green = low frequency)

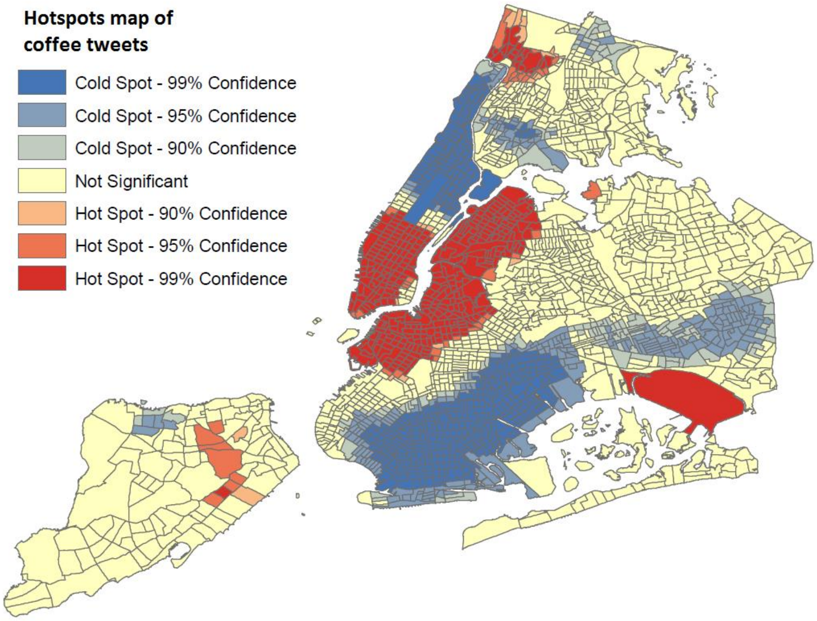

Hotspots map of coffee tweets
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 2.12 Hot spots (red) and cold spots (blue) of coffee tweets

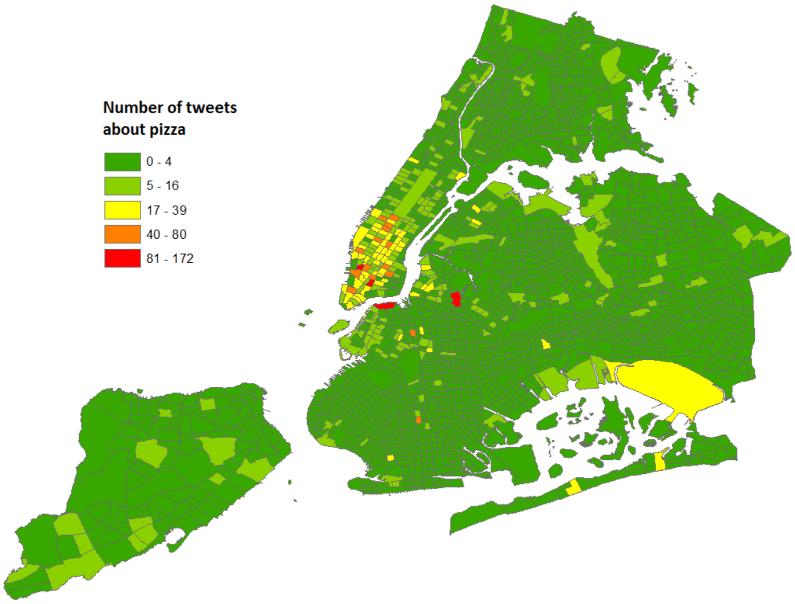
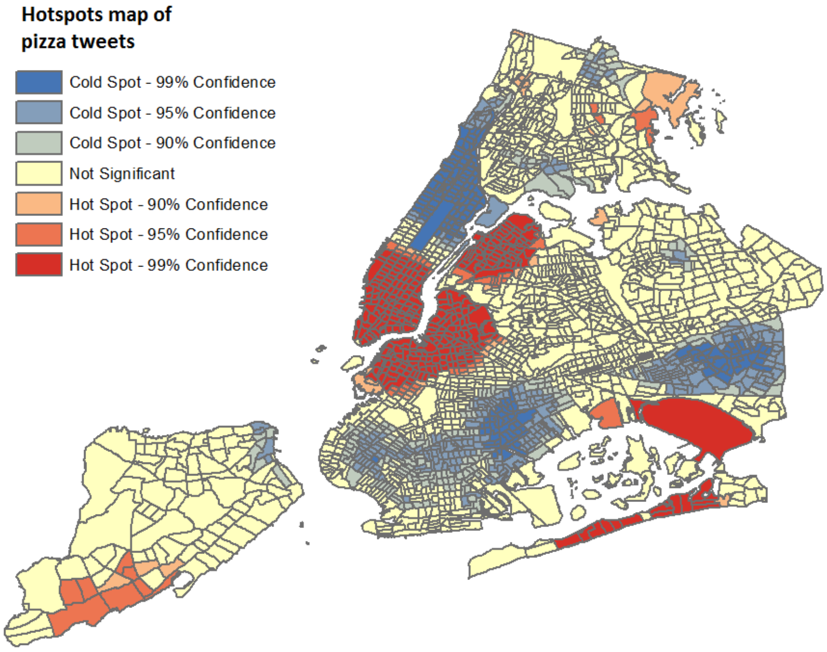


Figure 2.13 Frequency map of pizza tweets
(red = high frequency, green = low frequency)



Hotspots map of pizza tweets
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence

Figure 2.14 Hot spots (red) and cold spots (blue) of pizza tweets

Figure 2.15 shows the food-related words that were often mentioned together with pizza. The keyword, pizza, is placed in the center of the diagram. The four categories of concepts are marked by different colors. Food-related concepts nodes are blue, organization nodes are green, location nodes are brown, and meal nodes are purple. The size of each node corresponds to the frequency of this concept. For example, the diagram shows that artichoke, chicken, and cheese are usually mentioned together with pizza. Pizza is most mentioned in Brooklyn, and it is often consumed for lunch and dinner. The green nodes show the most popular pizza restaurants. Similarly, we can tell stories about coffee and beer using Figure 2.16 and Figure 2.17. More concept diagrams about any food can be produced using this method and the program written.

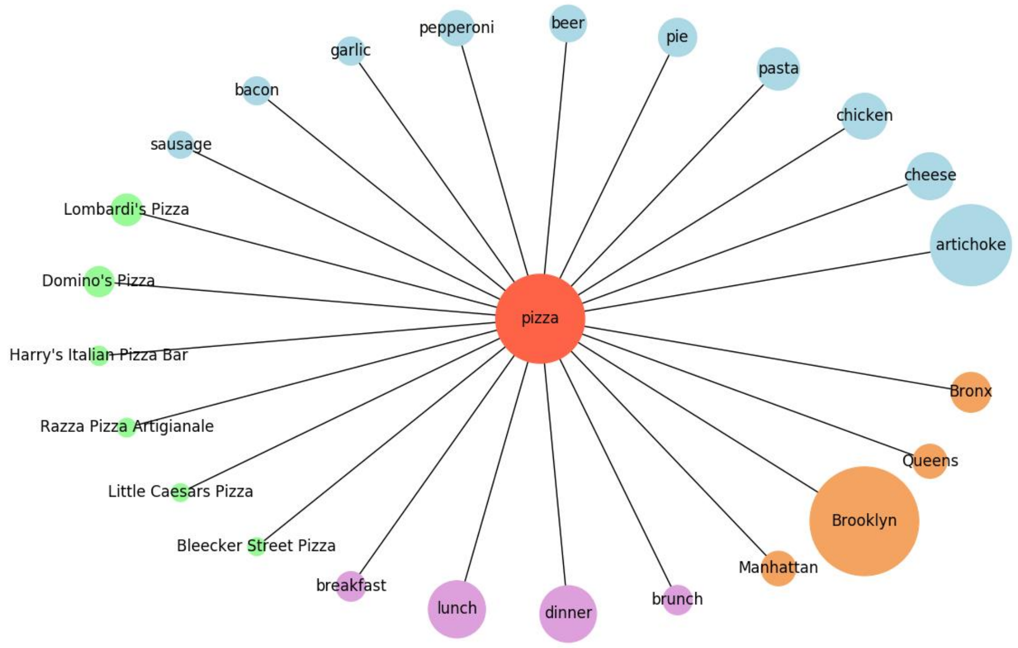


Figure 2.15 Concept diagram surrounding pizza

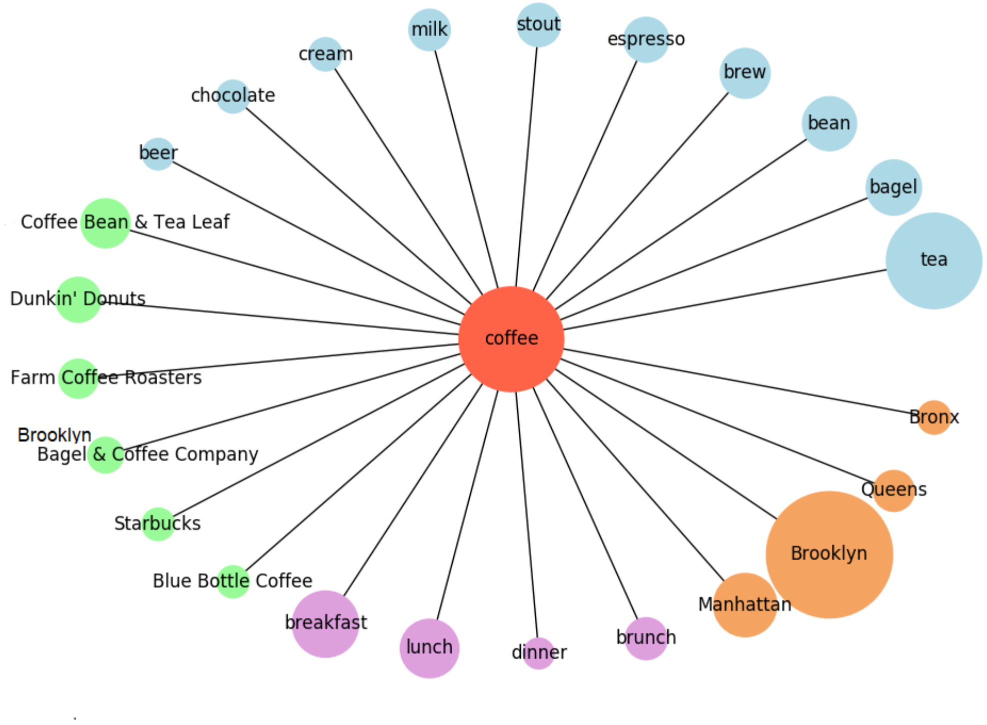


Figure 2.16 Concept diagram surrounding coffee

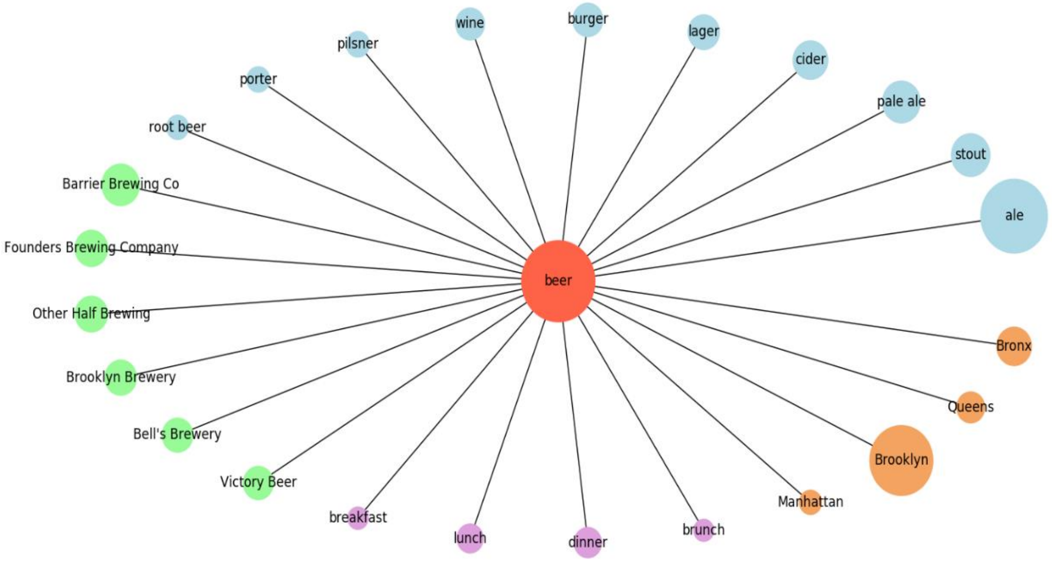


Figure 2.17 Concept diagram surrounding beer

2.6 Conclusion

This chapter studied food related discussions using social media data. Findings in this chapter can be used by restaurant owners to understand what kinds of foods people are interested in during different meals and in different areas, so that they can plan their food offering strategies. They can also use the data to find out competitors for certain foods. For example, restaurant owners could use the word cloud (e.g. Figure 2.10) to find out that demand for wine and sushi is higher in the afternoon, so they can offer more of those food during dinner. Furthermore, restaurant owners could learn from the concept diagram (e.g. Figure 2.16) that people like to mention bagel and coffee together, so they can offer meals that contain these two items; the diagram also shows popular competitors. Through the hot spot maps, restaurant owners could explore where people like to mention those food, which could imply the demand in those areas. Furthermore, the results can complement with results generated by more traditional methods (e.g. surveys and interviews) to provide greater information in near real time related to food in a large area or over extended periods of time.

The four questions raised at the beginning of this chapter are addressed. For the first question, the results show that breakfast is usually from 8 am to 11 am, lunch is usually from 12 pm to 3 pm, and dinner is usually from 7 pm to 10 pm. For the second question, different preferences on foods were found for different meals. It was found that coffee was more popular during the day, while alcoholic beverages were more popular at night. There are a few items that were only popular during a particular time, for example, bagel and donuts were found to be popular only for breakfast; burger was popular only for lunch; sushi was popular only for dinner. Pizza was found to be popular throughout the day, but it was comparably less popular for breakfast. These results and the methods used to find such results could be useful for restaurant owners and scholars interested in food related discussions.

By mapping such information, the third question was addressed. The two frequency maps (i.e. Figures 2.11 and 2.13) show that both coffee and pizza are frequently mentioned in Manhattan, however, the exact locations they were mentioned are different. To list a few findings, while most of the densest tracts about coffee are in Manhattan, two of them are in Brooklyn, across the East River, where there are cinemas, stores, and restaurants. The frequency map of pizza is able to show that pizza is not as popular as coffee in Manhattan, and the map can identify areas in Brooklyn where a lot of people tweeted about pizza. The hot spot maps of coffee and pizza show that most hot spots are located in Manhattan or areas in Brooklyn and Queens cross the East River from Manhattan. Besides, some areas in Bronx, Staten Island, and Long Island are also identified as hot spots. The hot spots for either coffee or pizza were different as shown in the maps, therefore it eliminated the chance that only places with high density of tweets will become hot spots. The maps can be produced given any food-related food name, therefore, this method allows people to explore the spatial distribution of hot spots about any food. For future studies, the hot spot maps could be compared with Census data to measure the difference between areas of different characteristics, such as percentage of residential population or the ambient population of an area.

The fourth question was addressed by the concept diagram, which provides information on the co-occurrence of foods that tend to occur at the same time. In general, foods that are supplemental to one another (e.g. pizza and cheese) and those in the same category (e.g. coffee and tea) are often mentioned together. The results of the co-occurrence analysis, however, do not necessarily show that people are eating both foods when they mention both. By reading samples of the tweets, it was found that foods in the same category are mentioned together in tweets when people debate between them, and when stores publish tweets to advertise both food items. Generally, the co-occurrence of two foods shows people's interest in both items at the same time. Furthermore, the concept diagrams are able to show many aspects of people's discussion on a specific food, including the locations, time periods, and restaurant names.

As such, this information could be useful to understand the dietary habits of New Yorkers, study public health-related problems, and inform restaurants' decisions. While what is presented in this chapter is only a preliminary analysis, such work could be combined with other social media platforms such as Yelp reviews for researchers who wish to study food consumption through user-generated content. For example, Yelp provides reviews and user experiences at specific establishments, while Twitter can show people's current desire and opinion on foods. The question is how do these two platforms and the content within each align.

Nonetheless, there are limitations with this approach that need to be considered and thus present areas for future research. As I found out in the data collected, the number of tweets in different months varies significantly, which will lead to seasonal change in word frequency. As a result, I cannot directly perform word frequency analysis on the data for seasonal patterns due to the limited number of tweets in some months (e.g. more tweets were collected before April). However, if data was collected for a longer period of time, or one supplemented the precisely geolocated tweets used within this study with tweets whose location is derived from the users' profiles or the text of the tweets, these issues could be resolved. Further, the meaning of tweets needs to be further explored. Name Entity Recognition algorithms currently do not work well on foods groups, because it requires building a new knowledge base about foods, and that could be an interesting area for future work. Another problem about understanding the tweets is that although many people tweet about what they eat, a problem discovered when reading samples of the tweets is the fact that not all of the tweets were about eating foods. Some tweets are advertising their foods, some are using another meaning of food-related words (e.g. "I am wearing an orange shirt"). When two foods were mentioned at the same time, it is even more difficult to determine the exact implication of the tweets without more in-depth linguistic analysis. To enhance this study, an area for future work is building a model to predict whether tweets are related to eating or not. By collecting more data and adding analytical models to this study, we can obtain more accurate results. Nevertheless, this study has used a simple and lightweight approach to show how through the analysis of social media data, one can gain insights into what people discuss pertaining to food topics in New York City. The approach undertaken in this chapter could be applied to other cities and therefore provides a means to gain insights into the food related discussions.

Chapter 3 : Linking Social Media and Traffic Data: The Impact of Events on Traffic Flow

The development of social media platforms and the increasing amount of open data sources provided by organizations and government authorities provide many areas of research for studying people living in cities. This chapter combines Twitter data and traffic data published by the local government authorities to study the impact of special events on traffic flow in New York City. Tweets are used to identify events that happened in the city, and then the events are linked to traffic data of the same spatiotemporal dimensions to explore any change in traffic flows during those events. Hashtag analysis and topic modeling are used to explore the themes of the events and categorize the tweets. This chapter concludes with a discussion on patterns in the traffic data, different categories of special events, and their impacts on traffic flows.

3.1 Introduction

Events, like music concerts and baseball games, can attract a lot of traffic to a certain area and significantly impact urban mobility. The consequences are of interest to scholars, city planners, and health departments. For example, as a result of the traffic flow, there could be traffic congestions problems in the areas, which has been a significant problem for cities. Urban congestion has become increasingly costly in terms of time, money, and fuel; the amount of CO_2 emitted is also concerning (Chang et al., 2014). What's more, massive gathering events can lead to the spread of diseases

(McCloskey et al., 2020; Muttalif, et al, 2019). Therefore, the ability to predict the impact of events on traffic flow is very important. Traditionally, people have predicted traffic flows by studying commute patterns using surveys, for example, the Metropolitan Washington Council of Governments publishes Commute Survey Report from the Metropolitan Washington Region every year; the Census (2019) also publishes commuting data, such as commuting patterns and travel time to work. In recent years, people are using cameras to monitor and analyze traffic flow in cities (Shan et al., 2015; Idé et al., 2016). With the development of information technology, a large amount of data are available from social media platforms (e.g. Twitter, Instagram) and government open data sources. This chapter explores the potential of studying the impact of events on traffic flow using these new sources of data.

In this chapter, I will identify events in the tweets by measuring the frequency of tweets in an area. When there is a high frequency of tweets in an area, it implies that people are gathering and discussing what is going on. Furthermore, I will use Topic Modeling to extract topics from the discussion and to understand the content of the events. Then, I will compare the traffic data (including taxi and Metro trips) during the events with normal traffic, to explore the impact of events on traffic. In Section 3.2, I will review related literature, which is the background of my research. Then, I will introduce the datasets in Section 3.3 and the methodology in Section 3.4. In Section 3.5, I will discuss the implications and limitations of this research.

3.2 Related Work

Two types of datasets: social media data and traffic data are compared to find out the impact of events on traffic flow. In this section, I will talk about previous studies about measuring and predicting traffic flow. Then I will talk about previous research on identifying events and extracting topics from social media data.

The first step of my research is to identify the events using Twitter data. User-contributed messages on social media sites such as Twitter data have been used by many scholars to detect events, and the methods have been summarized in papers (Weng et al., 2011; Hasan et al., 2018). Walther and Kaisser (2013) implemented an event detection method by clustering tweets based on their spatial and temporal features. Their method is simple and effective, since their approach is able to detect events that can be verified. In this research, I will use a similar approach to cluster tweets that are spatially and temporally close to each other, and identify an event when the frequency of tweets in the clusters is significantly higher than average. More details will be explained in Section 3.4.

There has been a lot of interest in exploring traffic-related problems using social media data. Traffic data such as traffic congestion and incidents were extracted from Twitter data (Wanichayapong et al., 2011). Their approach filtered tweets with traffic keywords such as "accident" or "congestion" and the location. It is also possible to identify small traffic accidents by training a machine learning model on microblogs data (Schulz et al., 2013). In another study, Gutierrez et al. (2015) built a system with tweet classification, event classification, named entity recognition, and event tracking to identify and track traffic-related events using geolocated tweets. While these methods are effective in identifying traffic congestion or accidents in real-time, they lack the ability to study traffic flows related to an event. On the other hand, some scholars have attempted to predict traffic flow using machine learning models and time-series data. Kriegel et al. (2008) introduced a statistical approach to predict to traffic density in traffic networks using by calculating the likelihood of any given individual in road networks to be located at a certain position and time. Traditional fixed sensory data was combined with Probe Vehicle data obtained from smartphone navigation applications to estimate traffic flow in areas with no stationary sensor coverage (Snowdon et al., 2018; Gkountouna et al., 2020). Deep learning techniques have been used to process time-series data collected by sensors for traffic forecast (Sun et al., 2020; Hu et al., 2020; Guo et al., 2019). Road coverage and contribution patterns for four US metropolitan areas were studied by crowdsourcing volunteered geographic information and volunteered street view imagery (Mahabir et al., 2020). Scholars also studied specifically public transport usage in special events (e.g. Rodrigues et al., 2017; Tempelmeier et al., 2019). With the increasing availability of social media data and government open data sources, we gain a new perspective on exploring and predicting traffic flow, by comparing events mined from events and changes in traffic data.

Another important element of my study is to understand the topics of the events, which can help us answer the question that why types of events attract more traffic. Topic modeling is a method for unsupervised classification of textual data, it is able to group the textual data into topics. Due to its unsupervised nature, we can use it to mine topics that we are not even aware of. Therefore, it is suitable for the task of understanding and summarizing the tweets published during events. Latent Dirichlet Allocation (Blei et al., 2003) is a probabilistic, generative model that represents each topic as a distribution over terms and represents each document as a mixture of the topics. It is used for discovering latent semantic topics in large collections of text documents. Hoffman et al. (2010) developed an online variational Bayes algorithm for LDA (Latent Dirichlet Allocation), which is based on variational inference and stochastic optimization, and it can be fit to large textual datasets. LDA identifies semantically related words and they together form topics that can be interpreted by people. This method has been widely used for extracting topics from tweets (e.g. Akhtar, 2017; Vosoughi et al.,2018; Rufai & Bunce, 2020) and have demonstrated that LDA is fast and effective in extracting topics from documents. In this study, I will develop a LDA topic model to categorize the events into 5 categories, and analyze their relationship with traffic flow.

3.3 Data

There are three datasets used for this study. The first dataset is the same dataset used in Chapter 2, it contains 1,938,982 precisely geo-located tweets collected from May 1st, 2015 to Dec 31st, 2015, with a geographic bounding box from -74.2721 W to -73.6262 W, and 40.4830 N to 40.9325 N. This dataset will be used to identify events in New York City and their topics. Table 3.1 shows an example of a tweet. The tweet includes information on time, location, and a message. The Twitter API gives the timestamp in UTC (Coordinated Universal Time), and they are converted to New York

City’s time zone, EST (Eastern Time Zone), for the analysis. In order to build the topic model, non-English tweets are removed.

Table 3.1 An example of a tweet and associated attributes

Time	ID	Language	Longitude	Latitude	Content
Wed Sep 16 23:37:12 2015	644XX	en	-73.9935	40.7505	I’m at Madison Square Garden - @thegarden for Madonna Rebel Heart Tour in New York

Another dataset used in this chapter is the TLC Trip Record Data (NYC Taxi and Limousine Commission, 2015) that consists of data of all taxi rides during 2015. This dataset provides timestamps and the precise geolocation of the taxi pickup points and drop-off locations. Last but not least, the third dataset is the Metro Turnstile Data (Metropolitan Transportation Authority of New York, 2015) that provides data on the name of the Metro stations, recording time, and turnstile entry and exist counts during the time period. The metro turnstile data was measured every 4 hours, which is a wide time period. Since this analysis is done by the hour, I divided the total number of entries and exists by 4, and evenly allocated them to each hour. While the taxi trips are precisely geo-located, the metro dataset does not have spatial information. In order to geocode the Metro stations, I obtained their coordinates using Google Maps. Each Metro station has multiple entrances and exits, to simplify the problem, all flows are combined into the one with the highest traffic.

All of the traffic datasets are available during the same timeframe as the tweets. Therefore, these three datasets could be used together to study the impact of events on traffic flows. To explore some characteristics of the datasets, I plotted them over months, day of the week, and hours of a day, as shown in Figures 3.1 to 3.3.

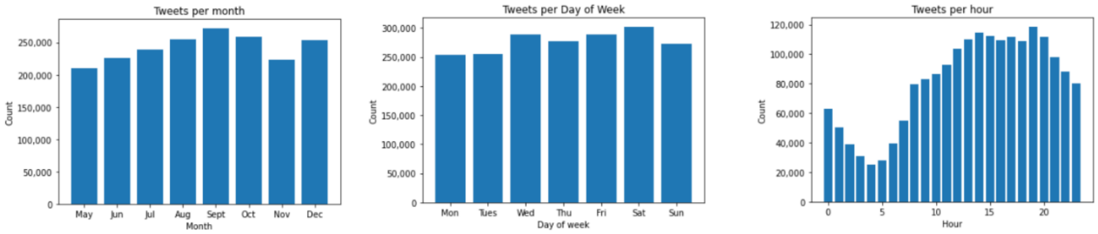


Figure 3.1 The Number of tweets over months, day of the week, and hours of a day

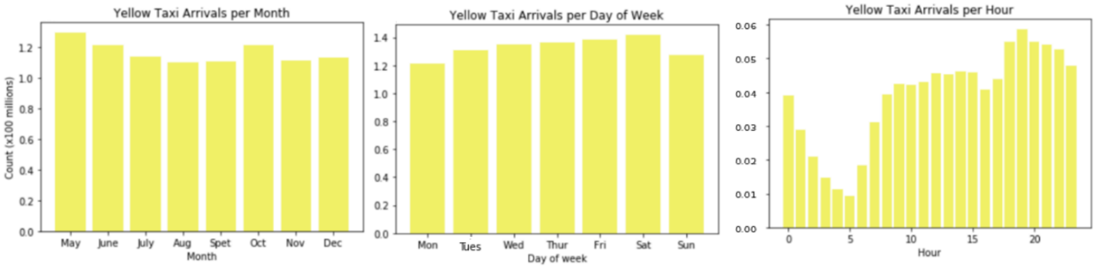


Figure 3.2 The Number of taxi trips over months, day of the week, and hours of a day

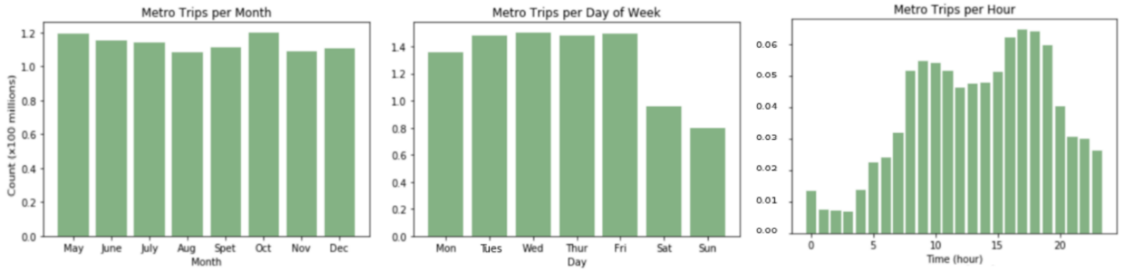


Figure 3.3 The Number of metro trips over months, day of the week, and hours of a day

For all three datasets, there is no significant difference among different months. However, the amount of metro trips is much lower during weekends. The frequency of tweets, metro, and taxi trips vary a lot over the course of a day. Taking these characteristics into consideration, for each event, I will compare event traffic with average traffic at the same hour of the day, in a 21 days window (41 days for weekends), and I will also differentiate weekdays and weekends. As Figure 3.1 shows, there is variance in the number of tweets over months and days of week, and even during the same month, the number for each day can have high variance. Therefore, data are only compared within a 21 days window to control for variance. For example, if an event happened on September 16th at 8 pm, its comparison dataset will be September 6th to September 26th, all weekdays, at 8 pm. This approach eliminates much fluctuation irrelated to the events.

3.4 Methodology

In this section, I will introduce the methodology used to identify the events, measure corresponding traffic flows, and categorize events using topic modeling. Figure

3.4 illustrates the overall process. The three input datasets were cleaned and preprocessed, and then the tweets were used to detect the events and compared with the traffic data. Next, a topic model is developed to explore the topics in the event related tweets.

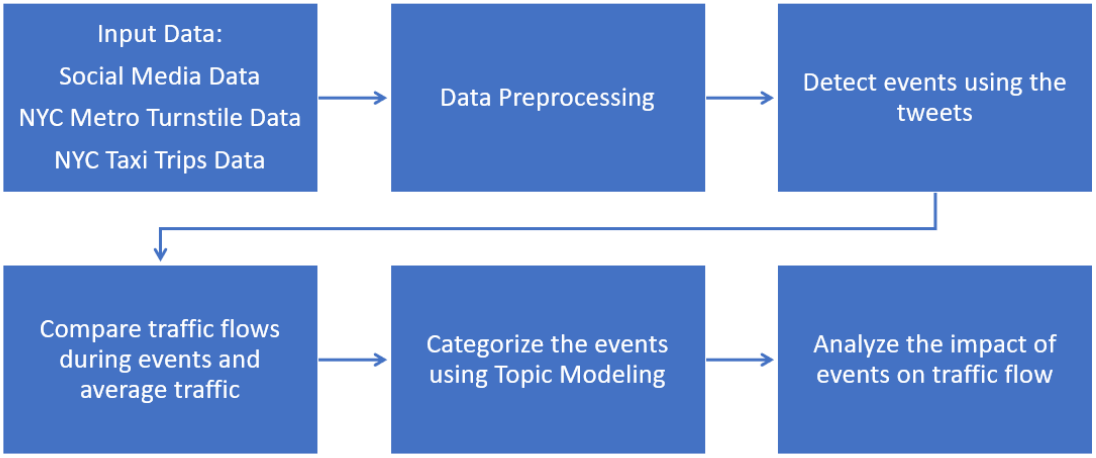
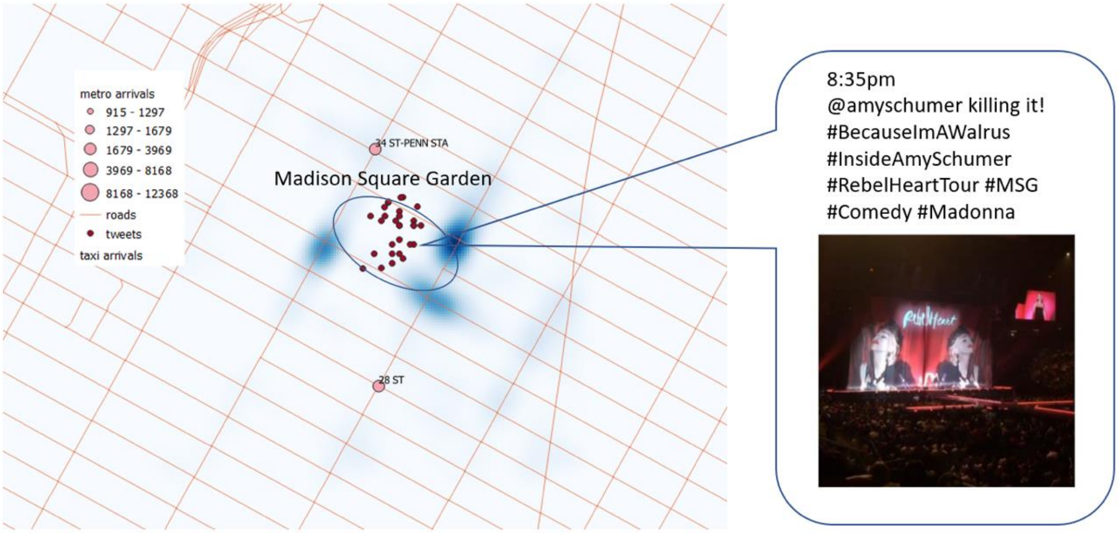

Input Data:
Social Media Data
NYC Metro Turnstile Data
NYC Taxi Trips Data
Data Preprocessing
Detect events using the tweets
Compare traffic flows during events and average traffic
Categorize the events using Topic Modeling
Analyze the impact of events on traffic flow

Figure 3.4 Flow chart of the overall process

3.4.1 Identifying Events

In order to identify events in the tweets, the map was divided into cells that are 200 meters by 200 meters, which is approximately the size of the blocks in New York City. Then, tweets are grouped based on their geographical location and timestamp. That means, tweets in the same cell during the same hour of the day will be put in the same group. This process is performed on every cell and every hour in the dataset. The number of tweets in each group is counted and compared. When the number of tweets in a cell is significantly higher (at least 10 and higher than the mean plus two standard deviations) than the average of the most recent 3 weeks during the same hours, it shows there is an increased number of tweets in that area. I am choosing a 20-day window for comparison to control for variance over time, as shown in Figures 3.2 and 3.3, the number of tweets differs monthly. I identify these observations as the events in the dataset. Events can last for more than one hour, therefore, when the volume of tweets is statistically significantly high for a few consecutive hours, they are grouped into one single event. Figure 3.5 shows a picture of the tweets published at Madison Square Garden from 8 to 9 pm on September 16th, 2015. As the map shows, the tweets form a cluster because they are close to each other spatially and temporally.



8:35pm
@amyschumer killing it! #BecauseImAWalrus #InsideAmySchumer #RebelHeartTour #MSG #Comedy #Madonna

Figure 3.5 Event at Madison Square Garden during 8 to 9 pm on September 16th, 2015

3.4.2 Compare Traffic during Event

Next, both the Metro data and the taxi arrival data are processed into the same format: arrival per hour. Then, for each of the cells on the map, I calculated the average weekend and weekday traffic flows during each hour of the day. When an event is detected, I search for its nearby traffic, including taxi arrivals within 500 meters, and up to 3 Metro stations within 1000 meters. According to a study, the average walking distance to public transport is 573 meters, and the average walking distance to trains is 805 meters (Daniels & Mulley, 2003). People taking taxis will arrive closer to the destinations. Areas with no taxi data or no Metro station with 1000 meters are excluded from the analysis. I compared the traffic flow at the events' beginning (the hour before the event starts plus the hour after the event starts) with average traffic during the same hours in the 3 most recent weeks. Since people usually arrive before the events, I am looking at traffic data an hour before the events and during the first hour of the events. For metro trips, I also control for weekend and weekdays, since they are very different as shown in Figure 3.3. Figure 3.5 also shows the Metro stations and taxi arrivals. The sizes of the stations are proportional to the number of arrivals, and the darkness of the blue color on the map represents the number of taxi arrivals.

3.4.3 Data Preprocessing

The event-related tweets are preprocessed using standard Natural Language Processing (NLP) methods for topic modeling. The data preprocessing includes a few steps: remove @usernames, URLs, stop-words, punctuation, numbers, and special characters. Then, lowercase all words, tokenize the tweets, perform lemmatization. Only

English tweets are used on building the model. Table 3.2 shows an example of the original text (on the left) and processed text (on the right). The preprocessed text is transformed into a matrix of token counts. The columns in the matrix are the words in the entire corpus, and each row is one tweet. To reduce the dimension of the matrix, only words that occurred in more than 10 tweets are included. A TF-IDF transformation is not needed for LDA, because LDA addresses the shortcomings of the TF-IDF model (Blei et al., 2003).

Table 3.2 Tweets pre-processing example

(the original text on the left, processed text on the right)

text	processed text
#RedEye starts now @user	#redeye start
These No No commercials #redeye	commercials #redeye
#InstaSize @ The Metropolitan Museum of Arts	#instasize metropolitan museum arts
Hockey is back. Let's go Devils! - Drinking at 1787 Abbey Single @user	hockey back let go devil drink abbey single
Never forget. 9.11.01. @ Freedom Tower at World Trade Center	never forget freedom tower world trade center
#Recordbroken #Arod #681 @ Yankee Stadium	#recordbroken #arod yankee stadium

3.4.4 Topic Modeling

The matrix is then passed to a LDA model to find n topics in the event-related tweets. Latent Dirichlet Allocation (Blei et al., 2003) is a probabilistic, generative model that represents each topic as a distribution over terms and represents each document as a mixture of the topics. I am using a python library. The model returns two matrix θ and β, θ(i,j) represents the probability of the i-th document containing the j-th topic, and β(i,j) represents the probability of i-th topic containing the j-th word. The number of topics k is a user specified parameter, and I determined it by testing k in range of 3 to 10, and manually selected k = 5 as this number generated the most meaningful topics. The topics are selected such that each category has a theme (e.g., most words are related to fashion shows). I also created word cloud for each topic with word sizes proportional to their frequency in the corpus, to show the most frequent words in each topic. Then, I read the frequent words in each topic and decided their meaning and event category. I used the θ(i,j) to assign a most likely topic to each document (tweet). The scripts used in this methodology and to generate the results are available on Github, see https://github.com/YangZhouCSS/cssphd. The results of topic modeling will be used to study the impact of each category of events on traffic flow.

3.5 Results

Using the method discussed above, 742 events and 15,064 event-related tweets over 2015 are detected in the tweets. Figure 3.6 shows the most frequent hashtags in the event-related tweets. From the chart, we can see that the most frequent hashtags are related to NYC, and then we see some hashtags related to fashion (#nyfw), and then sports (#usopen, #yankees), arts (#art), music, and singers (#paulinagoto, #llevamedespacio).

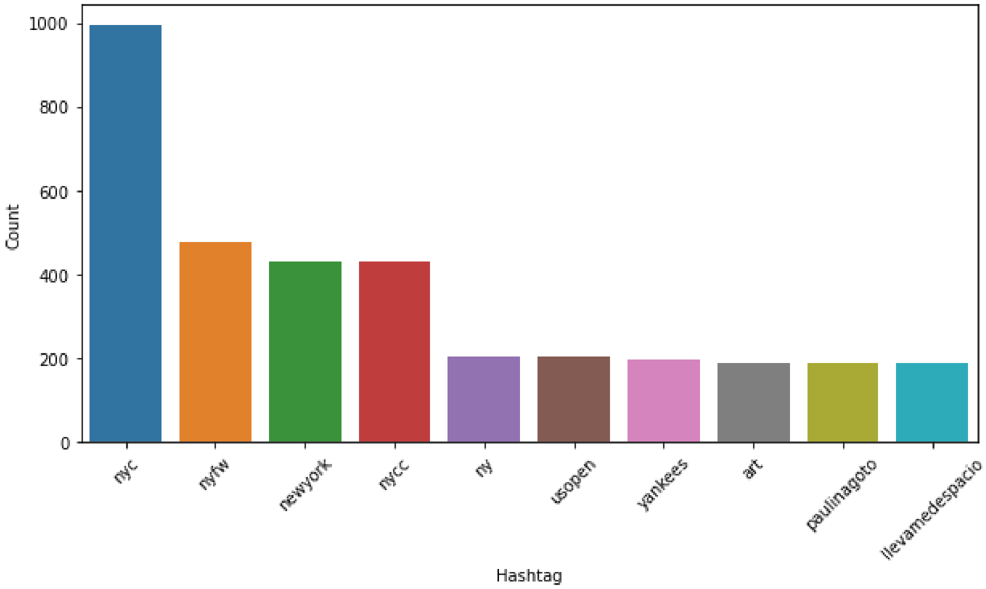
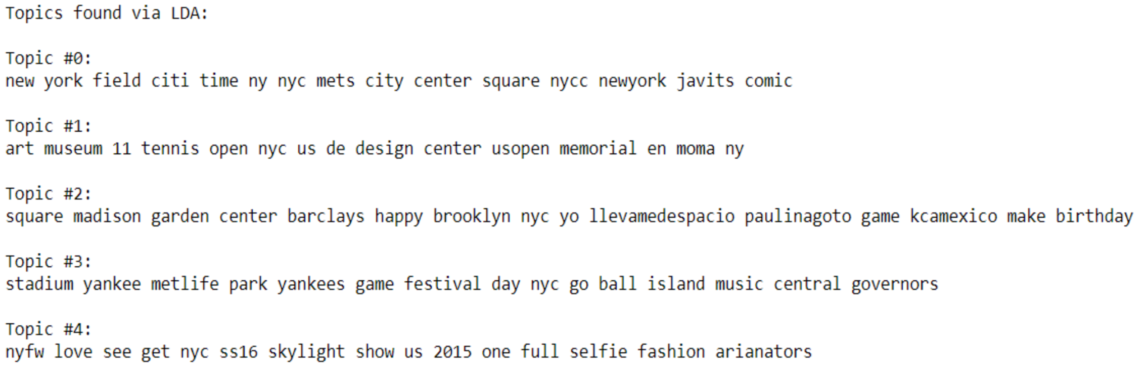


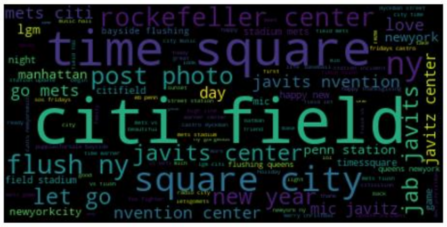
Figure 3.6 Most frequent hashtags in event-related tweets

The number of topics in the tweets was tested so that each topic is meaningful and they match findings from the hashtags. In the end, five topics are found in the event-related tweets. Figure 3.7 shows the five most frequent terms of each topic. For these 5 topics, I interpret their themes as the following: General NYC, Arts and museum, Concerts and other indoor events, Sports and other outdoor events, and Fashion shows. This is also consistent with the findings in exploring hashtag frequency. Then, I predicted the most likely category of each tweet. To visualize the frequent words in each category of tweets, I created word clouds of frequent words in tweets that belong to the topics (Figure 3.8), with the word sizes proportional to their frequency.

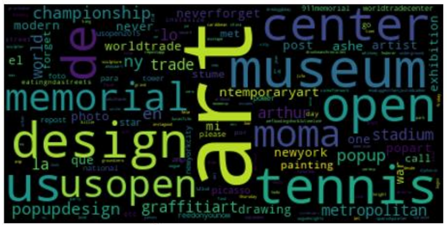


```
Topics found via LDA:

Topic #0:
new york field citi time ny nyc mets city center square nycc newyork javits comic

Topic #1:
art museum 11 tennis open nyc us de design center usopen memorial en moma ny

Topic #2:
square madison garden center barclays happy brooklyn nyc yo llevamedespacio paulinagoto game kcamexico make birthday

Topic #3:
stadium yankee metlife park yankees game festival day nyc go ball island music central governors

Topic #4:
nyfw love see get nyc ss16 skylight show us 2015 one full selfie fashion arianators
```

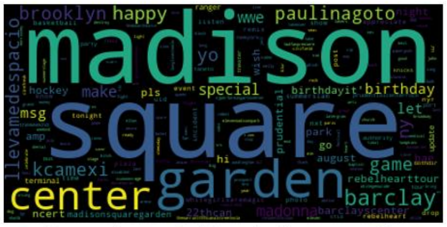
Figure 3.7 Five topics found in the event-related tweets



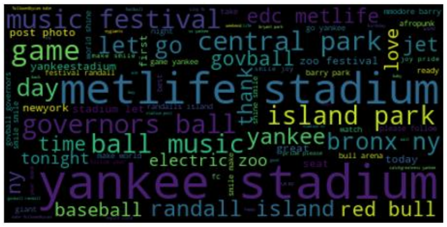
General NYC



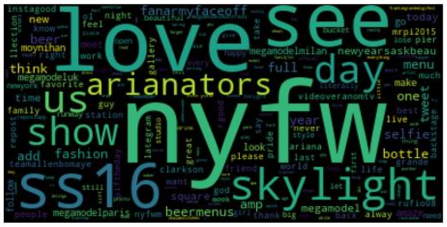
Arts, museum



Concerts and other indoor events



Sports and other outdoor events



Fashion shows

Figure 3.8 Word clouds showing frequent words in each topic

The change in traffic by event category are compared and shown in Table 3.3. Out of the 742 events, 198 events belonged to General NYC, 92 events belong to Arts and museum, 133 events belonged to Concerts and other indoor events, 169 events belonged to Sports and other outdoor events, and 150 events belonged to Fashion shows. The table showed the statistics (mean, median, 10th and 90th percentiles) of change in traffic flow during the events, for Metro and Taxi respectively. A positive percentage represented an increase in traffic, and a negative percentage represented a decrease in traffic. For example, during Sports and other outdoor events, Metro arrivals nearby increased by 44.96%, while taxi arrivals nearby increased by 139.57% on average. Figure 3.9 shows the distributions in graphs.

These results show that all types of events have an impact on traffic flow, both Metro and taxi traffic increase during events. Out of the five categories, Sports and other outdoor events have the highest influence on traffic flow, especially taxi trips. This is reasonable because these activities usually attract a large number of people. The results also show that events generally have a higher influence on taxi trips than Metro trips.

Furthermore, this study compares the change in traffic flow by location. I select 5 frequent locations that appeared in the tweets: Citi Field, Madison Square Garden, Museum, Times Square, and Yankee Stadium. For this task, I filter out event-related tweets that mentioned these locations and explore their influence on traffic. The results can be found in Table 3.4. Figure 3.10 shows the distributions in graphs. Yankee Stadium and Citi Field are both outdoor facilities for sports games and competitions. When the games are held, a large number of people come to these locations by public transportation or taxi, as we can see, they have a large impact on traffic flow. On the other hand, Madison Square Garden usually holds concerts and indoor activities, which are much smaller than sports games in terms of people attracted to the area during the event period. For the same reason, events at museums have a smaller impact on traffic flow compared to outdoor sports games. Interestingly, events at Times Square do not seem to attract much traffic. There are a few explanations. It could be due to the fact that there are so many different Metro stations nearby, people could be arriving at different locations and walking to the event area. The area is also very congested, so people taking taxis may be dropped off further away. Besides, after exploring the events at Times Square, I find that some of them are related to the new year countdown event, and during that time, the street is blocked and Metro stations are closed in this area, therefore, that may explain the decrease in Metro and Taxi arrivals.

Table 3.3 Change in traffic flow by event category

	Metro flow change				Taxi flow change				Count of events
Category	mean	median	p10	p90	mean	median	p10	p90	
arts	7.76%	1.81%	-20.34%	44.72%	13.53%	3.81%	-40.03%	78.30%	92
concerts	8.04%	2.41%	-8.39%	24.36%	22.54%	9.15%	-37.17%	59.24%	133
fashion show	11.24%	0.80%	-10.35%	27.71%	20.85%	7.52%	-30.07%	69.62%	150
sports	44.96%	0.44%	-6.65%	183.89%	139.57%	41.07%	-28.09%	433.39%	169
General NYC	24.12%	2.11%	-13.28%	108.28%	42.31%	8.82%	-37.10%	123.04%	198

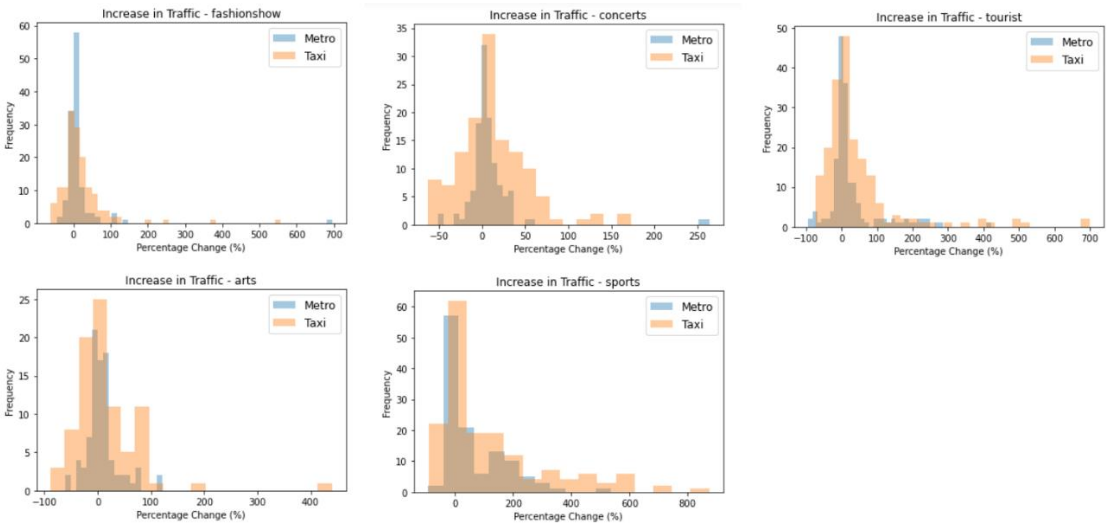


Figure 3.9 Distribution of the change in traffic by categories

Table 3.4 Change in traffic flow by location

	Metro flow change				Taxi flow change				Count of events
Location	**mean**	**median**	**p10**	**p90**	**mean**	**median**	**p10**	**p90**	
Citi field	111.74%	101.37%	8.96%	236.35%	148.17%	72.33%	-25.17%	454.66%	46
Madison square garden	0.94%	1.44%	-11.95%	14.83%	6.43%	7.55%	-17.80%	40.07%	48
Museum	10.36%	4.97%	-16.25%	36.15%	5.38%	-1.08%	-45.48%	46.56%	40
Times Square	-1.51%	2.31%	-10.60%	9.41%	-2.32%	-1.96%	-49.24%	37.32%	46
Yankee Stadium	120.65%	134.08%	4.99%	241.92%	242.29%	168.12%	24.34%	553.45%	56

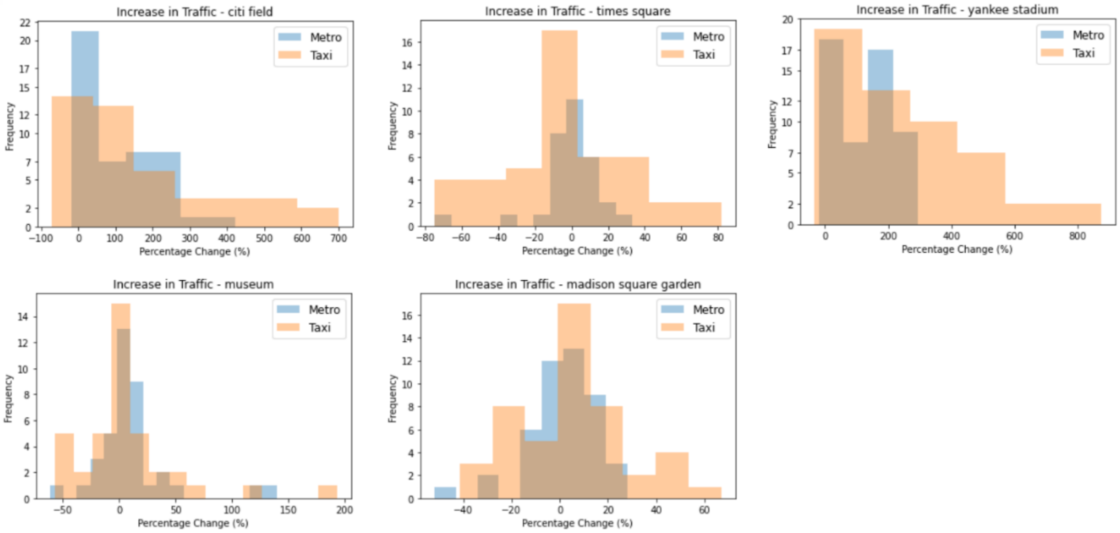


Figure 3.10 Distribution of the change in traffic by locations

3.6 Conclusion

Nowadays, we are able to gain a large amount of data from social media platforms and government open data sources. These resources provide new opportunities for research about social science problems using computational methods. My study shows that it is possible to study the influence of events on traffic flow by analyzing social media data and traffic data from government sources. This approach utilized open data sources that have been collected, and it supplements traditional methods such as surveys by providing a low-cost method to get insights into events and traffic flow. Furthermore, using hashtag analysis and topic modeling, I summarized the themes of the events, and through comparison with the traffic data, I am able to gain insight into their influence on traffic flow. I can also study the relationship between the locations of the events and the change in traffic.

This study is able to find several patterns in the data. Firstly, the Metro is more frequently used during weekdays than weekends, while taxi trips are more evenly distributed throughout the week. For Metro trips, we observe peak hours around 8 am and 5 pm; for taxi trips, peak hours are around 7 pm. It is very likely that people tend to utilize the Metro system for work, and they use taxis more often for other purposes, for example, leisure and entertainment. Therefore, the increase in taxi trips due to events could be higher than the increase in Metro trips due to events. In my study, I am comparing average Metro traffic on weekdays and weekends separately to avoid bias. Secondly, my study identifies five topics from the event-related tweets: General NYC, Arts and museum, Concerts and other indoor events, Sports and other outdoor events, and Fashion shows. These topics are observed through both hashtag analysis and topic modeling. Thirdly, by comparing the events with the corresponding traffic flow (which is Metro and taxi arrivals one hour before the event and during the event), the results show an increase in traffic flow during all types of events. According to the results, outdoor sports games have the largest impact on traffic flow, especially taxi arrivals. Outdoor events tend to attract more people. Yankee Stadium has a capacity of 54,251 people and Citi Field has a capacity of 41,922 people. When large sports games are held there, a lot of people will be gathering there to watch the games, which explains the high impact on traffic. When I look at the change in traffic flow by location, it shows the same patterns, that locations for outdoor sports games and activities tend to attract a lot more traffic. It is also worth mentioning that stadiums are built in areas with lower population density, while music halls, museums, and Times Square are located in more dense areas. In areas with high density, a lot of things are going on and the traffic flow is almost always high, therefore, an event is going to have a smaller impact on traffic flow in terms of the percentage change. While this chapter studies old events in 2015, the methods can be used to study streaming social media data to find real-time events, especially unscheduled events, and help people with traffic planning. This will require the use of the Twitter API to collect streaming data and identify areas that has significantly high number of tweets as compared to statistics of previous days in this area. Then, the topic model can be run to identify the type of event and estimate the impact on traffic.

This study also has a few shortcomings. To begin with, there is bias in social media data for a few reasons. First, there are demographic bias and representation issues in social media. Social media data has significant biases because of demographic differences in people who use the platforms, and what they choose to share (Cesare et al., 2019). Longley et al. (2015) studied the geo-temporal demographics of Twitter usage and discovered bias in ethnicity, age groups, and gender. They found out that Twitter is highly used by teenagers and there are more male users. As a result, the events I detected from the tweets could be biased towards these users, and neglect events popular among other groups of people who are not on Twitter. Second, when using an API to obtain data, due to the way a social media site makes its data available, we may not receive full data (Morstatter and Liu, 2017). This could also cause issues in missing some events. However, I do believe that the results I got from the data cover different types of events, and while they are certainly not all the events that happened in NYC during the study time, they do show the potential to explore traffic-related problems using social media data, and the results present some patterns in what types of events attract more traffic.

Furthermore, this study did not account for Uber and Lyft, which are alternatives to taxis. My study uses 2015 data, while ride-sharing (e.g. Uber and Lyft) was not as popular at that time, as ride-sharing apps are more popular now, it becomes more important to combine those data in the future. Another problem with the data source is that the Metro arrivals were recorded every four hours as turnstile entries and exist. I divided the arrivals by four and assumed uniform distribution during the four hours, because no further details are available. This problem could cause bias in the estimated Metro traffic flow change and cause the percentage to be lower than the true value. If we are going to obtain more detailed data from the NYC government in the future, it is possible to get more accurate estimates.

Another limitation is by using statistically significant increase in tweets to identify events, we are still getting 2.1% of the events as false positives. Though it does not have large impact on the conclusions of this chapter. Even with these limitations and potential areas of further work, the research presented demonstrates a novel approach to study events and their impact on traffic flow by combining social media data with traffic data from government sources.

Chapter 4 : Simulate the Impact of Events on Traffic Flow: An Agent-based Modeling Approach

This chapter continues the study introduced in Chapter 3 by using the results discovered in tweets and traffic data to develop an Agent-based Model (ABM) to simulate the impact of events on traffic flow. The model mimics the area around Madison Square Garden in New York City and measures traffic speed as well as the number of stopped vehicles to measure traffic congestion caused by events. Three base cases of the low, medium and high regular traffic volumes are built to measure these variables when no event is present. Then, nine additional scenarios are built with small, medium, and large events combined with different regular traffic volumes. By comparing the nine scenarios with the base cases, the model draws meaningful conclusions on the impact of special events on traffic flow under different situations. It also provides a stage to further study the impact of events on traffic flow in other areas and situations.

4.1 Introduction

Events in cities can significantly impact urban mobility, and the ability to understand and predict the impact is very important for residents, travelers, city planners, and traffic engineers. Studies have shown that there are two types of congestion: recurring and non-recurring (e.g. Dowling et al., 2004; Bremmer et al., 2004), and non-recurring congestion is caused by unpredictable incidents such as traffic accidents, weather conditions, special events, etc. In this research, I am exploring congestion caused by special events as studied in Chapter 3. I studied the impact of events on traffic flow using social media data as well as traffic data, and I found out that special events, especially large outdoor events, do have a significant impact on traffic flow. In this chapter, I would like to utilize some of the results from the last chapter to build an Agent-based Model (ABM) to simulate the impact of events on traffic flow under a few different scenarios. I am interested in studying the impact of events on traffic flow based on different characteristics of the events and the normal traffic in the area.

4.2 Related Work

Scholars have studied public transport usage in special events using spatial data and statistical prediction methods (Rodrigues et al., 2017; Tempelmeier et al., 2019). Agent-based modeling is another approach to explore this problem. Agent-based modeling is able to capture the geometry and network connection of roads and simulate human mobility behaviors between people as well as the environment they interact in, it is suitable for studying traffic problems and can explore the problems from a different perspective besides statistical problems.

Agent-based modeling has been widely used to study traffic problems (Bazzan & Klüg, 2014). In 2003, a large-scale agent-based traffic microsimulation was developed by Cetin et al. to explore traffic dynamics. The MATSim-T model was an ABM built by Balmer et al. (2009) to simulate traffic demand in a large area using an activity-based approach. In another study, Auld and Mohammadian (2012) took a different approach in simulating travel demand and developed an ABM based on the dynamic decision-making of agents. The influence of social networks on destination choice was simulated in an agent-based traffic demand model (Horni et al., 2011). While Yamashita & Kurumatani (2009) designed a collaborative navigation system for cars, in which they can communicate and share route information to improve traffic efficiency for each vehicle and the whole system, do Amarante and Bazzan (2012) built an ABM to investigate the benefits of drivers using communication and replanning routes. ABMs on traffic flow specifically, study traffic phenomena as a result of some agent behaviors or change in the environment. One of the earliest agent-oriented traffic simulation models was built in 1997 (Burmeister et al.). While Doniec et al. (2008) built a traffic model that emphasizes the behavior rules for the agents while moving through intersections, Luo and Boloni (2012) model driving strategies with lane changes on highways. An ABM was developed by Ben-Dor et al. (2018) to study the impact of dedicated bus lanes on urban traffic congestion. ABMs were also built to study traffic flow during emergency evacuations. For example, Wang et al. (2016) present an agent-based tsunami evacuation simulation model to investigate different scenarios. West et al. (2019) model the behavior of drivers sharing the road with autonomous systems in evacuation situations. Zhang et al. (2020) develop an ABM model to optimize the dynamic bus stop-skipping and holding strategies. Scholars also attempted to explore human mobility using ABM. Batty et al. (2003) developed a pedestrian model to explore mobility in carnivals and street parades. An agent-based simulation framework was developed to simulate human mobility (Kim et al., 2019) and the model was further developed to not only capture the location of people over time, but also their interactions via social networks (Kim et al., 2020).

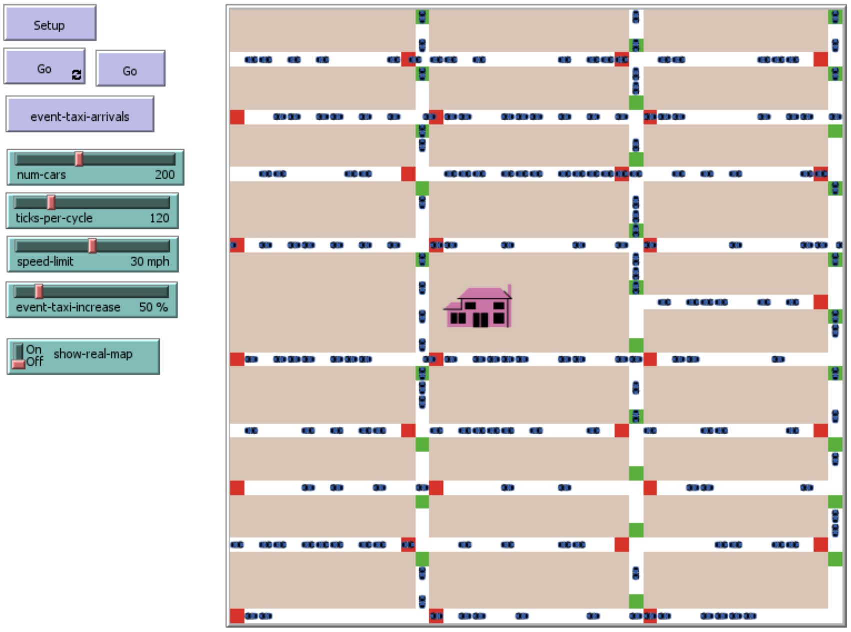
All the works discussed above have shown the potential to use ABM to explore traffic problems, including the impact of behaviors or change of environment on traffic flow. A lot of the models focus on the decision-making of agents under different circumstances, and how their individual behaviors and their interactions lead to change in the system. However, there is a lack of agent-based models that focus on simulating the influence of non-disaster events on traffic flow using findings from real-world data. In this chapter, I will develop an ABM based on the findings from the last chapter. The study area will be an area in New York City where special events are held, such as Madison Square Garden. The model is meant to be a simple model for proof of concept, and can be modified to study different areas and scenarios.

4.3 Conceptual Model

The Environment. This ABM simulates traffic flow on the streets of New York City, and the influence of special events on traffic flow. The streets in the model are designed to mimic the streets in New York City, as shown in Figure 4.1. The standard block in Manhattan is about 264 ft by 900 ft (274 m by 80 m), this data is obtained by measuring the distances between blocks corners on Google Map. Therefore, the blocks in my model will be 4 cells by 13 cells, with each cell equal to 20 m by 20 m. As shown in Figure 4.2, you can also turn on the real map (captured on OpenStreetMap) of the study area to see that the streets do approximately match with the real map. Besides, the directions of each street in the model also match the real map. All the streets are one-way streets. There are usually multiple lanes in reality, but in this model, they are single lanes, because this is a conceptual model in which we are studying the percentage of speed influence, and change in the traffic flow on one lane can represent the overall traffic flow. The purple house in Figure 4.1 represents Madison Square Garden, and it will be the destination of cars coming for the event. However, the use of the model should not be limited to the area around Madison Square Garden, as it could be modified to study other areas and situations.

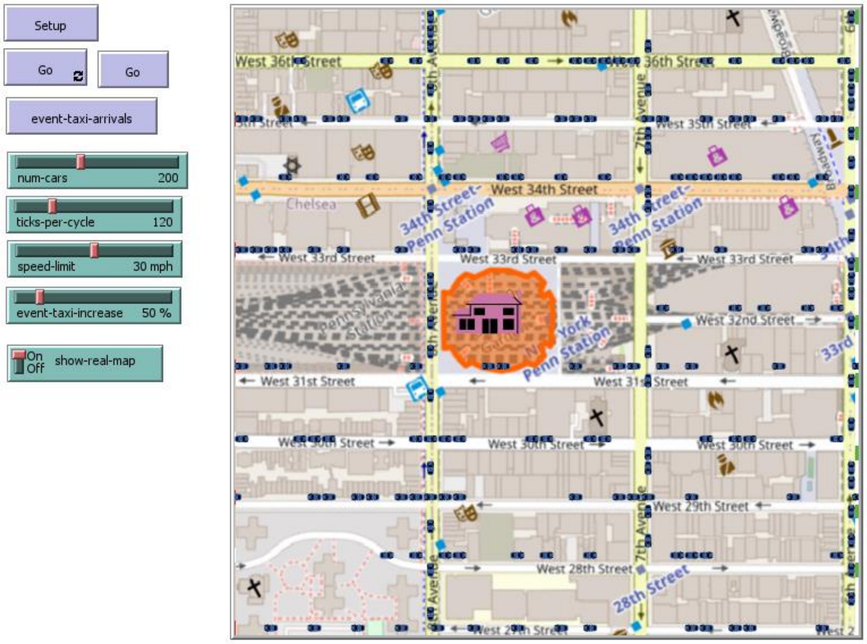
The Agents. The agents in the model will be cars, and there are two types of them: normal cars (including all types of vehicles) and event-related cars. Normal cars represent the normal traffic in the area. Event-related cars are added to the system when an event happens. The number of added event-related cars is determined by the type and size of the event, which can be adjusted by a slider. According to Chapter 3, event traffic usually arrive during the hour before the event starts. Therefore, if there is a special event, event cars will be coming to the map during a one-hour window, following a uniform distribution.

Behavior rules. The speed limit in New York City is 30 mph, so I will use the same speed limit in the model. Cars will speed up until they meet the speed limit, or if there is a vehicle in front of them with a slower speed, in that case, they will slow down. Event-related cars will try to reach their destination (Madison Square Garden). At each crossing, event cars consider making a left or right turn, and if they get closer to their destination after turning, they will make the turn (Figure 4.3). Normal cars do not have destinations, they represent the normal traffic that is always there, and their number can be adjusted as well. The model is built in Netlogo based on the Traffic Grid model (Wilensky, 2003) and available on Github: https://github.com/YangZhouCSS/cssphd.



num-cars 200
ticks-per-cycle 120
speed-limit 30 mph
event-taxi-increase 50 %

Figure 4.1 Interface the Agent-based Model. Regular cars are blue, event-related cars are yellow. Green and red cells represent green and red lights.



num-cars 200
ticks-per-cycle 120
speed-limit 30 mph
event-taxi-increase 50 %

Figure 4.2 Interface of the Agent-based Model, agents displayed on top of the map of Madison Square Garden (Map source: OpenStreetMap)

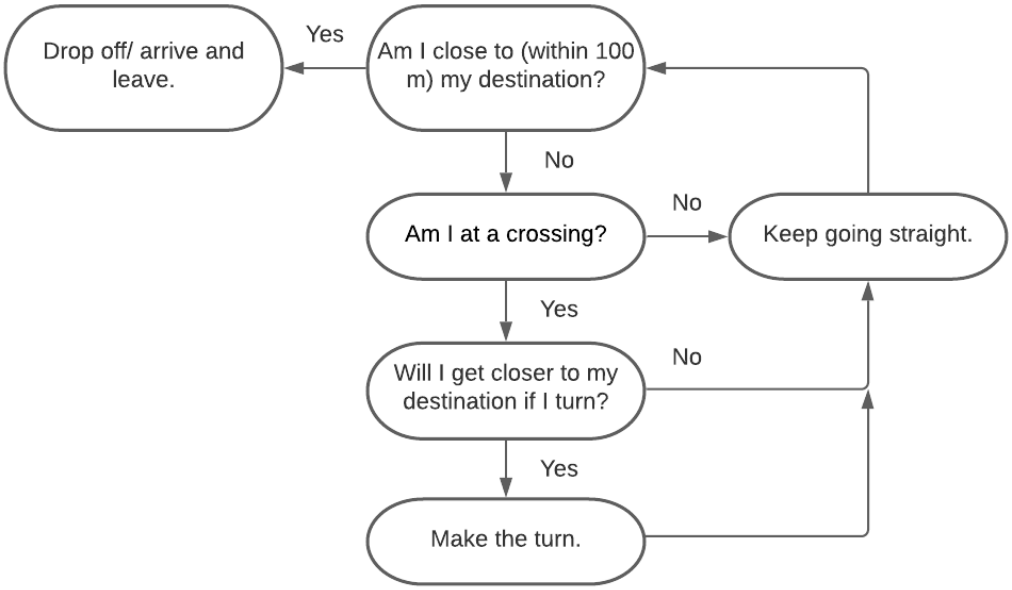

Drop off/ arrive and leave.
Yes
Am I close to (within 100 m) my destination?
No
Am I at a crossing?
No
Keep going straight.
Yes
Will I get closer to my destination if I turn?
No
Yes
Make the turn.

Figure 4.3 Flow chart that shows the decision making of event-related cars

As shown in Table 4.1, I designed 12 scenarios to study the influence of events of various sizes under three different traffic conditions (low, medium, high volume). The number of regular cars represents the regular traffic in this area, this number can vary during different periods, for example, it's higher during peak hours. Scenario 1 to 3 will be the base case when there is no event. Scenario 1 represents the case when there is low regular traffic, Scenario 2 represents normal hours and medium traffic, and Scenario 3 represents peak hours with high traffic. When there is an event, a number of event cars (cars arriving due to the event) will be arriving at this area and moving towards their destination. In the last chapter, I analyzed the percentage increase in taxis. In this model, the event cars include both taxis and non-taxis arriving due to the events. We can expect the number of non-taxi vehicles to increase during events too, because there will be people driving themselves or their friends to the events, and they should be considered as well. My study in the last chapter shows that special events can attract more taxis because of the events. In this ABM, I have three types of events: small events, medium events, and large events, and they represent the art and museum events, General NYC, and sports and outdoor events, respectively. The impact on taxis trips is obtained from Table 3.3 and used to build the ABM. Specifically, small events, such as such as special Museum exhibitions, attract 13.5% more traffic to the area. Medium-size activities include special events held at Times Squares and Rockefeller center, music tours and fashion shows. An example would be the Rebel Heart Tour, a music tour that lasts for a few days and attracted a large number of fans. In this ABM, I put this type of events into the medium events category, and they will attract 42.3% more traffic. Last but not least, there are large events, mainly popular sports games held at outdoor stadiums and fields, they could attract 139.6% more taxis to the area. Examples of large events are the baseball games at Yankee Stadium.

Table 4.1 Scenarios to test in the agent-based model

Scenario	Number of normal cars	Event type	Event cars (as a percentage of normal cars)
1	100	None	0
2	200	None	0
3	300	None	0
4	100	Small	13.5%
5	200	Small	13.5%
6	300	Small	13.5%
7	100	Medium	42.3%
8	200	Medium	42.3%
9	300	Medium	42.3%
10	100	Large	139.6%
11	200	Large	139.6%
12	300	Large	139.6%

Each run lasts for 3 hours (each tick in the model represents one second) which is a period around the event and it allows me to measure and analyze the traffic flow before, during, and after the event. If there is an event, it will start at the end of the first hour. Each scenario is run 100 times. For each scenario, the average of the 100 runs is taken for comparison and analysis. To measure the influence on traffic flow, I measured the average speed of cars over time, the average speed of cars near the destination (Madison Square Garden) over time, and the number of stopped cars over time. By measuring the number of stopped cars over time, we are also measuring the wait time of cars. The model is verified through face validity, by checking its animation and graphical representation to make sure the model behaves reasonably and as intended (Xiang et al., 2005). Furthermore, the model could be validated by comparing its results with the taxi dataset in the last chapters. Approximate trip distances and durations of the taxi trips can be derived from the taxi dataset, which gives the approximate speed. As a result, we can estimate the change in speed during events from the taxi data.

4.4 Results

The simulation results are visualized in Figures 4.4 to 4.7. Figure 4.4 shows the average speed of all cars in blue and cars near the destination (Madison Square Garden) in orange. Figure 4.5 shows the number of event-related cars and the number of stopped cars. Figures 4.6 and 4.7 compare the results during the event traffic arrivals hour and regular hours to calculate the percentage change in car speeds and the number of stopped cars. Figure 4.6 shows the percentage changes in average speeds, and Figure 4.7 shows the percentage changes in the number of stopped cars. Each scenario was run 100 times, and the charts show the average of the 100 runs to control the randomness in each simulation run. We can see that normally the traffic near the destination is a little faster than the average of all cars, however, the simulation results show that it could decrease more during events. The results show that the influence on traffic flow is higher with large events during peak hours with high regular traffic, and the influence is also higher for streets near the destination.

Scenarios 1 to 3 are the base cases when there is no event and different traffic conditions (low, medium, high volume of regular cars). The traffic speed in these scenarios is relatively steady over time. Scenarios 4 to 6 show the results of scenarios with small events held during different traffic conditions. In this case, the small events have very little impact on average traffic speed and cars stopped.

Next, I simulate the situations when there are medium events held during different traffic conditions in scenarios 7 to 9. While the event has a small impact on the average speed during low and medium traffic volume, the impact significantly increases as the regular traffic volume increases. The event has a larger impact on car speed near the destination, and car speed near the destination becomes lower than the average speed at some point. The average car speed is slowed by 8.7%, and the car speed near the destination is slowed by 13.1%. At the same time, the number of stopped cars increases by 10.9% in the scenario with high traffic volume.

Last but not least, Scenarios 10 to 12 show the impact of a large event during different traffic conditions. From the charts, we can observe an obvious decrease in car speed and an increase in cars stopped in these three scenarios. Again, the impact is much higher when the regular traffic volume is high, and the impact on streets near the destination is even higher than average. The average car speed is slowed by 28.7%, and the car speed near the destination is slowed by 44.7%. The number of stopped cars increases by 47.2% in this scenario.

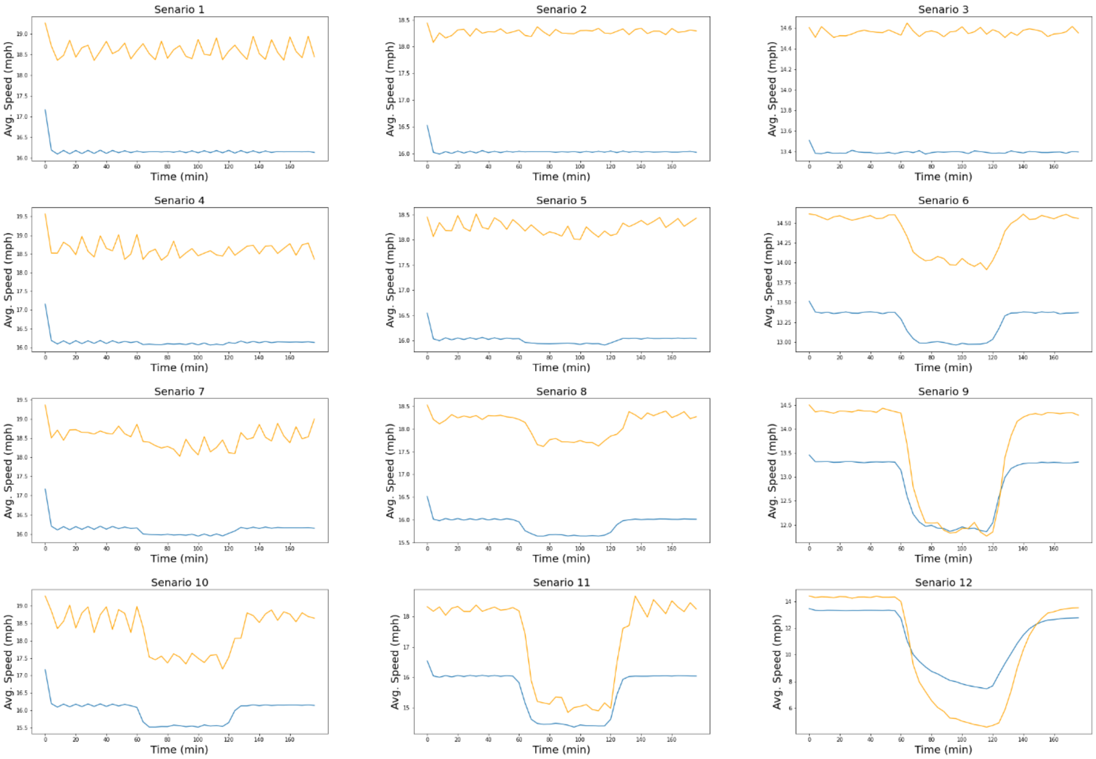


Figure 4.4 Simulation results – the average speed of cars and the average speed of cars near the destination over time.

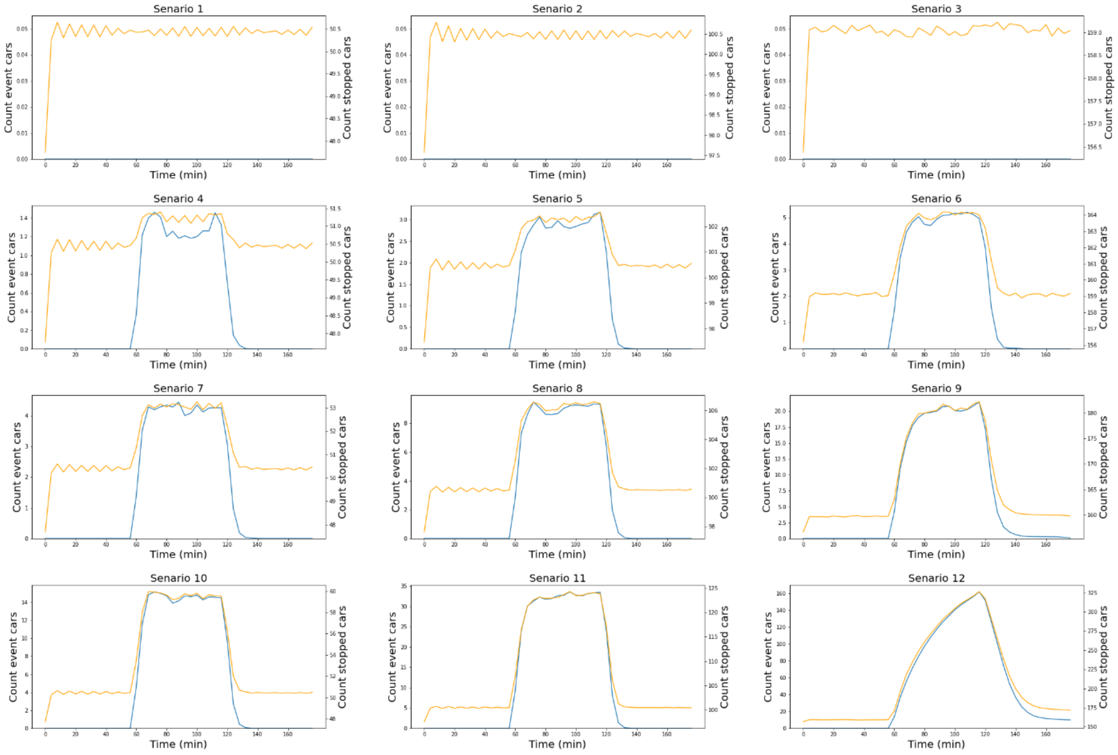


Figure 4.5 Simulation results – the number of event related cars and stopped cars over time.

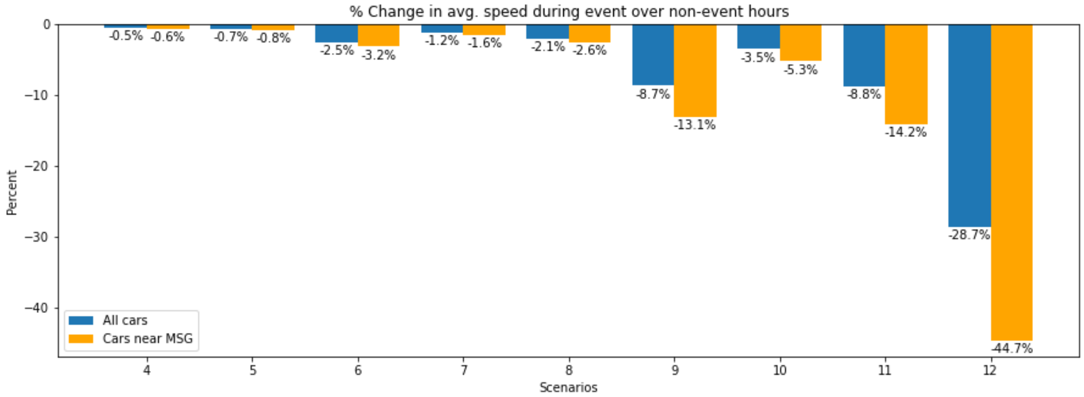


Figure 4.6 Simulation results – percentage change in average speed.

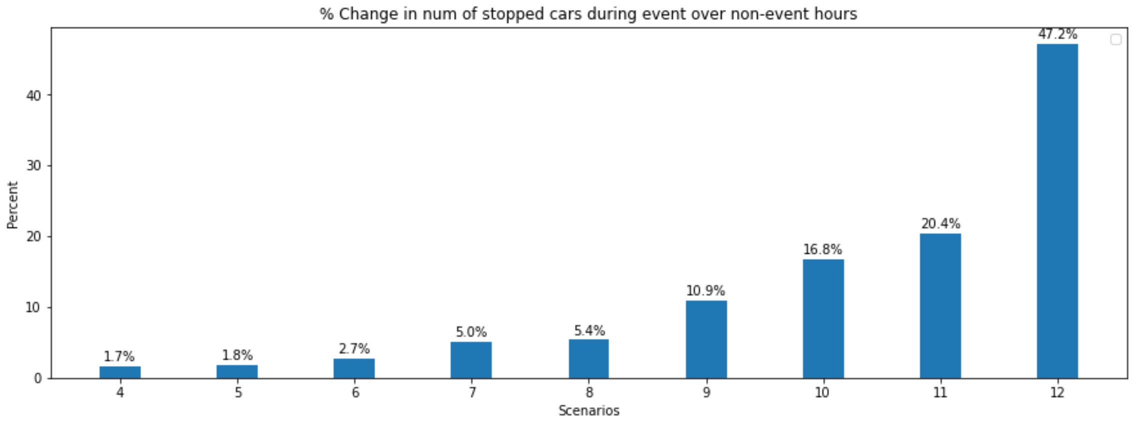


Figure 4.7 Simulation results – percentage change in stopped cars

4.5 Conclusion

In this chapter, I built an ABM to simulate the influence of events on traffic flow. It has been well established that agent-based modeling is a powerful simulation modeling technique for studying traffic problems. However, there is a lack of agent-based models that focus on simulating the influence of non-disaster events on traffic flow using findings from real-world data. This research utilizes the findings from the previous chapter to set the scenarios to test in this ABM, and it shows the potential of combining social media data, spatial-temporal traffic data, and agent-based modeling to study human activities and their effects.

The results of the study show that the magnitude of influence depends on the size of the events and the normal traffic condition of the study area. Small events have very little impact on traffic speed and wait time. Medium events have a more significant impact on the traffic flow when the streets are busy. However, large events will impact traffic flow no matter what the traffic condition is, even when the regular traffic volume is small, the number of cars stopped (which represents wait time for drivers) still increase by 16.8%. When regular traffic volume is high, the large events could decrease traffic speed by 28.7% and increase the wait time by 47.2% as well.

How can the simulation results benefit the field of traffic planning and engineering? When a medium or large event is held during peak hours when traffic volume is high, we will need to plan for the possibility of traffic congestion. In reality, these kinds of large events are often held in large stadiums (such as Yankee Stadium and Citi Field) that are located with enough distance from downtown areas, and their nearby streets usually have lower traffic volume compared to downtown areas. However, we may have a large event or multiple medium events at the same time in the downtown area, such as Madison Square Garden. In this case, we face potential traffic congestion problems. The model could help to study different scenarios and forecast the severity of traffic congestion by comparing the results with the base cases. Then actions could be taken, such as inform the drivers, or plan for short-term traffic flow management (e.g., detour plans for pedestrian and vehicular traffic).

This model could be further developed into a more complex ABM in the future, by improving a few elements. First, it could use a large and realistic map with buildings and roads constructed using shapefiles. This will allow the ability to study the influence more precisely. It will also allow us to simulate more cars in the system and their interactions. For example, we could have two events happening in two nearby buildings, how will they affect each other and the overall traffic speed? Secondly, I am interested in developing a dynamic decision-making system for the drivers in the future. Drivers can learn about the events beforehand to plan for their arrivals (e.g., come earlier to avoid traffic congestion), or they can communicate with other drivers on the traffic conditions when they are driving and re-route to select a faster path. How will this system change the situation? By adding this decision-making system, we can explore ways to forecast and alleviate traffic congestion problems. Therefore, this model has a lot of room for improvement, but at the same time, it also has a lot of potentials. After some modifications to adapt the model to the area of interest, it could be used to analyze more urban mobility problems.

This simple model demonstrates the concept of analyzing the influence of special events on traffic flow using ABM, social media data, and spatial-temporal traffic data. Although this model is not able to predict precise car speed or the number of cars like more complex models, it provides a method to estimate potential traffic congestion severity by comparing the scenarios of interest against the base cases.

Chapter 5 : Conclusion

5.1 Summary of Dissertation Results

Overall, this dissertation explored human activities in cities and their effects by analyzing data obtained from social media and other online sources. Specifically, this dissertation answers the three research questions provided in Chapter 1 and provides contributions to the fields of computational social sciences by examining the potential of using computational methods to study human activities in cities, and the results of the studies also contribute to the fields of geography and urban planning when exploring cities.

Chapter 2 of this dissertation examined the potential of using new sources of data generated through Web 2.0 technology, specifically that of social media data with spatial-temporal information to gain insights into the food related discussions of people living in New York. In this study, I was able to get a few findings from the social media data, specifically, the timeframes of each meal, popularity of different types of foods during different meals, the spatial hot spots of popular foods, and the foods that often appear together. What's more, this study has used a lightweight approach to show how to combine and utilize the three dimensions - textual, spatial, and temporal dimensions of social media data to gain insights into people's thoughts on food topics in New York City, and use them to understand their thoughts about food.

The research in this dissertation was furthered by Chapter 3 by linking social media data with traffic data to explore the impact of special events on traffic flow. Similar to Chapter 2, I was able to extract special events from the tweets by analyzing the tweets. The spatial and temporal information combined allowed me to identify a statistically significant increase in tweets and identify potential events. Then, the textual content of the tweets provided information on the topics of the events. Next, this study examines the correlation between these events and traffic data to explore the impact of events on traffic flow. The research successfully identifies special events, explores their topics, and categorizes them according to their topics. It is also able to correlate the events with the change in traffic data, and summarize the impact of events on traffic flow, broken down by category of events or locations.

The results from Chapter 3 are used in Chapter 4 to simulate the impact of events on traffic flow using an agent-based modeling approach and it was discussed how the results could be applied to solve problems in traffic planning and engineering. This model mimics the streets of New York. The study tests 12 different scenarios and control for regular traffic volume and sizes of events. The first 3 scenarios are the base cases with low, medium, high regular traffic volumes and without events. The other 9 scenarios are the base cases with 3 types of events: small events, medium events, and large events. These events and their impact on traffic flow are defined following the results of Chapter 3. Using the results from Chapter 3, each type of event causes a certain percentage increase in traffic flow. The simulation results show that medium and large events can significantly slow down the traffic and increase wait time, especially when regular traffic volume is high, a large event could slow the traffic by about 29% and increase stopped vehicles by about 45%. Moreover, the speed of vehicles near the destination received a greater impact and decreased by about 47%. This simple model provides a method to estimate potential traffic congestion under different scenarios.

5.2 Contributions of Dissertation

The major contribution of this dissertation is that it showed the potential and methods to explore human activities in cities using social media data together with other open data sources, and through ABM simulation and analysis, the results can help planners and engineers to study, plan, and manage our cities. The methods presented in this dissertation can be used to explore human activities in cities, and the simulation model can be used to study the impact of such activities on traffic flow. This dissertation research used a multi-disciplinary computational social science approach to study human activities in cities, as illustrated in Figure 5.1. This approach utilized data mining and computer simulation techniques to analyze the large amount of data generated and available on the internet to study human activities in cities and their impact. The studies presented in this dissertation also contributes to the field of “digital twins”, which refers to monitoring and understanding digital representations of physical entities to provide continuous feedback to improve quality of life (El Saddik, 2018). Urban digital twins are data models that simulate the complete system of cities (Dembski et al., 2020). The research in my dissertation shows the methods to model cities through analyzing open data sources and simulation models.

Chapters 2 and 3 are two case studies based on New York City, and they studied the food related discussions of New Yorkers and the special events that people talked about in New York City. The results from Chapter 3 were then used to simulate the impact of events on traffic flow using an agent-based modeling approach in Chapter 4. Chapter 2 focused on exploring food related discussions in New York suing social media data. This chapter showed that it is possible to mine human activities in social media data with spatial-temporal dimensions, and using the data we can explore the topics, locations, and time of the activities. This approach could be applied to other cities and even other research topics of interest, and therefore, it provided a means to gain insights into the activities and opinions of people over large areas. In the next chapter, I used a similar approach to extract special events from social media data, and it furthered the study by linking social media data with traffic data to explore the impact of events on traffic flow. This chapter demonstrated the means to link social media data with other data sources to study the impact of human activities. Furthermore, I utilized the results to build an agent-based model that simulated the impact of events on traffic by testing 12 different scenarios and comparing them against the base cases. While a number of ABMs have been developed to study traffic-related problems (see Bazzan & Klüg, 2014 for a review), but none of them have built an ABM to study non-disaster events' impact on traffic using results from social media data analysis. This simple model is a proof of concept for simulating the impact of events on traffic flow using an ABM, and it provided a method to estimate potential traffic congestion severity by comparing the scenarios of interest against the base cases. It could also be modified to study other areas and scenarios. By using open data sources, the models and methods presented in this dissertation provides cost efficient methods to explore human activities in large geographical areas during long periods of time, which complements traditional methods to study human activities, such as interviews and surveys.

For those who are interested in extending the research, scripts created for this dissertation are shared on Github: https://github.com/YangZhouCSS/cssphd, so that they can be used by others to solve similar problems and advance the studies presented in this dissertation.

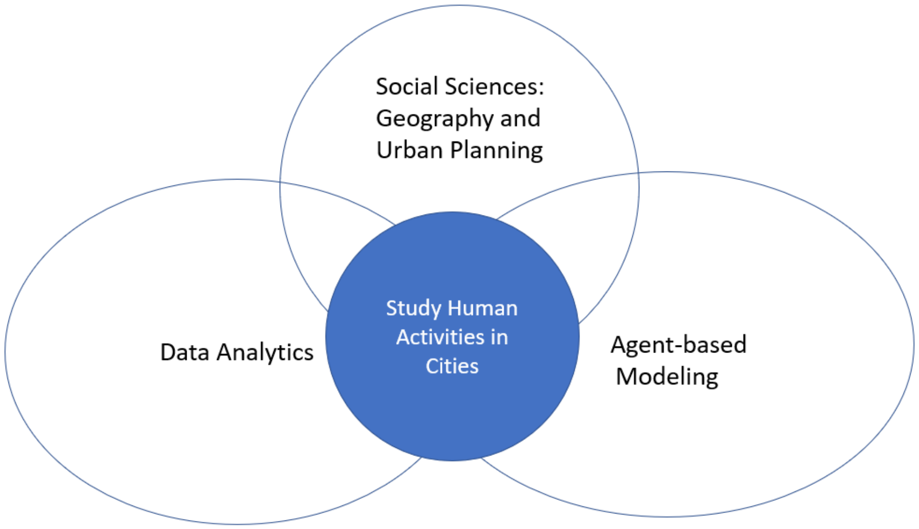


Figure 5.1 The multi-disciplinary computational social science approach to study human activities in cities.

5.3 Limitations and Future Work

While all research has limitations, this research is no exception. It could be expanded in several aspects. Firstly, more and newer data could be collected and added to the study. The Twitter dataset used in this study was collected in 2015, therefore, more recent data will reveal change and new trends in the patterns I discovered, but this study sets the stage for such further study. Moreover, data could be collected on multiple social media platforms such as Instagram to get more data and lower the participation bias. We could also supplement the precisely geolocated tweets used within this study with tweets whose location is derived from the users' profiles or the text of the tweets to get more data points for the analysis. The traffic data could also be supplemented by Uber and Lyft data, bus data, and bike sharing data to reduce participation bias. For example, Uber data has been used to study the impact of ride-sharing on traffic congestion in cities (Li et al., 2016), and they found that Uber's entry into the market significantly reduced traffic congestion in urban areas. In another study, Uber and Lyft data has been used to study the impact of ride-sharing on vehicle miles traveled in cities, and it was found that vehicle miles traveled increased due to the addition of dead-head miles before each pick-up (Schaller, 2021). Bike-sharing data in Chicago was explored to find out patterns in biking behaviors, such as peak hours and spatial clusters of bike demands (Zhou, 2015). Other open data sources such as Metro tap-in and tap-out data, cellphone trajectory data are available in some areas. The research presented in this dissertation can be furthered by combing even more data sources.

Secondly, the meaning of tweets could be further explored. In Chapter 2, if we develop a knowledge base of foods, it is possible to recognize foods in tweets more accurately than using a list of food names. Furthermore, tweets that mentioned food may be talking about eating the food, advertising the food, or just expressing their opinions on the food. If we manually label some of the tweets by why they mentioned foods, and train a classifier, we could potentially better understand the food related discussions and provide more information, such as the dietary habits of people.

Thirdly, the agent-based model can definitely be expanded into a larger area with a real-world map to provide more accurate predictions. Another area for improvement is to allow agents to make decisions and change the environment of the model. For example, if drivers can learn about events and choose to detour, how will that change the results? What if traffic regulations take place? Adding these functions will allow traffic planners to better study the impact of events on traffic flow.

Nonetheless, this study builds lightweight models to extract meaningful results from the data and demonstrates new means to explore human actives using social media data generated by new web technologies. Thanks to the open data initiatives, the datasets used to build the models in this dissertation can be obtained with no extra costs, therefore, they can be used to supplement traditional methods when exploring large areas during a long time period.

References

Abbar, S., Mejova, Y., & Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3197-3206).

Akhtar, N. (2017). Hierarchical summarization of news tweets with twitter-lda. *Applications of Soft Computing for the Web* (pp. 83-98).

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can we predict a riot? Disruptive event detection using Twitter. *ACM Transactions on Internet Technology (TOIT)*, *17*(2), 1-26.

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. *Proceedings of the 2011 Conference on empirical methods in natural language processing* (pp. 1568-1576).

Auld, J., & Mohammadian, A. K. (2012). Activity planning processes in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. *Transportation Research Part A: Policy and Practice*, *46*(8), 1386-1403.

Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., & Nagel, K. (2009). MATSim-T: Architecture and simulation times. *Multi-agent Systems for Traffic and Transportation Engineering* (pp. 57-78).

Batty, M., DeSyllas, J., & Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, *17*(7), 673-697.

Bazzan, A. L., & Klügl, F. (2014). A review on agent-based technology for traffic and transportation. *The Knowledge Engineering Review*, *29*(3), 375.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

Ben-Dor, G., Ben-Elia, E., & Benenson, I. (2018). Assessing the impacts of dedicated bus lanes on urban traffic congestion and modal split with an agent-based model. *Procedia Computer Science*, *130*, 824-829.

Benhardus, J., & Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, *9*(1), 122-139.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of machine Learning research*, *3*, 993-1022.

Bremmer, D., Cotton, K. C., Cotey, D., Prestrud, C. E., & Westby, G. (2004). Measuring congestion: Learning from operational data. *Transportation Research Record*, *1895*(1), 188-196.

Burmeister, B., Doormann, J., & Matylis, G. (1997). Agent-oriented traffic simulation. *Transactions of the Society for Computer Simulation International*, *14*(2), 79-86.

Census. (2019). Commuting (Journey to Work). Retrieved from: https://www.census.gov/topics/employment/commuting.html

Cesare, N., Grant, C., & Nsoesie, E. O. (2019). Understanding demographic bias and representation in social media health data. *Companion Publication of the 10th ACM Conference on Web Science* (pp. 7-9).

Cetin, N., Burri, A., & Nagel, K. (2003). A large-scale agent-based traffic microsimulation based on queue model. *Proceedings of Swiss Transport Research Conference.*

Chang, Y. S., Lee, Y. J., & Choi, S. S. B. (2017). Is there more traffic congestion in larger cities?-Scaling analysis of the 101 largest US urban centers. *Transport Policy*, 59, 54-63.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, *6*(2), e19273.

Cioffi-Revilla, C. (2017). Computation and social science. *Introduction to Computational Social science* (pp. 35-102).

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science, 27*(12), 2483-2508.

Croitoru, A., Stefanidis, A., Radzikowski, J., Crooks, A., Stahl, J., & Wayant, N. (2012). Towards a collaborative geosocial analysis workbench. *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications* (pp. 1-9).

Crooks, A. T., Patel, A., & Wise, S. (2014). Multi-agent systems for urban planning. *Technologies for Urban and Spatial Planning: Virtual Cities and Territories* (pp. 29-56).

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124-147.

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

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, *47*(1), 217-238.

Daniels, R., & Mulley, C. (2013). Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use*, *6*(2), 5-20.

Dembski, F., Wössner, U., Letzgus, M., Ruddat, M., & Yamu, C. (2020). Urban digital twins for smart cities and citizens: the case study of Herrenberg, Germany. *Sustainability*, *12*(6), 2307.

do Amarante, M. D. B., & Bazzan, A. L. (2012). Agent-based simulation of mobility in real-world transportation networks: effects of acquiring information and replanning en-route. In *AAMAS* (pp. 1351-1352).

Doniec, A., Mandiau, R., Piechowiak, S., & Espié, S. (2008). A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, *21*(8), 1443-1454.

Dowling, R., Skabardonis, A., Carroll, M., & Wang, Z. (2004). Methodology for measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record*, *1867*(1), 60-68.

El Saddik, A. (2018). Digital twins: The convergence of multimedia technologies. *IEEE Multimedia*, *25*(2), 87-92.

Food Ingredients A-Z. (2016). Retrieved from: www.bbc.co.uk/food/ingredients.

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., & Bell, D. (2014). Analyzing the language of food on social media. *2014 IEEE International Conference on Big Data (Big Data)* (pp. 778-783).

Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. *Perspectives on Spatial Data Analysis* (pp. 127-145).

Gkountouna, O., Pfoser, D., & Züfle, A. (2020). Traffic Flow Estimation using Probe Vehicle Data. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 579-588).

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the US Congress. *Journal of the American Society for Information Science and Technology*, *61*(8), 1612-1621.

Goldenkoff, R. (2010). 2010 Census: Preliminary Lessons Learned Highlight the Need for Fundamental Reforms. Retrieved from: https://www.gao.gov/products/gao-11-496t

Grinberg, N., Naaman, M., Shaw, B., & Lotan, G. (2013). Extracting diurnal patterns of real world activity from social media. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 7, No. 1).

Guidry, Jeanine D., et al. (2015) From# mcdonaldsfail to# dominossucks: An analysis of Instagram images about the 10 largest fast food companies. *Corporate Communications: An International Journal* 20.3: 344-359.

Guo, S., Lin, Y., Li, S., Chen, Z., & Wan, H. (2019). Deep spatial–temporal 3D convolutional neural networks for traffic data forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3913-3926.

Han, J., Kamber, M., & Pei, J. (2012). 13-data mining trends and research frontiers. *Data Mining (Third Edition)*, 585-631.

Hasan, M., Orgun, M. A., & Schwitter, R. (2018). A survey on real-time event detection from the twitter data stream. *Journal of Information Science*, *44*(4), 443-463.

Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, *56*(3), 1146-1165.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. A*dvances in Neural Information Processing Systems* (pp. 856-864).

Hojati, A. T., Ferreira, L., Washington, S., Charles, P., & Shobeirinejad, A. (2016). Reprint of: Modelling the impact of traffic incidents on travel time reliability. *Transportation Research Part C: Emerging Technologies*, *70*, 86-97.

Horni, A., Nagel, K., & Axhausen, K. W. (2011). High-resolution destination choice in agent-based demand models. *Arbeitsberichte Verkehrs-und Raumplanung*, *682*.

Hu, R., Chiu, Y. C., & Hsieh, C. W. (2020). Crowding prediction on mass rapid transit systems using a weighted bidirectional recurrent neural network. *IET Intelligent Transport Systems*, *14*(3), 196-203.

Hu, Y., & Han, Y. (2019). Identification of urban functional areas based on POI data: A case study of the Guangzhou economic and technological development zone. *Sustainability*, 11(5), 1385.

Huang, D., Xing, J., Liu, Z., & An, Q. (2020). A multi-stage stochastic optimization approach to the stop-skipping and bus lane reservation schemes. *Transportmetrica A: Transport Science*, 1-33.

Idé, T., Katsuki, T., Morimura, T., & Morris, R. (2016). City-wide traffic flow estimation from a limited number of low-quality cameras. *IEEE Transactions on Intelligent Transportation Systems*, *18*(4), 950-959.

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PloS one*, *11*(4), e0152932.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59-68.

Kim, J. S., Jin, H., Kavak, H., Rouly, O. C., Crooks, A., Pfoser, D. & Züfle, A. (2020, June). Location-based social network data generation based on patterns of life. *2020 21st IEEE International Conference on Mobile Data Management (MDM)* (pp. 158-167).

Kim, J. S., Kavak, H., Manzoor, U., Crooks, A., Pfoser, D., Wenk, C., & Züfle, A. (2019, November). Simulating urban patterns of life: A geo-social data generation framework. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 576-579).

Koren, O., Bagozzi, B. E., & Benson, T. S. (2021). Food and water insecurity as causes of social unrest: Evidence from geolocated Twitter data. *Journal of Peace Research*.

Kriegel, H. P., Renz, M., Schubert, M., & Züfle, A. (2008). Statistical density prediction in traffic networks. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 692-703).

Ku, P. J., & Lee, K. (2000). A survey on dietary habit and nutritional knowledge for elementary school children's nutritional education. *Journal of the Korean Society of Food Culture*, *15*(3), 201-213.

Kumar, A., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research*, 1-32.

Kywe, S. M., Hoang, T. A., Lim, E. P., & Zhu, F. (2012). On recommending hashtags in twitter networks. *International Conference on Social Informatics* (pp. 337-350).

Lampos, V., & Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. *2010 2nd International Workshop on Cognitive Information Processing* (pp. 411-416).

Li, Z., Hong, Y., & Zhang, Z. (2016). Do on-demand ride-sharing services affect traffic congestion? *Evidence from Uber Entry*.

Liu, W., Luo, X., Gong, Z., Xuan, J., Kou, N. M., & Xu, Z. (2016). Discovering the core semantics of event from social media. *Future Generation Computer Systems*, *64*, 175-185.

Liu, Z., Ma, T., Du, Y., Pei, T., Yi, J., & Peng, H. (2018). Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS*, *22*(2), 494-513.

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

Louf, R., & Barthelemy, M. (2014). How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports*, *4*(1), 1-9.

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. *Harvard Business School NOM Unit Working Paper*, (12-016).

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427.

Luo, Y., & Bölöni, L. (2012, June). Modeling the conscious behavior of drivers for multi-lane highway driving. *7th International Workshop on Agents in Traffic and Transportation (ATT-2012)* (pp. 95-103).

Mahabir, R., Schuchard, R., Crooks, A., Croitoru, A., & Stefanidis, A. (2020). Crowdsourcing Street View Imagery: A Comparison of Mapillary and OpenStreetCam. *ISPRS International Journal of Geo-Information*, *9*(6), 341.

McCloskey, B., Zumla, A., Ippolito, G., Blumberg, L., Arbon, P., Cicero, A., Borodina, M. (2020). Mass gathering events and reducing further global spread of COVID-19: a political and public health dilemma. *The Lancet*, *395*(10230), 1096-1099.

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An “infodemic”: leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. *Open forum infectious diseases* (Vol. 7, No. 7, p. ofaa258).

Mejova, Y., Haddadi, H., Noulas, A., & Weber, I. (2015). # foodporn: Obesity patterns in culinary interactions. *Proceedings of the 5th International Conference on Digital Health 2015* (pp. 51-58).

Morstatter, F., & Liu, H. (2017). Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media*, *1*, 1-13.

Moshfegh, A., Goldman, J., Ahuja, J., Rhodes, D., & LaComb, R. (2009). What we eat in America, NHANES 2005–2006: usual nutrient intakes from food and water compared to 1997 dietary reference intakes for vitamin D, calcium, phosphorus, and magnesium. *US Department of Agriculture, Agricultural Research Service*.

Muttalif, A. R., Presa, J. V., Haridy, H., Gamil, A., Serra, L. C., & Cané, A. (2019). Incidence and Prevention of Invasive Meningococcal Disease in Global Mass Gathering Events. *Infectious diseases and therapy*, *8*(4), 569-579.

Myslín, M., Zhu, S. H., Chapman, W., & Conway, M. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research*, *15*(8), e174.

Noori, M. A. R., & Mehra, R. (2020). Fire Emergency Detection from Twitter Using Supervised Principal. *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (pp. 403-408).

NYC Taxi and Limousine Commission. (2015). TLC Trip Record Data. Retrieved from: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge-an introduction to the special issue. *Telecommunications Policy*, *39*(9), 745-750.

Panteras, G., Wise, S., Lu, X., Croitoru, A., Crooks, A., & Stefanidis, A. (2015). Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS*, *19*(5), 694-715.

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

Park, Y. M., & Kwan, M. P. (2017). Multi-contextual segregation and environmental justice research: Toward fine-scale spatiotemporal approaches. *International Journal of Environmental Research and Public Health*, *14*(10), 1205.

Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

Poblete, B., Guzmán, J., Maldonado, J., & Tobar, F. (2018). Robust detection of extreme events using Twitter: Worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, *20*(10), 2551-2561.

Rodrigues, F., Borysov, S. S., Ribeiro, B., & Pereira, F. C. (2016). A bayesian additive model for understanding public transport usage in special events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2113-2126.

Rodriguez, P. (2017). Food Classification with Deep Learning in Keras / Tensorflow. Retrieved from: https://github.com/stratospark/food-101-keras

Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health*, *42*(3), 510-516.

Sadilek, A., & Kautz, H. (2013). Modeling the impact of lifestyle on health at scale. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 637-646).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web* (pp. 851-860).

Schaller, B. (2021). Can sharing a ride make for less traffic? Evidence from Uber and Lyft and implications for cities. *Transport Policy*, 102, 1-10.

Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. *Extended semantic web conference* (pp. 22-33).

Shan, Z., & Zhu, Q. (2015). Camera location for real-time traffic state estimation in urban road network using big GPS data. *Neurocomputing*, *169*, 134-143.

Shou, Z., Cao, Z., & Di, X. (2020). Similarity Analysis of Spatial-Temporal Mobility Patterns for Travel Mode Prediction Using Twitter Data. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-6).

Sikder, A., & Züfle, A. (2019, November). Emotion predictions in geo-textual data using spatial statistics and recommendation systems. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising* (pp. 1-4).

Snowdon, J., Gkountouna, O., Züfle, A., & Pfoser, D. (2018). Spatiotemporal traffic volume estimation model based on GPS samples. *Proceedings of the Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data* (pp. 1-6).

Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P. L., Jacobsen, K. H., ... & Crooks, A. (2017). Zika in Twitter: temporal variations of locations, actors, and concepts. *JMIR Public Health and Surveillance*, *3*(2), e22.

Stunkard, A. J. (1959). Eating patterns and obesity. *Psychiatric Quarterly*, *33*(2), 284-295.

Sun, S., Wu, H., & Xiang, L. (2020). City-wide traffic flow forecasting using a deep convolutional neural network. *Sensors*, *20*(2), 421.

Tassone, J., Yan, P., Simpson, M., Mendhe, C., Mago, V., & Choudhury, S. (2020). Utilizing deep learning and graph mining to identify drug use on Twitter data. *BMC Medical Informatics and Decision Making*, *20*(11), 1-15.

Tempelmeier, N., Dietze, S., & Demidova, E. (2020). Crosstown traffic-supervised prediction of impact of planned special events on urban traffic. *GeoInformatica*, *24*(2), 339-370.

The Metropolitan Transportation Authority. (2015). Turnstile Data. Retrieved from: http://web.mta.info/developers/turnstile.html

Tracy McGraw. (2020). Spending 2020 Together on Twitter. Retrieved from: https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter.html

Traffic Flow Estimation using Probe Vehicle Data. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 579-588).

Traffic Flow Estimation using Probe Vehicle Data. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 579-588).

Twitter. (2021). Q4 And Fiscal Year 2020 Letter to Shareholders. Retrieved from: https://investor.twitterinc.com/events-and-presentations/event-details/2021/Twitter-Fourth-Quarter-2020-Earnings-Conference-Call/default.aspx

U.S. Census Bureau. (2019). City and Town Population Totals: 2010-2019. Retrieved from: https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html

U.S. Census Bureau. (2019). QuickFacts: New York city. Retrieved from: https://www.census.gov/quickfacts/newyorkcitynewyork

United Nations. (2018). 2018 Revision of World Urbanization Prospects. Retrieved from: https://population.un.org/wup/.

van Dam, R. M., Rimm, E. B., Willett, W. C., Stampfer, M. J., Hu, F. B. (2002). Dietary patterns and risk for type 2 diabetes mellitus in US men. *Annals of Internal Medicine*, *136*(3), 201-209.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the twitter stream. In *European Conference on Information Retrieval* (pp. 356-367).

Wang, H., Mostafizi, A., Cramer, L. A., Cox, D., Park, H. (2016). An agent-based model of a multimodal near-field tsunami evacuation: Decision-making and life safety. *Transportation Research Part C: Emerging Technologies*, *64*, 86-100.

Wang, Z., Ye, X., & Tsou, M. H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523-540.

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., Chaovalit, P. (2011). Social-based traffic information extraction and classification. In *2011 11th International Conference on ITS Telecommunications* (pp. 107-112).

Wayant, N., Crooks, A. T., Stefanidis, A., Croitoru, A., Radzikowski, J., Stahl, J., & Shine, J. (2012). Spatiotemporal clustering of social media feeds for activity summarization. *Proceedings of the Seventh International Conference for Geographical Information Science*.

Weng, J., & Lee, B. S. (2011). Event detection in twitter. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

West, J.M., Hieb, M.R. and Crooks, A.T. (2019), Modeling Behavior in Drivers of Autonomous Vehicles in Evacuation Situations, *87th Symposium of the Military Operations Research Society (MORS)*, 17th-20th June.

Wilensky, U. (1999). NetLogo. Retrieved from: http://ccl.northwestern.edu/netlogo/.

Wilensky, U. (2003). NetLogo Traffic Grid model. Retrieved from: http://ccl.northwestern.edu/netlogo/models/TrafficGrid.

Xiang, X., Kennedy, R., Madey, G., & Cabaniss, S. (2005). Verification and validation of agent-based scientific simulation models. *Agent-directed Simulation Conference* (Vol. 47, p. 55).

Yamashita, T., & Kurumatani, K. (2009). New approach to smooth traffic flow with route information sharing. *Multi-Agent Systems for Traffic and Transportation Engineering* (pp. 291-306).

Yang, C., Clarke, K., Shekhar, S., & Tao, C. V. (2020). Big Spatiotemporal Data Analytics: A research and innovation frontier. *International Journal of Geographical Information Science Volume 34.*

Yuan X., Crooks, A.T. and Züfle, A. (2020), A Thematic Similarity Network Approach for Analysis of Places Using Volunteered Geographic Information, *ISPRS International Journal of Geo-Information*, 9(6), 385

Zhang, H., Zhou, X., Tang, G., Xiong, L., & Dong, K. (2020). Mining spatial patterns of food culture in China using restaurant POI data. *Transactions in GIS*.

Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. *PloS One*, *10*(10), e0137922.

Biography

Yang Zhou is from Guangzhou, China. She graduated from the Affiliated High School of Guangzhou University in 2009. Then, she came to New York and received her Bachelor of Science in Civil Engineering from New York University in 2013. During her studies, she developed an interest in urban planning and building better cities. As a result, she went to Cornell University to study Regional Sciences with a concentration in Urban Economics. She received her Master of Science in Regional Sciences from Cornell University in 2015. After graduation, she received a Presidential Scholarship at George Mason University to pursue a PhD degree in Computational Social Science. During her PhD studies, she developed an interest in studying cities using data analytics and simulation models. In 2020, She started her career as a data scientist for a consultant firm to develop models to study the characteristics of capital projects.