Fall 2014

Spatial And Temporal Patterns Of Geo-Tagged Tweets

Yue Li
Purdue University

Graduate School Form 30
(Revised 08/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Yue Li

Entitled

SPATIAL AND TEMPORAL PATTERNS OF GEO-TAGGED TWEETS

For the degree of Master of Science

Is approved by the final examining committee:

Jie Shan

James Bethel

Hao Zhang

Ningning Kong

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Jie Shan

Approved by Major Professor(s): _____

_____

Approved by: Dulcy Abraham 12/01/2014

Head of the Department Graduate Program Date

SPATIAL AND TEMPORAL PATTERNS OF GEO-TAGGED TWEETS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Yue Li

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

December 2014

Purdue University

West Lafayette, Indiana

To my dear parents

Xiaoling Wang and Minghai Li

ACKNOWLEDGEMENTS

I am sincerely grateful to my advisor, Dr. Jie Shan, for his valuable scholarly advice, support and encouragement.

I also thank my committee members: Dr. James Bethel, Dr. Ningning Kong, and Dr. Hao Zhang for their assistance, valuable comments and support.

I am eternally grateful to my parents for their endless love, trust and encouragement. My sincerest gratitude also goes to my dearest friends, who are like my family during my stay at Purdue: He Zhang, Ximeng You, Jue Gu, and Shuang Wei; you were always there to listen, offer me advice, and provide support during this entire process.

I thank my colleagues in Geomatics Engineering: Qinghua Li and Chen Ma for their help as well as other friends for discussions and assistance throughout my study. I am also grateful to Sarah Menefee, Dana Gottfried, Paul Briggs and other members of Kossuth Baptist Church for their friendship and support during my study.

Finally, I could not have completed my thesis without the help of those who provided the data, and I therefore thank the Tippecanoe County Area Plan Commission, the GIS team at the City of Bloomington, the City of Ann Arbor, and the City of Columbus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Li, Yue. M.S., Purdue University, December 2014. Spatial and Temporal Patterns of Geo-tagged Tweets. Major Professor: Jie Shan.

With over 500 million current registered users and over 500 million tweets per day, Twitter has caught the attention of scientists in various disciplines. As Twitter allows users to send messages with location tags, a massive amount of valuable geo-social knowledge is embedded in tweets, which can provide useful implications for human geography, urban science, location-based service, targeted advertising, and social network studies. This thesis aims to determine the lifestyle patterns of college students by analyzing the spatial and temporal dynamics in their tweets. Geo-tagged tweets are collected over a period of six months for four US Midwestern college cites: 1) West Lafayette, Indiana (Purdue University); 2) Bloomington, Indiana (Indiana University); 3) Ann Arbor, Michigan (University of Michigan); 4) Columbus, Ohio (The Ohio State University). The overall distribution of the tweets was determined for each city, and the spatial patterns of representative individuals were examined as well. Grouping the tweets in time domains, the temporal patterns on an hourly, daily, and monthly basis were analyzed. Utilizing detailed land use data for each city, further insight about the thematic properties of the tweeting locations was obtained, leading to a deeper understanding about the life, mobility and flow patterns of Twitter users. Finally, space-time clusters and anomalies within tweets,

which were considered events, were found with the space-time statistics. The results generally reflected everyday human activity patterns including the mobile population in each city as well as the commute behaviors of the representative users. The tweets also consistently revealed the occurrence of anomalies or events. The results of this thesis therefore confirmed the feasibility and promising future for using geo-tagged micro-blogging services such as Twitter in understanding human behavior patterns and other geo-social related studies.

.

CHAPTER 1. INTRODUCTION

1.1 General

Twitter is the most popular micro-blogging service in the world. Millions of people use this online social network to socially connect with friends, family members and co-workers (Milstein et al., 2008), and use it to let others know what they are doing or thinking. A status update message is called a "tweet" and each tweet is limited to 140 characters. All users can follow other users, and they can read the tweets they post. Users who are being followed by others do not need to follow them back. The number of Twitter users has increased rapidly since Twitter's launch in 2006; and as of 2014, there were over 500 million registered users, which is more than 2.9 percent of the inhabitants of the Earth (Twitter, 2014). Remarkably, 9.1 percent of the U.S. population "has become the pulse of a planet–wide news organism, hosting the dialogue about everything from the Arab Spring to celebrity deaths" (Stone, 2012). In the last seven years, over 170 billion tweets have been sent, totaling 133 terabytes, with more than 500 million tweets posted each day (Lunden, 2012; Leetaru et al., 2013). Twitter offers "an unprecedented opportunity to study human communication and social networks" (Miller, 2011), and has caught the attention of social researchers. Furthermore, Twitter provides real-time programmatic access to a massive seven-year archive via APIs, and its ease and availability of use have turned Twitter into one of the favorite data sources of social researchers' (Leetaru

et al., 2013).

One important feature about Twitter is its availability from cell phones, which may have embedded location sensors such as GPS, allowing users to send messages with their geographic coordinates (Fujisaka et al., 2010). Also, since August 2009, Twitter has permitted users to manually indicate their city or neighborhood location (Twitter, 2014). On average, two percent of all tweets include location information (Leetaru et al., 2013), which translates to around ten million tweets per day. Therefore, Twitter is becoming a key source of open and free volunteered geographic information (VGI), which is the digital spatial data generated by citizens to gather and disseminate their geographic information and observations (Goodchild, 2007). Geo-tagged tweets have been utilized in a variety of fields, including disaster management (Sakaki et al., 2010), event detection (Nakaji and Yanai, 2012), politics (Tsou et al., 2013), health science (Ghosh and Guha, 2013), crime analysis (Malleson and Andresen, 2014) and human mobility pattern analysis (Fujisaka et al., 2010; Hawelka et al., 2014). The immense volume and diversified information available in tweets have made them a promising or even better alternative to traditional survey data collection, opening new avenues for discovering geo-social knowledge and, thus providing novel research approaches in a number of areas.

1.2 Objectives

In consideration of the characteristics of Twitter data and its potential in geosocial knowledge discovery, the objectives of this thesis are as follows:

- The main objective is to explore the spatial and temporal patterns of geo-tagged tweets in Midwestern college cities by using different data analysis and mining methods.
- The second objective is to infer the mobility patterns of the users behind the tweets and compare the pattern of the four study areas.
- Finally, this research aims to provide a framework for geosocial media data mining and knowledge discovery, especially in the context of human behavior research.

1.3 Related Work

1.3.1 Volunteered Geographic Information

The way people create, use and share geographic information has changed in recent years due to innovate new technologies and online services (Elwood, 2008). The untrained general public can collect and produce spatial data due to the widespread use of hand-held GPS, geotags, high-resolution graphics and access to internet and Web 2.0 (Goodchild, 2007). Unlike the traditional methods of collecting spatial data, which required trained professionals, every human being now can serve as an intelligent sensor interpreting and synthesizing local geographic information. This phenomenon is called Volunteered Geographic Information (VGI) (Goodchild, 2007). VGI not only tremendously increases the volume of existing spatial data, but also alters its content and characteristics (Elwood, 2008). More diversified modes of spatial information, including geo-referenced images, videos and other digital formats, consequently have become available (Elwood and Leszczynski, 2011). This shift deeply impacts the disciplines of geography, sociology and politics with innovative alternative solutions to traditional data collection methods such as

surveys, interviews, and focus groups (Elwood, 2008; Tsou and Leitner, 2013). According to Crampton et al. (2013), the web is not only a collection of longitude-latitude coordinates with information, but a "socially-produced space that blurs the oft-reproduced binary of virtual and material spaces".

The capabilities of producing massive geodata in a short period of time, as well as allowing individuals to report on local and specific conditions make VGI a useful tool for disaster and emergency management (Zook et al., 2010). Three main frameworks in crisis management are map mashups aimed at informing the general public, contribution platforms and collaborative platforms such as Wikimap, OpenStreetMap, etc. (Zook et al., 2010). In the case of the Santa Barbara, California wildfires of 2007-2009, VGI appeared on the Web as text reports, photographs, and video (Goodchild & Glennon, 2010). For example, several individuals and groups set up mapping sites immediately after the Jesusita Fire ignited in May 2009, synthesizing the official information and the VGI. By the end of the fire, 27 volunteer map sites had been established, and the most popular one received over 600,000 hits and offered essential and timely information on the location of the fire, shelters available, evacuation plans, and other useful information (Goodchild & Glennon, 2010). Similarly, Zook et al. (2010) explored the role of web-based mapping services in Haiti relief efforts, and demonstrated the potential of crowdsourced online mapping by providing a way through which individuals can make a contribution without being physically present at the scene.

In addition to disaster management, researchers also have explored the role of VGI in event monitoring, and the possibility of using VGI in event detection. Crampton et al. (2013) focused on the widely reported riots after the University of Kentucky men's

basketball team's 2012 championship, and developed a large data analytic engine with geo-visualization functionality for geo-tagged tweets. Their system analyzed the geography of one specific hashtag #LexingtonPoliceScanner, which referred to the online feed of the Lexington Police Department, to evaluate the capability of using geo-referenced social media data in spatially determining events and news diffusion over time and space. Instead of focusing on one event, Nakaji and Yanai (2012) designed a visualization system for real-world events by utilizing the geotags of tweets as well as the visual features of attached photos. Similarly, Hiruta et al. (2012) used tweets with content relevant to the tagged locations to detect events.

Combined with topic modeling and semantic analysis, VGI has been used in other fields. Tsou et al. (2013) explored the spatial distribution of social media messages and web pages regarding the 2012 U.S. Presidential Election. Web pages and tweets related to "Barack Obama" or "Mitt Romney" were visualized on maps, which highly corresponded to the major campaign events. The results led to the conclusion that this approach was promising in studying human activities, social events and human thoughts quantitatively (Tsou et al., 2013). Ghosh and Guha (2013) aimed to map tweets related to obesity. They used topic modeling to find the topics associated with the keyword "obesity", and analyzed the spatial patterns of these topics with U.S. census data and the locations of fast food restaurants. This study provided a prototype for the use of large conversational datasets on health problems (Ghosh and Guha, 2013).

1.3.2 Human behavior research

Understanding human behavior patterns is important for a wide variety of fields including urban planning, traffic forecasting, spread of biological and mobile viruses, and

crisis management (Gonzalez et al., 2008; Kang et al., 2010). The traditional methods for determining individual human mobility patterns are using travel diary datasets collected by censuses and questionnaires (Kwan, 1999a; Kwan, 1999b). However, the traditional approaches seldom gained enough sample data, and were very time consuming and expensive (Kang et al., 2010). Researchers therefore have been seeking more effective data collection techniques; and due to the rapid advancements in information/ communication technology, cell phones as well as other handheld devices with GPS now have the attention of researchers. With respect to the size of the data, this data collection approach is becoming increasingly promising for exploring individual mobility on a large scale (Kang et al., 2010).

Some researchers have used GPS datasets consisting of cellphone data (Bayir et al., 2009), and metro card transactions (Hasan et al., 2013) among others to understand human mobility and urban characteristics. Bayir et al. (2009) used cellphone data from 100 people in a nine-month period to discover mobile user profiles, and also proposed a "cell clustering" method to filter out noise and improper handoffs. Hasan et al. (2013) used smart subway fare card transactions to model the spatial and temporal patterns of the mobility of individuals in a city. The model is capable of reproducing the frequency of visits as well as a sub-linear increase in the number of different locations visited as a function of time at the individual level, and it can generate the heterogeneous flows at the aggregated level (Hasan et al., 2013).

Researchers have also used VGI, especially social media data for human activity analysis (Li & Shan, 2013). Cheng et al. (2011) used footprints recorded by location sharing services including Foursquare, Gowalla, and Facebook to quantitatively assess

human mobility patterns by extracting its spatial, temporal, social and textural aspects. Similarly, Fujisaka et al. (2010) explored mass movement histories from geo-tagged tweets, and proposed an aggregation model to calculate how many new users entered the region as well as a dispersion model to compute those leaving the region. Hawelka et al. (2014) explored the global mobility pattern using geo-located tweets, and revealed the mobility profiles of different countries, as well as the peak or valley season of international travelers. They also validated the results with global tourism statistics and confirmed Twitter's capability in quantifying global mobility patterns. Besides social media data, Li et al. (2013) also took the socioeconomic characteristics of local people into consideration; and by analyzing their relationship with the density of the tweets, the authors discovered the spatial, temporal and socioeconomic patterns.

Instead of studying the general human activity pattern, Popescu et al. (2009) focused on a certain group of people – tourists. They introduced a method for extracting tourist information, such as the sites people visit and, how long, and panoramic spots from Flickr, covering 183 cities of different sizes from different parts of the world. On the other hand, Malleson and Andresen (2014) discussed the possibility of using VGI in analyzing a special behavior – crime. They discovered that, compared to the residential population, geosocial media data can potentially represent the mobile population, which can be a proxy for the population at risk; their approach was proven helpful to the analysis of the spatial patterns of crimes.

1.3.3 Summary

From the above discussion, we can see that geo-tag tweets and other forms of VGI have been used in a variety of applications such as emergency or crisis management (Zook

et al., 2010), event detection (Nakaji and Yanai, 2012; Crampton *et al.*., 2013), knowledge discovery combined with topic modeling and semantics analysis (Tsou et al., 2013; Ghosh and Guha, 2013). The potential of geo-tagged tweets and other VGI in social science research also has been proven. Researches on modeling human mobility at the individual level has been performed on GPS datasets, such as cellphone data (Bayir et al., 2009) and card transactions (Hasan et al., 2013) among others. Although human behavior research has used tweets and other forms of VGI, they either focus on a certain group of people such as tourists using photo-sharing services (Girardin et al., 2008; Popescu et al., 2009), or on the general public but on a regional scale (Fujisaka et al*.,* 2010), a county scale (Li et al., 2013), or even a global scale (Leetaru et al., 2013). Very limited work has focused on modeling human mobility patterns on a smaller scale such as a city or town. This thesis aims to fill this gap. Also, due to the great volume and public accessibility of tweets, the focus of this thesis is to utilize tweets rather than traditional GPS datasets to better depict human mobility patterns. Thus, the research of this thesis is expected to benefit a wide variety of applications, and inspire sociologists, anthropologists, policy makers, and geographers.

1.4 Organization of the Thesis

This thesis contains five chapters. The remaining chapters are organized as follows:

Chapter 2 describes the study areas, and Twitter data as well as other data used. Acquisition and pre-processing methods for Twitter data also will be discussed.

Chapter 3 describes the methodology for analyzing the spatial and temporal pattern of tweets.

Chapter 4 presents the results, and discusses the human activity patterns revealed.

Chapter 5 describes the generic findings, the limitations of the work, and possible future research directions.

CHAPTER 2. STUDY AREA AND DATA

2.1 Study Area

The study area of this thesis is college cities in the Midwestern U.S., and four particular sites were chose: 1) West Lafayette, IN, home of Purdue University; 2) Bloomington, IN, home of Indiana University; 3) Ann Arbor, MI, home of University of Michigan; 4) Columbus, OH, home of The Ohio State University.

2.1.1 West Lafayette, IN

West Lafayette is the most densely populated city in Indiana with a population of 29,596 as of the 2010 census (2010 Population Finder, 2010). It also is the most culturally-diverse city in the Midwest. The median age is 22.8 years, and 49.4% are between the ages of 18 and 24. The population density is 1,499.6/km^2 (West Lafayette, Indiana, 2014). The city lies in the center of Tippecanoe County, and overlooks the Wabash River (Figure 2.1). It covers 19.76km^2. Purdue University is located in West Lafayette, and has almost 39,256 students, 30,147 of which were undergraduate students in the fall semester of 2012 (Purdue University, 2014). The university has 15 residence halls, and in which approximately one-third of the single undergraduate students live (West Lafayette, Indiana, 2014).

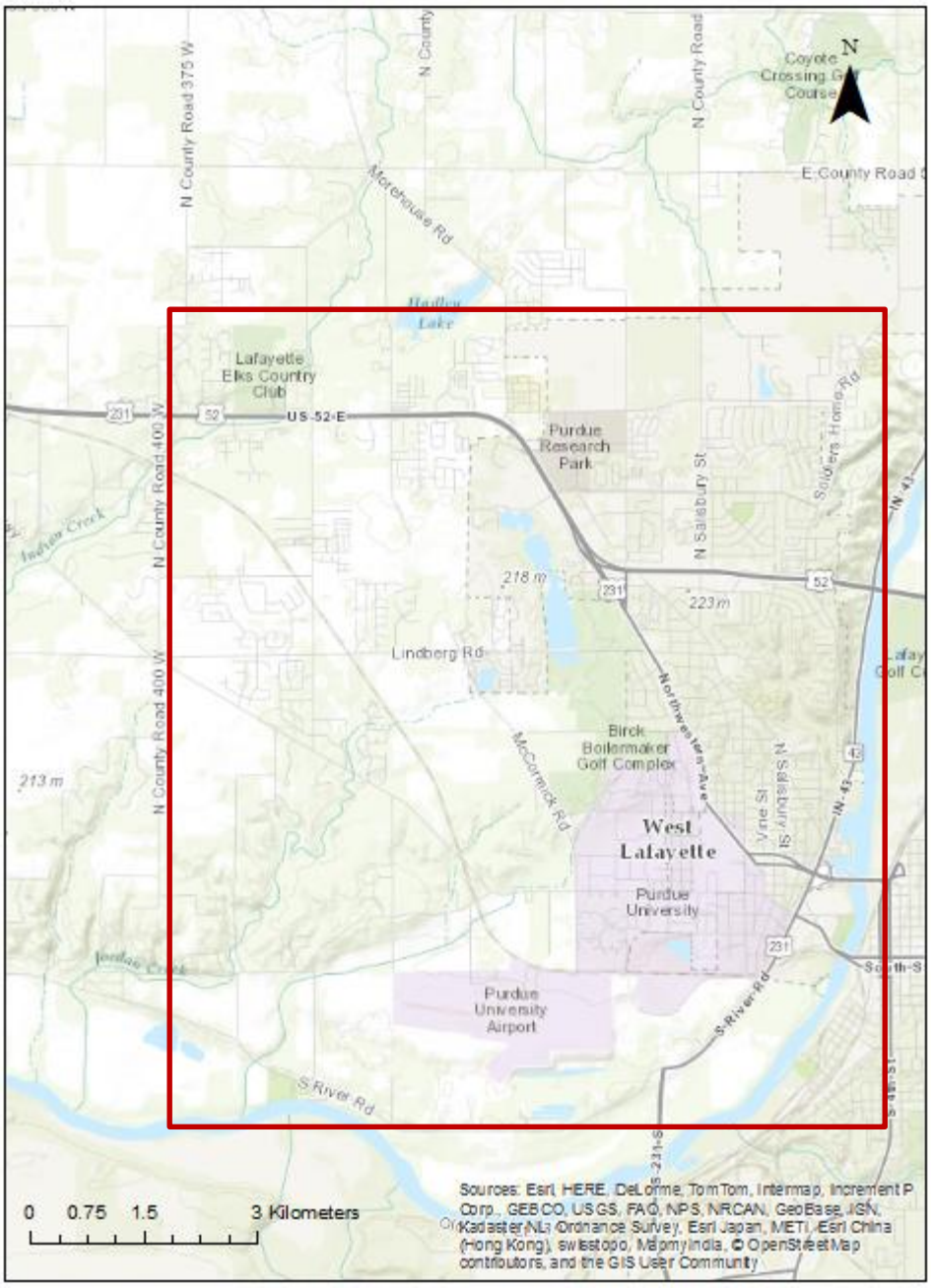


Figure 2.1 Topographic map of West Lafayette, IN
(red box indicates the boundary of tweets being analyzed)

2.1.2 Bloomington, IN

Bloomington is the county seat of Monroe County in the southern section of Indiana. It is the sixth largest city in Indiana, based on its population of 80,405 as of the 2010 census. The population density is around 1,340.4/km^2. The median age in the city is 23.3 years, and 44.5% are between the ages of 18 and 24 (Bloomington, Indiana, 2014). The city covers 60.50km^2. Indiana University Bloomington is located in Bloomington, and has 32,532 undergraduates out of a total student body of 42,731 (Bloomington, Indiana, 2014). 55.2% are from Indiana. There are 12 residence centers on campus which are clustered into three neighborhoods (Housing, 2014).

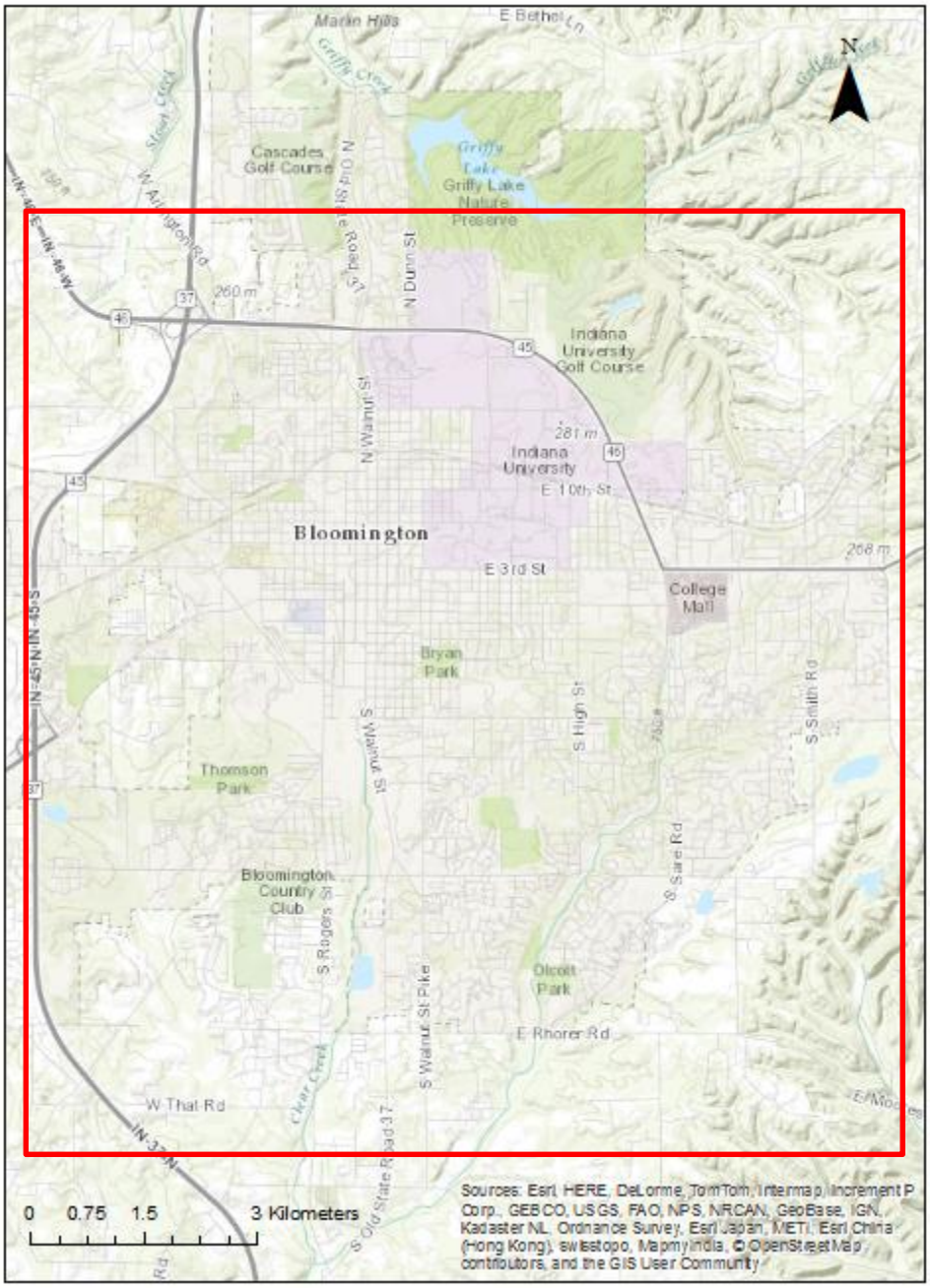


Figure 2.2 Topographic map of Bloomington, IN

(red box indicates the boundary of tweets being analyzed)

2.1.3 Ann Arbor, MI

Ann Arbor is the sixth largest city in Michigan with a population of 113,934 as of the 2010 census and a population density of 1,580.7/km^2. The median age in the city is 28 years, of which 26.8% are between the ages of 18 and 24, and 31.2% are between 25 and 44 (Ann Arbor, Michigan, 2014). The city covers 74.33km^2. Ann Arbor is the home of the University of Michigan, which shapes the city, lending a college-town character (Ann Arbor, Michigan, 2014). The university had 43,246 students as of the fall of 2012, among which 27,979 were undergraduate students. It has four main campuses (North, Central, Medical, and South). The on-campus housing is located on the Central Campus, the Hill Area and the North Campus; and nearly 40% of the undergraduate students live on campus (Housing Options, 2014). Besides the large student population, the university also employees about 30,000 employees, including about 12,000 in the medical center (Ann Arbor, Michigan, 2014). Besides the University of Michigan, Ann Arbor is also home to Concordia University Ann Arbor, a campus of the University of Phoenix, and Cleary University (Ann Arbor, Michigan, 2014).

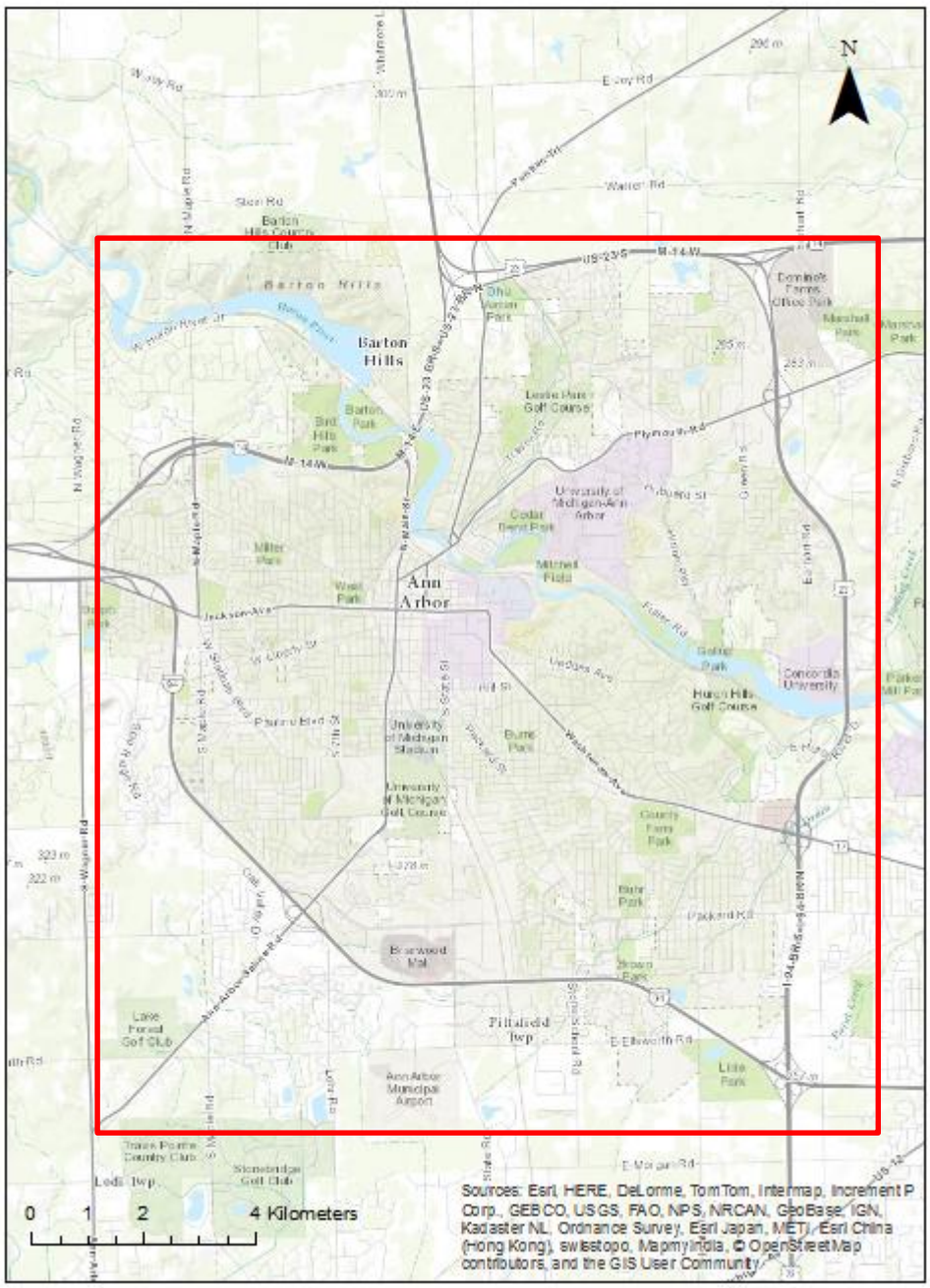


Figure 2.3 Topographic map of Ann Arbor. MI
(red box indicates the boundary of tweets being analyzed)

2.1.4 Columbus, OH

Columbus is the capital of the state of Ohio and its largest city. It is the 15th largest city in the U.S. with a population of 822,553 as of the 2010 census, making it the most populous city in Ohio. The city covers 577.85km^2. The population density is 1,399.2/km^2. The median age from the 2010 census was 31.2 years, of which 14% were between the ages of 18 and 24; and 32.3% were between 25 and 44. The city has a diversified economy, including education, insurance, banking, government, energy, health care, retail, technology, food, clothing, logistics, and health care; and five U.S. Fortune 500 corporation headquarters are located in Columbus as well. The Ohio State University, Columbus State Community College, and many private institutions are located in Columbus (Columbus, Ohio, 2014). The Ohio State University has 56,867 students in total, of which 42,916 are undergraduate students. There are 31 on-campus residence halls, located on the South, North, and West Campuses (The Ohio State University, 2014).

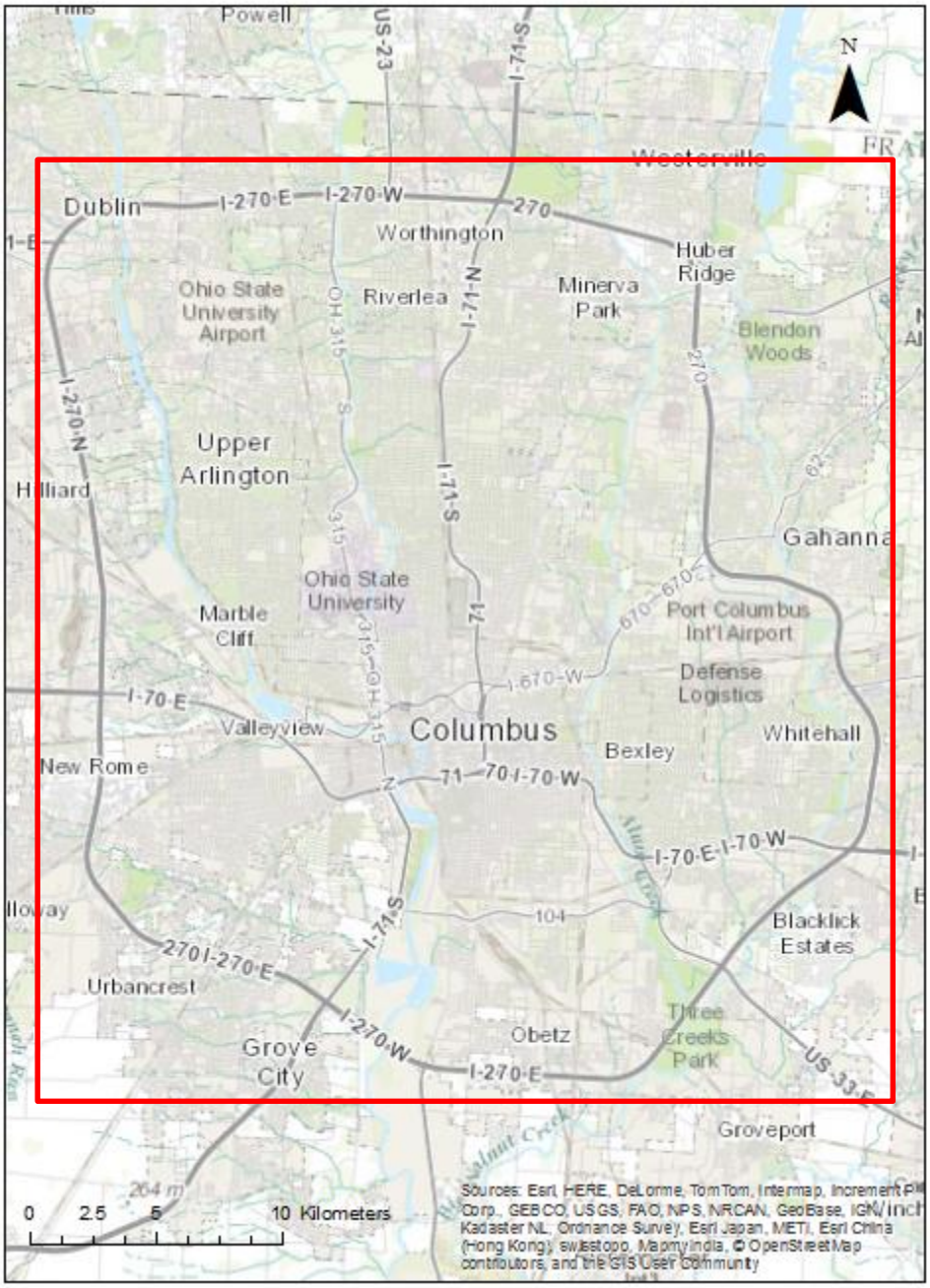


Figure 2.4 Topographic map of Columbus, OH
(red box indicates the boundary of tweets being analyzed)

2.2 Twitter Data

The Twitter data used in this analysis were downloaded using the Twitter Streaming Application Programming Interface (API), which provides developers low latency access to the global stream of Tweet data. There are three main streaming endpoints: 1) the public streams by which the streams of public data flowing through Twitter can be pushed; 2) the user streams by which a single-user's stream containing almost all of the data corresponding to the user's view can be accessed; 3) the site stream, which is a multi-user version of user streams (The Streaming APIs Overview, 2014). Because this thesis aims to understand the pattern of geo-tagged tweets in the four study areas and the tweets within the cites' boundaries were needed, the public stream method was used with two Python libraries, Tweepy and Twitter-Streamer. The search terms used were the coordinate boundaries of the study areas (Table 2.1). The only tweets included were those attached with longitude and latitude, which are usually generated from mobile phones by users who explicitly opt to publish their present locations. I found that around 70% ~ 80% of the tweets were sent from the iPhone OS platform, and 10% ~ 20% were from Android platform (Figure 2.5).

Table 2.1 Coordinates in degrees of the four study areas

Study Area	Southwest corner	Northeast corner
West Lafayette, IN	(-86.970374, 40.4144141)	(-86.895974,40.475314)
Bloomington, IN	(-86.623249,39.101675)	(-86.472874,39.196459)
Ann Arbor, MI	(-83.804226,42.221002)	(-83.673763,42.322620)
Columbus, OH	(-83.194656,39.842747)	(-82.773056,40.204509)

A total of 3,091,794 tweets were downloaded from November 18, 2013 to June 1, 2014, with about 70,000 from West Lafayette; about 300,000 each from Bloomington and Ann Arbor, and more than 2,600,000 from Columbus, which had more than 50,000 users. Columbus also had the highest average number of tweets per user, more than 50. Ann Arbor had the lowest, less than 20 tweets per user (Table 2.2).

Each tweet was downloaded as a JSON object with all the attributes (Figure 2.6). However, since the aim of this thesis it to explore the spatial and temporal patterns, only the attributes needed, such as the time the tweet was posted, its longitude and latitude at the time of posting, and a few relevant fields about the user posting the tweet were included. As the time recorded in a tweet is in Coordinated Universal Time (UTC), it was necessary to convert the posted time to the local time, which was Eastern Time (ET). The time in UTC first was converted to Unix time, or Epoch time, which describes instants in time, and is determined as the number of seconds since 00:00:00 UTC, Thursday, 1 January 1970 (Unix Time, 2014). Then the Epoch time was converted to Eastern Time, and stored in separate fields including “hour”, “day”, “month”, “year”, and “weekday”.

Table 2.2 Number of tweets and users

	West Lafayette	Bloomington	Ann Arbor	Columbus
# tweets	71,658	348,478	295,057	2,671,648
# users	2,884	8,336	15,394	52,149
Average tweets per user	24.85	41.80	19.17	51.23

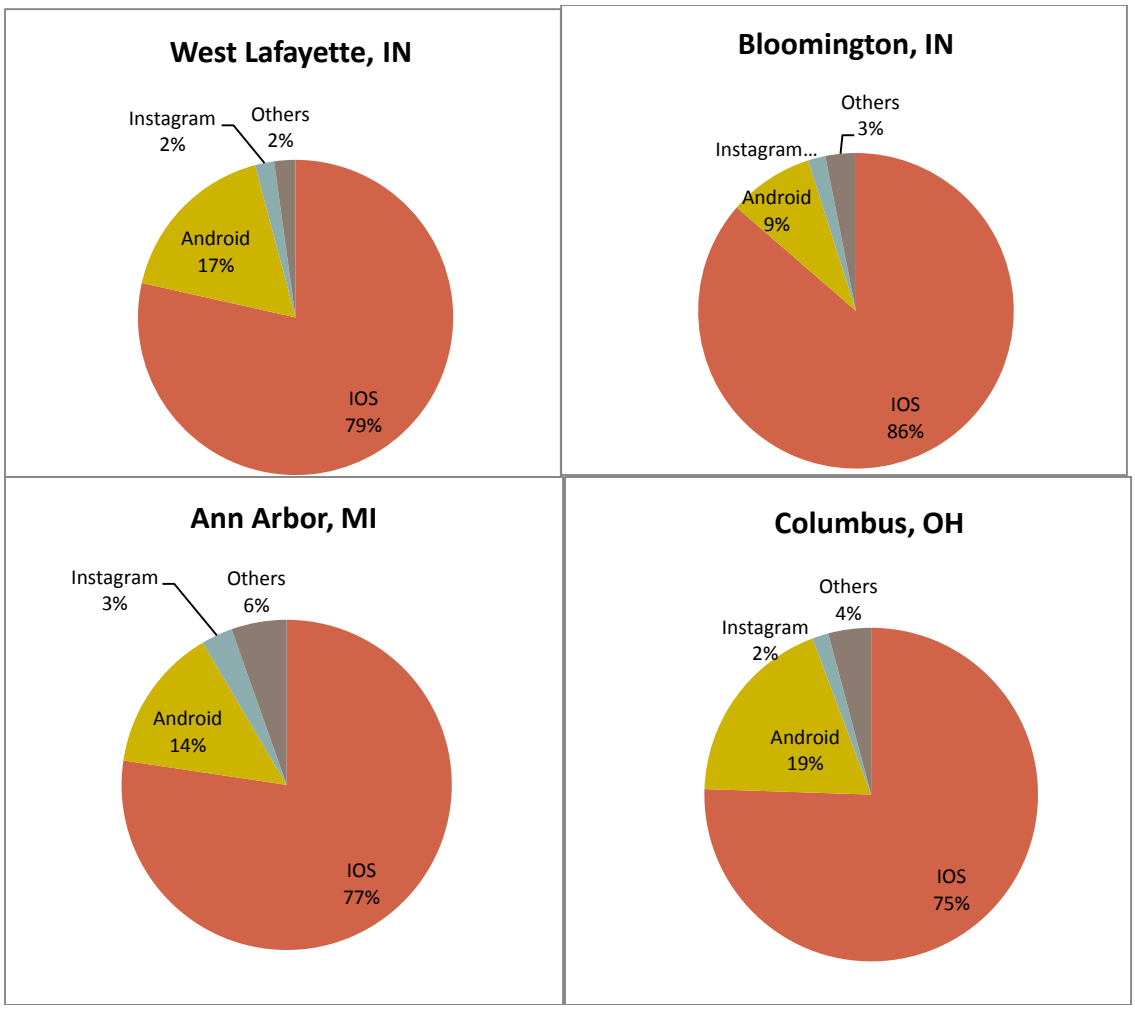


Figure 2.5 Distribution of tweets by platforms

```
{u'contributors': None,
 u'coordinates': {u'coordinates': [-87.6625628, 37.96927275],
  u'type': u'Point'},
 u'created_at': u'Mon Nov 18 03:09:39 +0000 2013',
 u'entities': {u'hashtags': [],
  u'symbols': [],
  u'urls': [],
  u'user_mentions': []},
 u'favorite_count': 0,
 u'favorited': False,
 u'filter_level': u'medium',
 u'geo': {u'coordinates': [37.96927275, -87.6625628], u'type': u'Point'},
 u'id': 402272363360043008,
 u'id_str': u'402272363360043008',
 u'in_reply_to_screen_name': None,
 u'in_reply_to_status_id': None,
 u'in_reply_to_status_id_str': None,
 u'in_reply_to_user_id': None,
 u'in_reply_to_user_id_str': None,
 u'lang': u'en',
 u'place': {u'attributes': {},
  u'bounding_box': {u'coordinates': [[[-88.097892, 37.771741999999996],
     [-88.097892, 41.761368],
     [-84.784662, 41.761368],
     [-84.784662, 37.771741999999996]]],
   u'type': u'Polygon'},
  u'contained_within': [],
  u'country': u'United States',
  u'country_code': u'US',
  u'full_name': u'Indiana, US',
  u'id': u'1010ecfa7d3a40f8',
  u'name': u'Indiana',
  u'place_type': u'admin',
  u'url': u'https://api.twitter.com/1.1/geo/id/1010ecfa7d3a40f8.json'},
 u'retweet_count': 0,
 u'retweeted': False,
 u'source': u'<a href="http://twitter.com/download/iphone" rel="nofollow">Twitte
 u'text': u'I love listening to lies when i know the truth !',
 u'truncated': False,
 u'user': {u'contributors_enabled': False,
  u'created_at': u'Sun Jan 22 12:23:41 +0000 2012',
  u'default_profile': True,
  u'default_profile_image': False,
  u'description': u'Take me as i am ; or watch me as i go !',
  u'favourites_count': 166,
  u'follow_request_sent': None,
  u'followers_count': 262,
  u'following': None,
  u'friends_count': 438,
  u'geo_enabled': True,
  u'id': 471029996,
  u'id_str': u'471029996',
  u'is_translator': False,
  u'lang': u'en',
  u'listed_count': 0,
  u'location': u'Evansville, IN',
  u'name': u'Slooh Alrumeih',
  u'notifications': None,
  u'profile_background_color': u'C0DEED',
  u'profile_background_image_url': u'http://abs.twimg.com/images/themes/theme1/
  u'profile_background_image_url_https': u'https://abs.twimg.com/images/themes/
  u'profile_background_tile': False,
  u'profile_image_url': u'http://pbs.twimg.com/profile_images/3592498625/0c766ac8
  u'profile_image_url_https': u'https://pbs.twimg.com/profile_images/3592498625/0
  u'profile_link_color': u'0084B4',
  u'profile_sidebar_border_color': u'C0DEED',
  u'profile_sidebar_fill_color': u'DDEEF6',
  u'profile_text_color': u'333333',
  u'profile_use_background_image': True,
  u'protected': False,
  u'screen_name': u'saalrumeih',
  u'statuses_count': 1941,
  u'time_zone': u'Baghdad',
  u'url': None,
  u'utc_offset': 10800,
  u'verified': False}}
```


Figure 2.6 A sample tweet downloaded

2.3 Land Use Data

Local land use data were included to assist with interpreting the human mobility patterns behind the spatial and temporal patterns of the tweets. To compare the patterns between the different study areas, the land use types in each city were grouped into more general categories. For West Lafayette, the land use data were digitized based on the zoning map provided by the Tippecanoe County GIS website; and the original zoning classes (Table 2.3) were clustered into five groups: institutional, residential, business, development and others. The Bloomington land use data were downloaded from the City of Bloomington GIS website; and the land use classes (Table 2.4) were regrouped into five groups: institutional, residential, commercial, planned unit development, and others. Ann Arbor's land use information was retrieved from the city's website; the classes (Table 2.5) were reclassified into five groups: institutional, residential, commercial, transportation, and others. The Columbus land use was obtained from the Columbus city GIS office; and the zoning classes (Table 2.6) were categorized into five groups: institutional, residential, commercial, downtown district, and manufacturing.

Table 2.3 Land use in West Lafayette, IN

Original Class	Grouped Class
A (agricultural), AA (select agricultural), AW (agricultural and wooded)	Agricultural
CB, CBW (central business), GB (general business), HB (highway business), NB, NBU (neighborhood business)	Business
PDCC (condominium conversion planned development), PDMX (mixed-use planned development), PDNR (nonresidential planned development), PDRS(residential planned development)	Development
R1, R1A, R1B, R1U (single family residential), R2, R2U (single family and two family), R3, R3U,R3W,R4W (single, two and multi-family), RE (rural estate)	Residential
I1, I2, I3 (industrial), FP (floodplain), MR, MRU (medical related), OR (OFFICE)	Others

Table 2.4 Land use in Bloomington, IN

Original Class	Grouped Class
IN (institutional)	Institutional
CA (arterial commercial), CD (downtown commercial), CG (general commercial), CL (limited commercial)	Commercial
MH (manufactured home), RH (residential high density), RE (residential estate), RS (residential single-family), RM (residential multi-family), RC (residential core)	Residential
PUD (planned unit development)	PUD
IG (industrial), MD (medical), BP (business park), QY (quarry)	Others

Table 2.5 Land use in Ann Arbor, MI

Original Class	Grouped Class
restaurants, general retail, auto service, trade retail, personal service, entertainment, wholesale	Commercial
assembly, cemetery, government, hospital, institution, organizations, religious, cultural, education	Public/quasi-public/institutional
assisted living, bed&breakfast, group housing, hotel/motel, mobile home park, multiple family, non-residential mixed use, single family, two family	Residential
communication facility, local transportation, parking, railroad, road transportation, utility facilities	Transportation/communication/utilities
warehousing, non-manufacturing, agricultural, heavy manufacturing, light assembly, research, residential/non-residential, financial/bank, medical, prof./general, indoor, mixed use, outdoor, lake, vacant	Others

Table 2.6 Land use in Columbus, OH

Original Class	Grouped Class
manufactured home, multi-family, neighborhood center, neighborhood edge, neighborhood general, residential manufacturing	Residential
institutional, research park	Institutional
Commercial	Commercial
Downtown District	Downtown District
east franklin district, excavation, parking, town center	Others

CHAPTER 3. METHODOLOGY

3.1 Overall Spatial Density

Knowing the locations where people usually tweet can be important for a variety of applications. However, due to the point aggregations resulting from the large volume of data, simply displaying all the tweets on a map would not be useful for revealing the patterns of interest in this. Therefore, proper methods were needed to extract the most useful information and to summarize the patterns. The density surface of the locations of the tweets in each study area were generated using ArcGIS. As Columbus dataset contained more than 2,600,000 points which exceeded the capability of ArcGIS, a random subset of the dataset was created with 150,000 tweets for the point density. The Point Density tool computed the density of point features within a neighborhood around each cell. The neighborhood was pre-defined, and the number of points within the neighborhood were summed up and divided by the area of the neighborhood.

Therefore, since the units for the maps were meters, the density values here represented the number of points per square meter. A change in radius may not greatly impact the computed density values because even though the number of points inside the neighborhood changes, the area by which the number will be divided changes as well. Thus, a larger radius would only result in more points being considered in the density

calculation, which would lead to a more generalized output raster and a smaller radius of results in a more detailed density surface raster (Point Density (Spatial Analyst)). The radiuses of the neighborhood were carefully chosen considering the diagonal length of the study area (Table 3.1.1). Specifically, radiuses were around 0.25% of the diagonal length of the study area, and the cell size was the same as the radius.

Table 3.1 Radius (km) of neighborhood in point density calculation

	West Lafayette	**Bloomington**	**Ann Arbor**	**Columbus**
Radius	0.020	0.030	0.025	0.100
Diagonal length of the study area	8	12	10	40

3.2 Spatial Clustering

An Expectation-Maximization (EM) algorithm was used for clustering the tweets of individual users, and the tweets are assumed to follow Gaussian Mixture Model (GMM). For each individual Twitter user, the EM algorithm took all the user's tweets (x); the total number of clusters (M) which was defined as 5 in this analysis; the accepted error to converge (e) which was 10^{-10} degree here; and the maximum number of iterations, which was set as 3000. For each iteration, the first step, the E-step (E-xpectation), assessed the probability of each point belonging to each cluster. Then, in the second step, the M-step (M-aximiation), the parameter vector of the probability distribution of each class was re-estimated. The algorithms were run until the distribution parameters converged or reached the maximum number of iterations (Dempster et al., 1977). Following are the details in implementing this algorithm for one Twitter user:

1) Initialization: each cluster j in the M clusters consisted of a parameter vector (θ). The vector consisted of the mean (μ_j), the covariance matrix (σ_j) and the average responsibility which cluster C_j takes for explaining the data point x_k (π_k). The following represents the features of the Gaussian probability distribution to describe the observed and unobserved entities of the data point x.

$$\theta_j(t) = \mu_j(t), \sigma_j(t), \pi_j(t)\ j = 1 \dots M$$

Initially (t=0), the random values of the mean (μ_j), covariance matrix (σ_j), and probability of occurrence of each cluster (π_j) were generated. This algorithm estimated the parameter vector of the real distribution.

2) E-Step approximated the probability of each point belonging to each cluster ($P(C_j|x_k)$). Each point as composed by an attribute vector (x_k), in this case, the longitude and latitude. The relevance degree of the points of each cluster was calculated as the likelihood of each point attribute compared with the attributes of the other points of the clusters C_j (Equation 3.1).

$$P(C_j|x_k) = \frac{\pi_j \cdot |\sigma_j|^{-\frac{1}{2}} \cdot \exp[-\frac{1}{2}(x_k-\mu_j)^t \cdot \sigma_j^{-1} \cdot (x_k-\mu_j)]}{\Sigma_{i=1}^{M} \pi_j \cdot |\sigma_i|^{-\frac{1}{2}} \cdot \exp[-\frac{1}{2}(x_k-\mu_i)^t \cdot \sigma_i^{-1} \cdot (x_k-\mu_i)]} \quad (3.1)$$

3) M-Step estimated the parameters of the probability distribution of each cluster for the next step. The mean (μ_j) of the cluster j was computed as the mean of all the points in the function of the relevance degree of each point. Suppose there were N points in C_j (Equation 3.2).

$$\mu_j(t+1) = \frac{\Sigma_{k=1}^{N} P(C_j|x_k) x_k}{\Sigma_{k=1}^{N} P(C_j|x_k)} \quad (3.2)$$

The covariance matrix for the next iteration was calculated with the Bayes Theorem (Equation 3.3).

$$\sigma_j(t+1) = \frac{\Sigma_{k=1}^{N} P\left(C_j \middle| x_k\right)\left(x_k - \mu_j(t)\right)\left(x_k - \mu_j(t)\right)^T}{\Sigma_{k=1}^{N} P(C_j|x_k)} \quad (3.3)$$

The probability of occurrence of each cluster was calculated as the mean of the probabilities (C_j) in the function of the relevance degree of each point from the cluster (Equation 3.4).

$$\pi_j(t+1) = \frac{1}{N}\Sigma_{k=1}^{N} P(C_j|x_k) \quad (3.4)$$

The attributes were the parameter vector θ, which describes the probability distribution of each cluster and was used in the next iteration.

4) A convergence test verified whether the difference of the attribute vector of the iteration to that of the previous iteration was smaller than the defined error tolerance after each iteration (Nasser et al., 2006).

Then, since the clusters were places of frequent visits, which very likely were users' homes and workplaces, the distance between the cluster centers could approximate the commute distance of users. For each Twitter user, the average of all the distances between any two centers was calculated as the user's average commute distance.

3.3 Temporal Analysis

With the time stamp associated with the data, various temporal analyses were performed to uncover the temporal patterns in each study area. The analysis was conducted in three stages: 1) by the hour of day; 2) by the day of the week; 3) by the month. First, the tweets were summarized by hour to reveal the people's dynamics during a day and to find

the peak and valley times of their Twitter use. The tweets posted anytime within an hour were totaled. Then by counting the number of tweets on each day of the week, it was possible to determine the day of the week that users were most likely to use Twitter as well as the day with the least usage. Finally, the total numbers of tweets in each month between December 2013 and May 2014 were calculated for each study area and then compared to discover the potential patterns. November 2013 was not included since the data only contain tweets after Nov 18.

The land use data gave further insight about locations of tweet incidents, leading to a deeper understanding about the population mobility, lifestyles and flow patterns of Twitter users. The land use data were spatially joined to the tweet incidents in ArcGIS 10.1, and an analysis of how the number of tweets in each land use type changed with time was conducted. Similar to the temporal analysis performed above, three time intervals were used: 1) the hour of day; 2) the days of the week; and 3) the month.

3.4 Event Detection

In this analysis, space-time scan statistics (STSS) was used to identify the space-time locations of tweet clusters, and thus to determine the occurrence of events. It was assumed that when an event occurred, the users would tweet more than usual to spread the word and describe the event, which would lead to clusters of tweets. STSS has been applied in various situations, such as analysis of crime (Nakaya and Yaho, 2010), forest fires (Vadrevu, 2008), and construction (Stevenson et al., 2010). STSS perceives data points, known as incidences or cases, in a space-time cube. In this thesis, each tweets is a case. A cylindrical window of varying radii (space) and heights (time) moves across the study area,

which is repeated until all possible space-time locations have been visited (Block, 2007). Each window is a candidate for cluster. The number of incidences in each window is compared to the number of expected incidences for that window. Then the significance of each cluster is tested, and a p-value, showing the likelihood that it occurs by chance, is calculated (Cheng and Wicks, 2014).

The STSS method is implemented via SaTScan 9.3 Software (Kulldorff, 2009) and is used retrospectively. The retrospective method searches for clusters across all possible time periods in the data, thereby discovering historic clusters. The other option is to apply STSS to the data prospectively, where only ongoing clusters in the most recent time period can be discovered (Kulldorff, 2014). As this analysis aims to find possible events during the time period, the retrospective method was used.

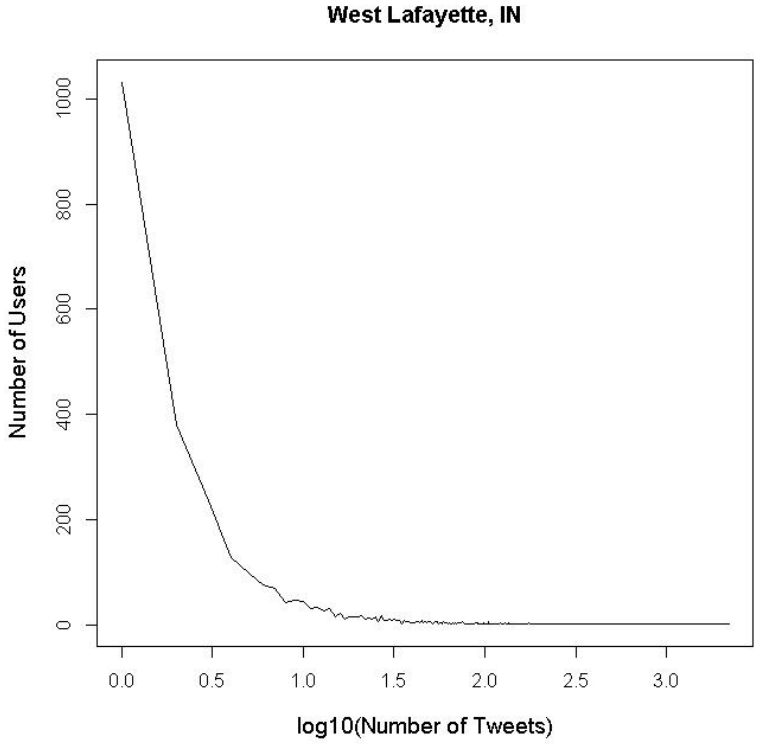
Moreover, STSS method is used with different models. In this thesis, space-time permutation model (STPM) and Poisson model are utilized. STPM only requires data to have spatial and temporal attributes but no other information. As the tweets are going to be clustered only with space and time regardless of the content, STPM was the most suitable method. For likelihood ratio test, STPM uses the same function as the Poisson model (Kulldorff et al., 2005).

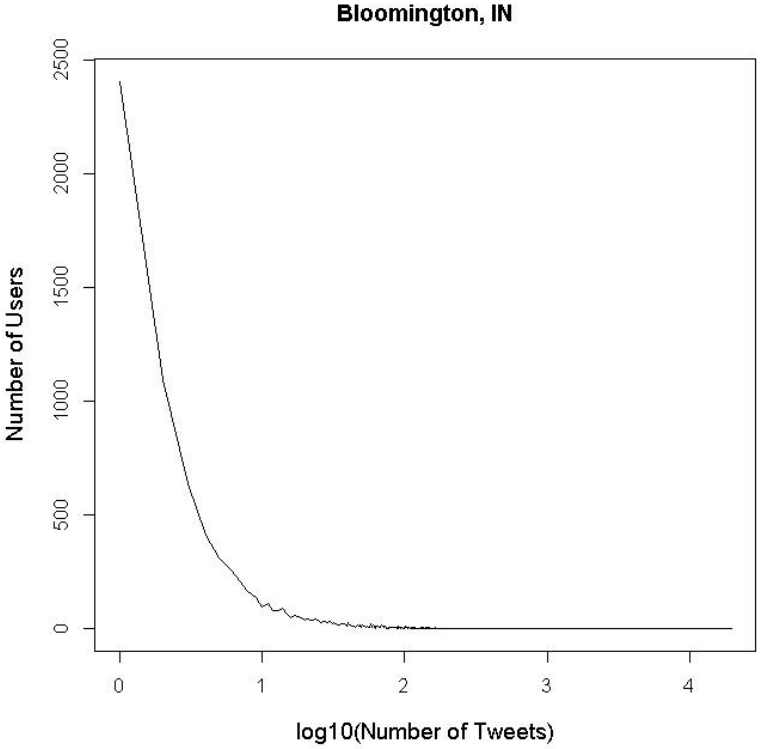
STSS method can also be used for purely temporal clusters, meaning that the bottom of the cylindrical window covers the whole study area. Poisson model is utilized where the number of points in each window is recorded and compared to its distribution under the null hypothesis of a purely random Poisson process (Kulldorff, 1997).

CHAPTER 4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

First, in order to determine how many tweets Twitter users posted in these four areas, the relationship between the number of users and the number of tweets was analyzed. A long tail was discovered in the distribution of the number of users vs. the number of tweets (Figure 4.1). The long tails included relatively fewer users who had posted most tweets, which made up the majority of the distribution (Figure 4.1). Even though these "long tail" users were a small portion of the total number of users, they had posted the majority of the tweets (Table 4.1). According to the first quartile statistics, 25% of the tweets were tweeted from users with less than 55, 111, 98, and 224 tweets for the four study areas (Table 4.2). Thus, it was possible to infer that the "long tail" users made a large contribution; and by analyzing their tweets, a great deal of information was found. As these long tail users posted relatively more tweets than other users, they could be regarded as "frequent Twitter users". Also, due to the large number of tweets posted from these users, determining their mobility patterns and the frequent places they visited became possible. In this thesis, users with more than 100 tweets were defined as frequent Twitter users. Although only around 4% ~ 8% of the total Twitter users were included in this analysis, about 40%~70% of all the tweets were used (Table 4.2).


West Lafayette, IN
Number of Users
0
200
400
600
800
1000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log10(Number of Tweets)


Bloomington, IN
Number of Users
0
500
1000
1500
2000
2500
0
1
2
3
4
log10(Number of Tweets)

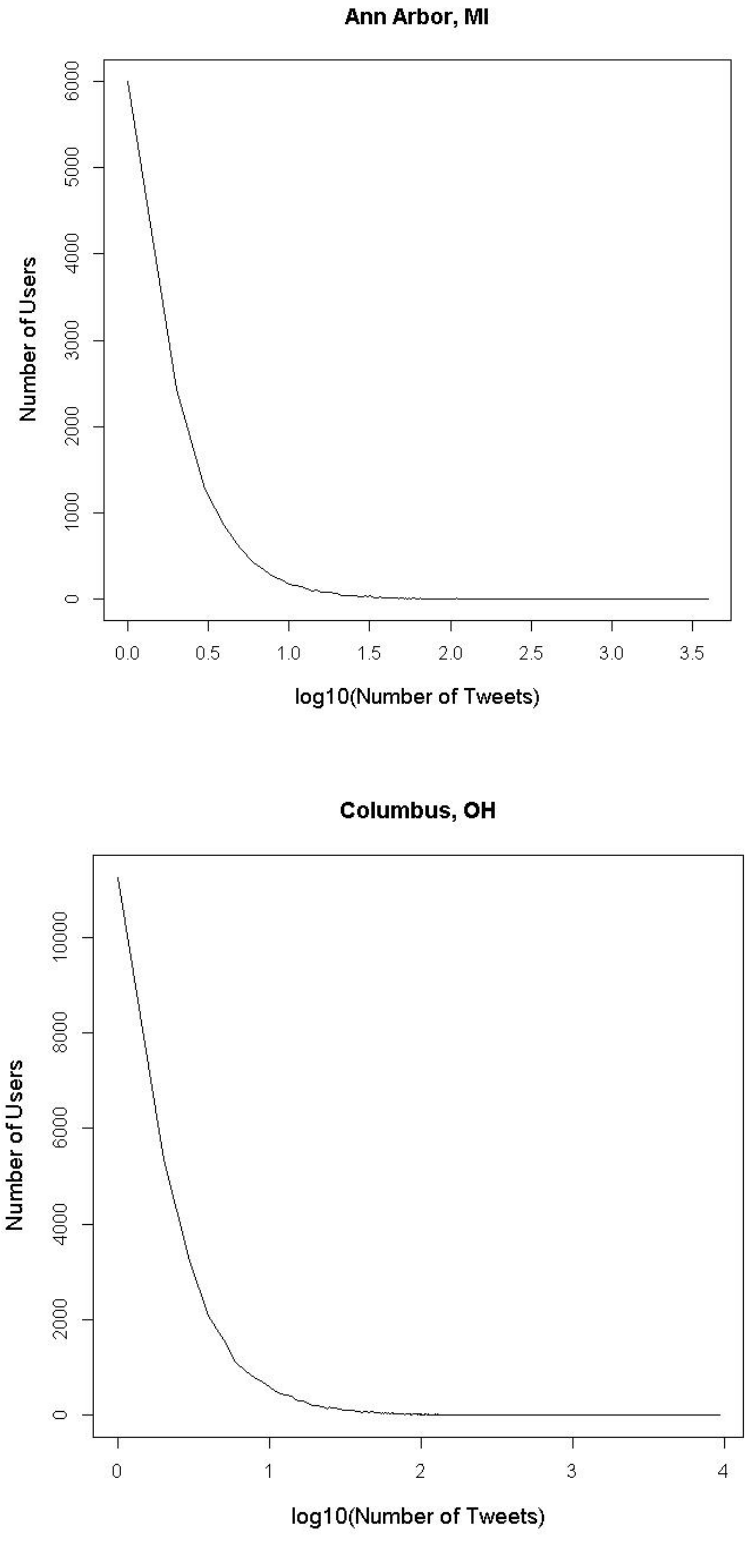


Figure 4.1 Number of users against log10(number of tweets)

Table 4.1 Summary statistics about distribution of number of users against number of tweets

Number of Tweets	West Lafayette	Bloomington	Ann Arbor	Columbus
Min	1	1	1	1
1st Quartile	55	111.8	98.25	224.5
Median	117	241.5	202.5	465.0
Mean	192	415.4	316.5	725.6
3rd Quartile	224	454.2	375.8	879.0
Max	2206	19520	3918	9287

Table 4.2 Number of Twitter users and number of tweets in frequent user analysis

	West Lafayette, IN	Bloomington, IN	Ann Arbor, MI	Columbus, OH
# Twitter Users with more than 100 tweets	153	725	571	2661
Total of Twitter users	2,884	8,336	15,394	52,149
Percentage	5.3%	8.6%	3.7%	5.1%
# Tweets from users with more than 100 tweets	41,402	248,549	168,138	1,071,941
Total of tweets	71,658	348,478	295,057	2,671,648
Percentage	57.7%	71.3%	56.9%	40.1%

4.2 Overall Spatial Density

Bloomington had the most densely distributed tweets with more than four tweets per square meter (Figure 4.3). The highest density of tweets in Columbus should have been 1.06, which was 13 times the density calculated since the sample dataset used was a subset. Columbus was similar to West Lafayette, which had a highest density of 1.05 (Figure 4.1); but compared to West Lafayette, Bloomington, and Ann Arbor, where most of the tweet clusters appeared around the campuses (Figure 4.2, Figure 4.3, and Figure 4.4), the

locations of the clusters in Columbus were scattered all over the city and were more evenly distributed (Figure 4.5). A closer look at each study area follows.

The tweets in West Lafayette were geographically concentrated on the Purdue University campus and its surroundings, especially in the classroom buildings and in the on-campus dorms for undergraduate students (Figure 4.2). Also, a few hot spots appeared at a few apartment complexes such as the Avenue South and Willowbrook, where the majority of the residents were Purdue students (Figure 4.2). Similarly, most of the hot spots in Bloomington occurred on the Indiana University campus and its surrounding areas, which covers the area bounded by Union Street and College Avenue as well as Third Street and IN-45 (Figure 4.3). Other hot spots included Woodbridge Apartment at John Hinkle Place, Campus Corner Apartments, the Village at Muller Park Apartments and others on Muller Parkway (Figure 4.3).

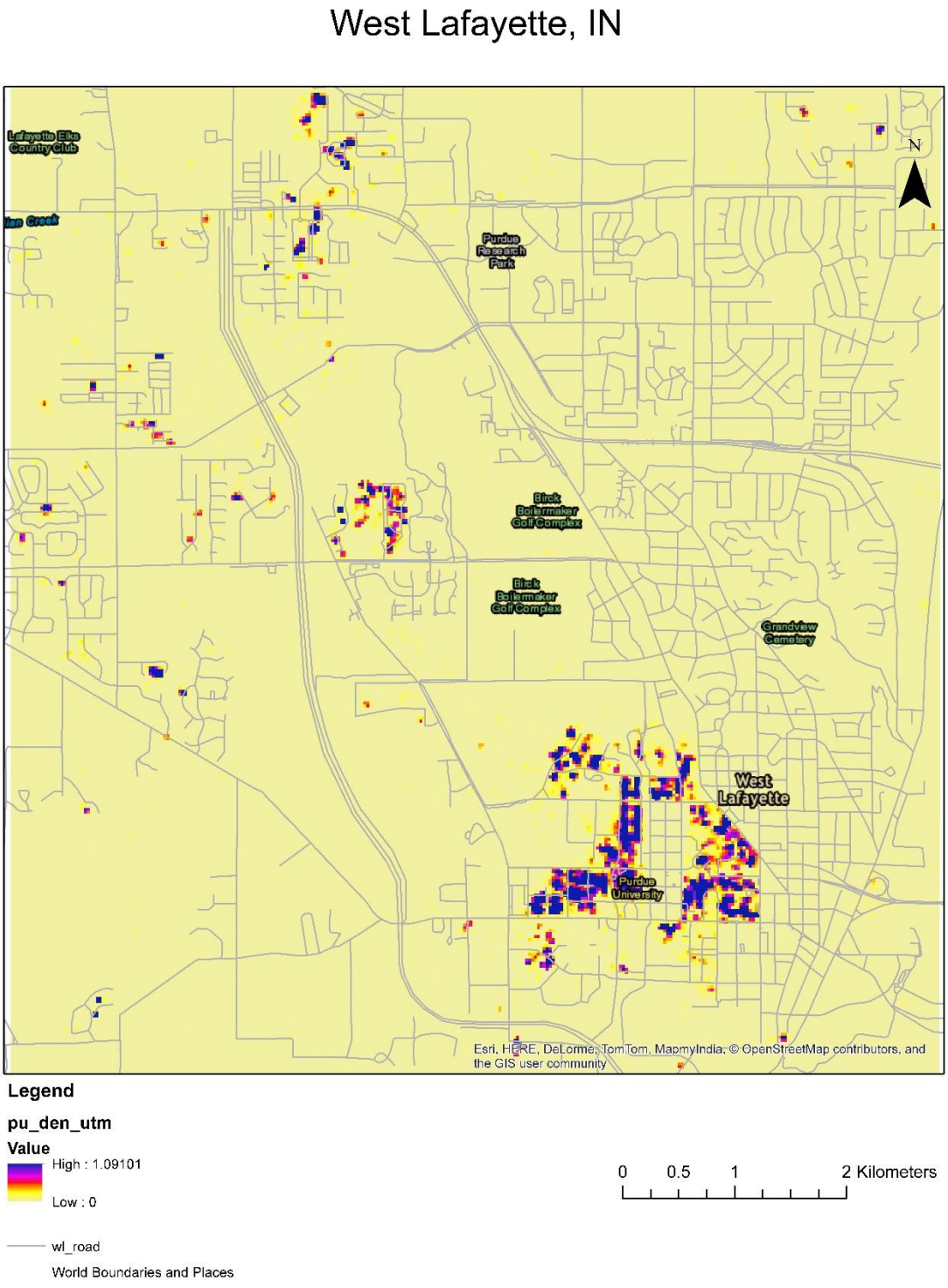


Figure 4.2 Point density raster of West Lafayette, IN

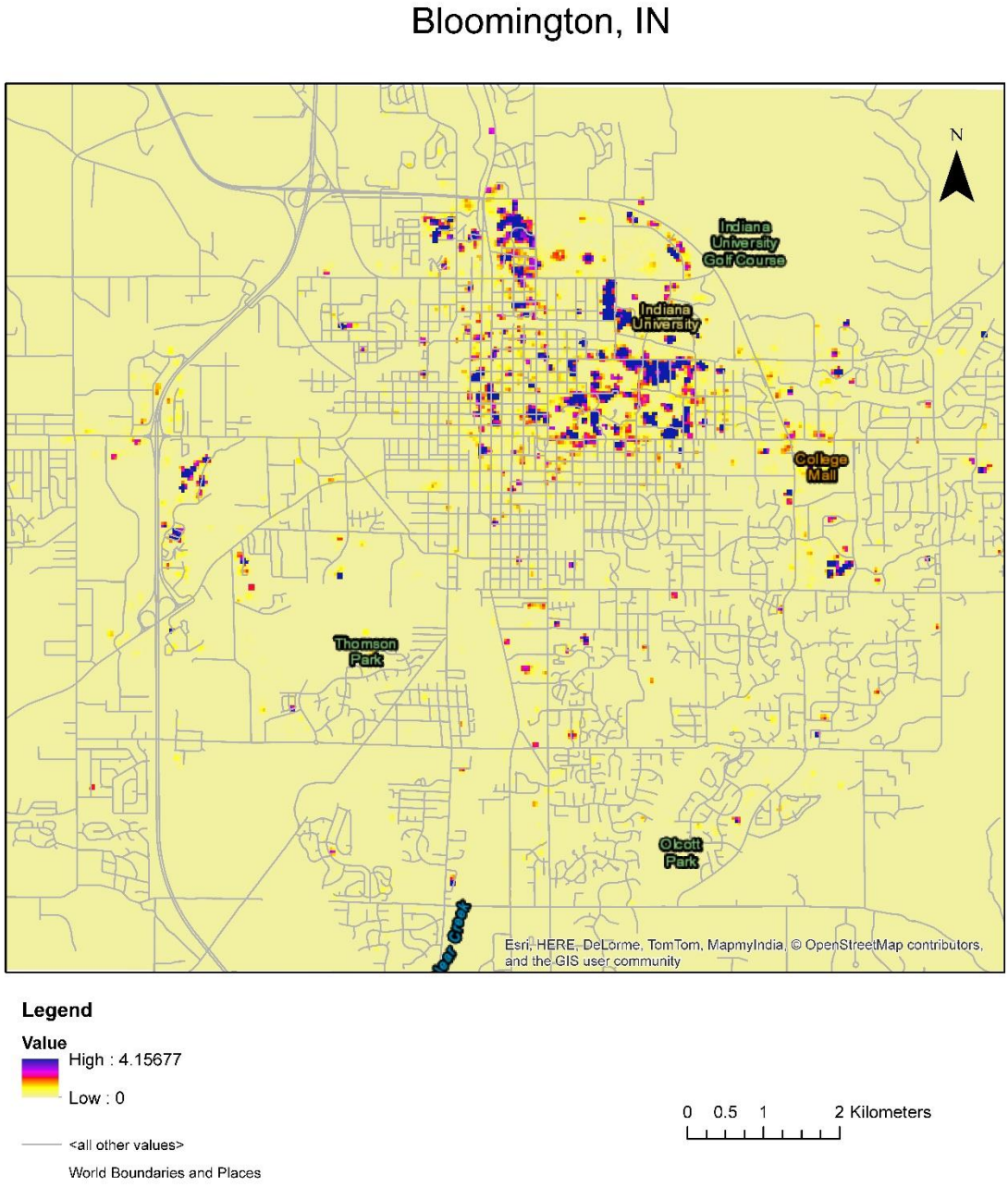


Figure 4.3 Point density raster of Bloomington, IN

The biggest tweet cluster in Ann Arbor was on the University of Michigan campuses including the north, central, and medical campuses. Concordia University also had a concentration of tweets (Figure 4.4). Besides the clusters on campuses, tweets were also concentrated in a few apartment complexes, such as the Pine Valley Apartments, the Ponds at Georgetown, and Park Place Apartments. However, Ann Arbor differed from West Lafayette and Bloomington in that significant clusters of tweets were found at the Briarwood Mall and the Georgetown Country Club (Figure 4.4).

Similar to Ann Arbor, the biggest tweet cluster in Columbus was on The Ohio State University campus (Figure 4.5). However, the downtown district also had a large cluster. A few apartment clusters in the north, the southwest and the south also had a higher concentration of tweets. Furthermore, clusters of tweets were found at Easton Town Center where there is a shopping mall and theaters. Compared to West Lafayette, Bloomington, and Ann Arbor, however, Columbus had more hot spots, which were spread around the city (Figure 4.5), which indicates that the active Twitter users were scattered throughout the cities, or the users traveled to different places in the city.

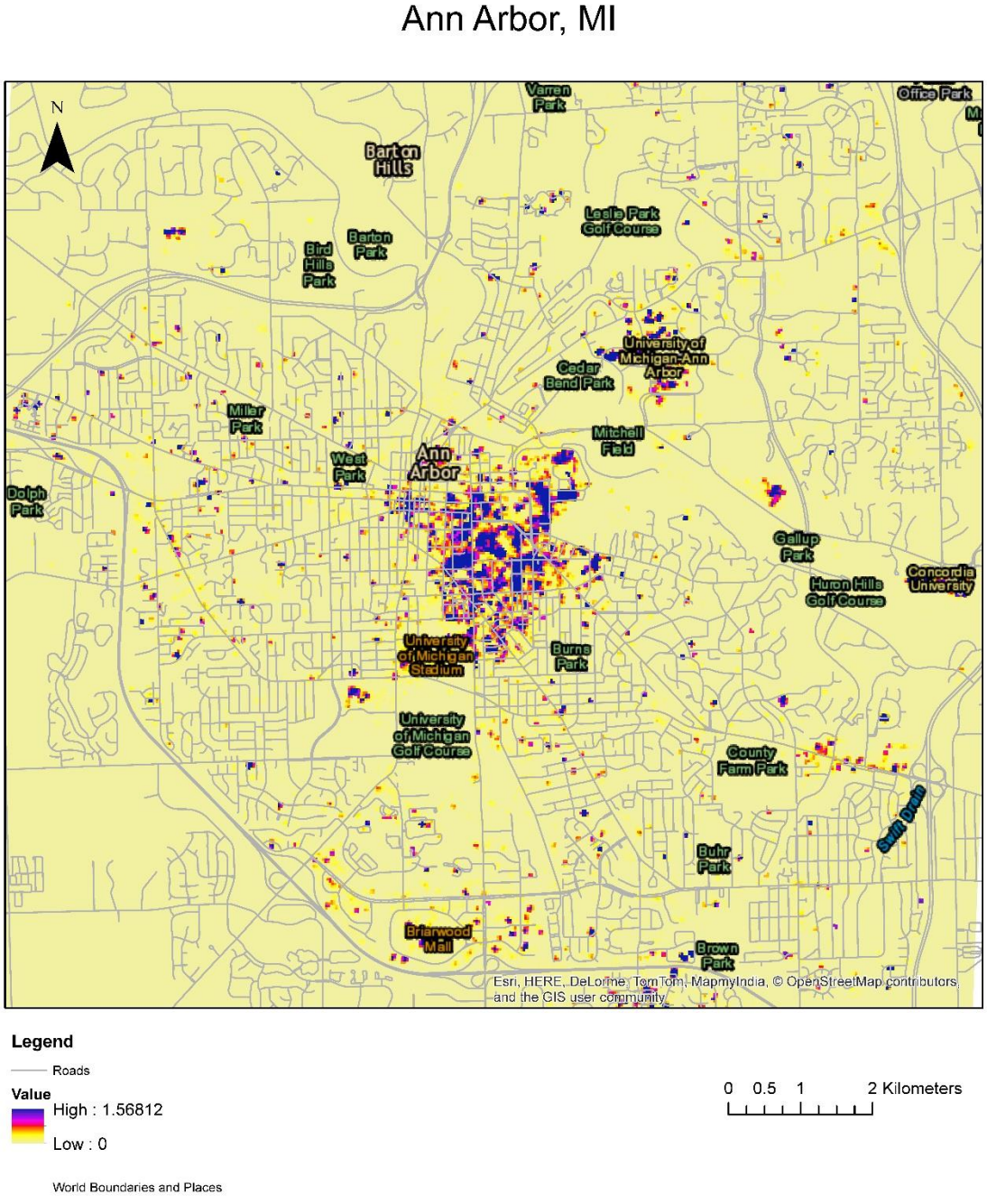


Figure 4.4 Point density raster of Ann Arbor, MI

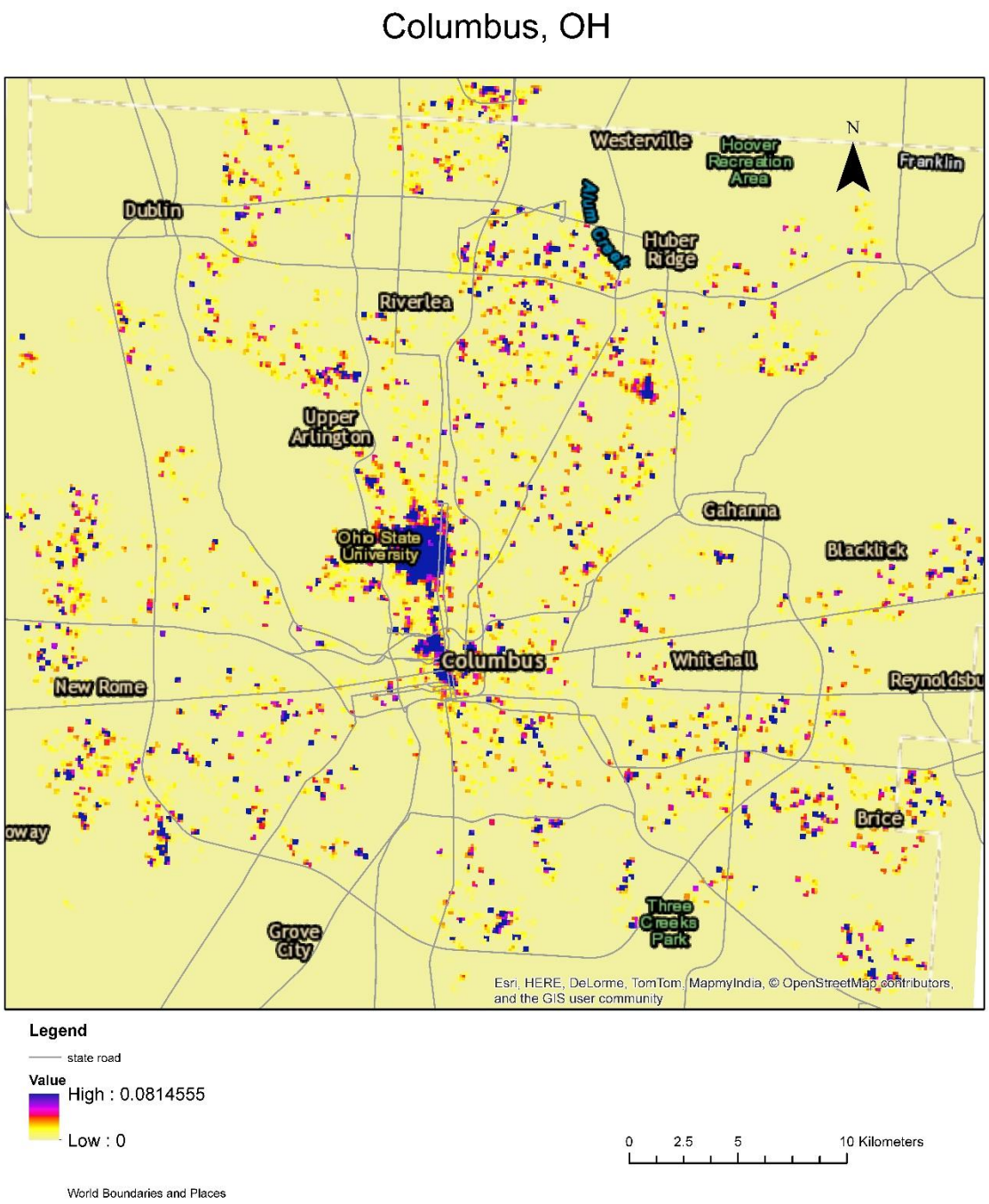


Figure 4.5 Point density raster of Columbus, OH

4.3 Spatial Clustering

To understand the spatial patterns of the tweets of individual users, the tweets of a few frequent Twitter users from Ann Arbor were plotted on the map. Several typical patterns of spatial distribution were found by considering the number of clusters in the user's tweets as well as the land use, time and content of the tweets: 1) work-home pattern with two main clusters, one probably the home of the user and the other the workplace or school (Figure 4.6); 2) work-road-home pattern with two main clusters at the workplace and home as well as a few tweets along the road between them (Figure 4.7); 3) work-home-short visit pattern with three main clusters (i.e. the home, the workplace and the place visited in a short time such as a weekend, but not frequently) (Figure 4.8); 4) multiple places frequently visited with more than three clusters whose purposes were hard to determine (Figure 4.9). It can be inferred that when the Twitter users had posted enough tweets, tweet clusters emerged that very likely were his/her home, workplace or a place of frequent visits. It was therefore important to determine the cluster locations in the users' tweets to understand their spatial pattern.

Work-Home

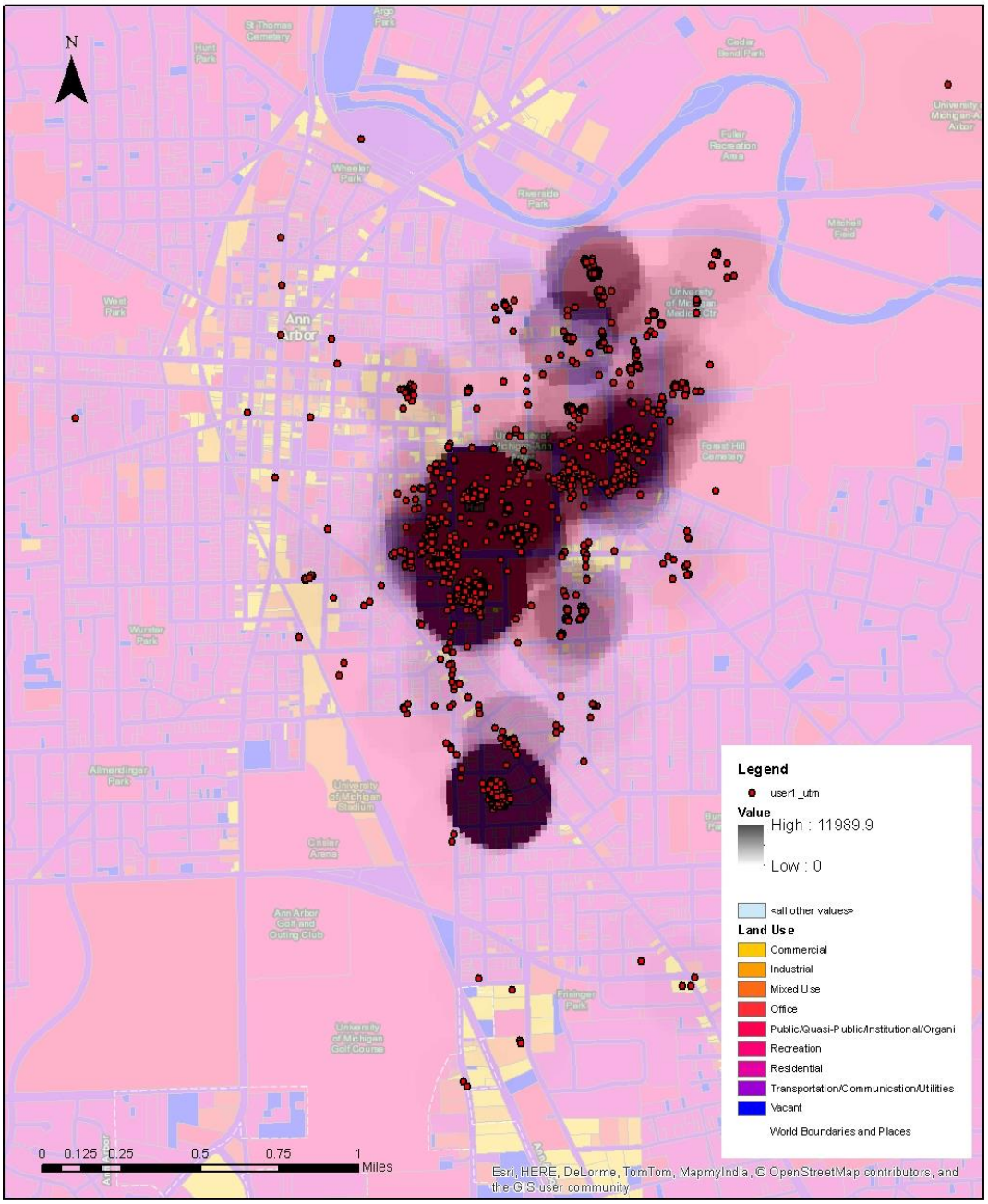


Figure 4.6 Work-Home pattern

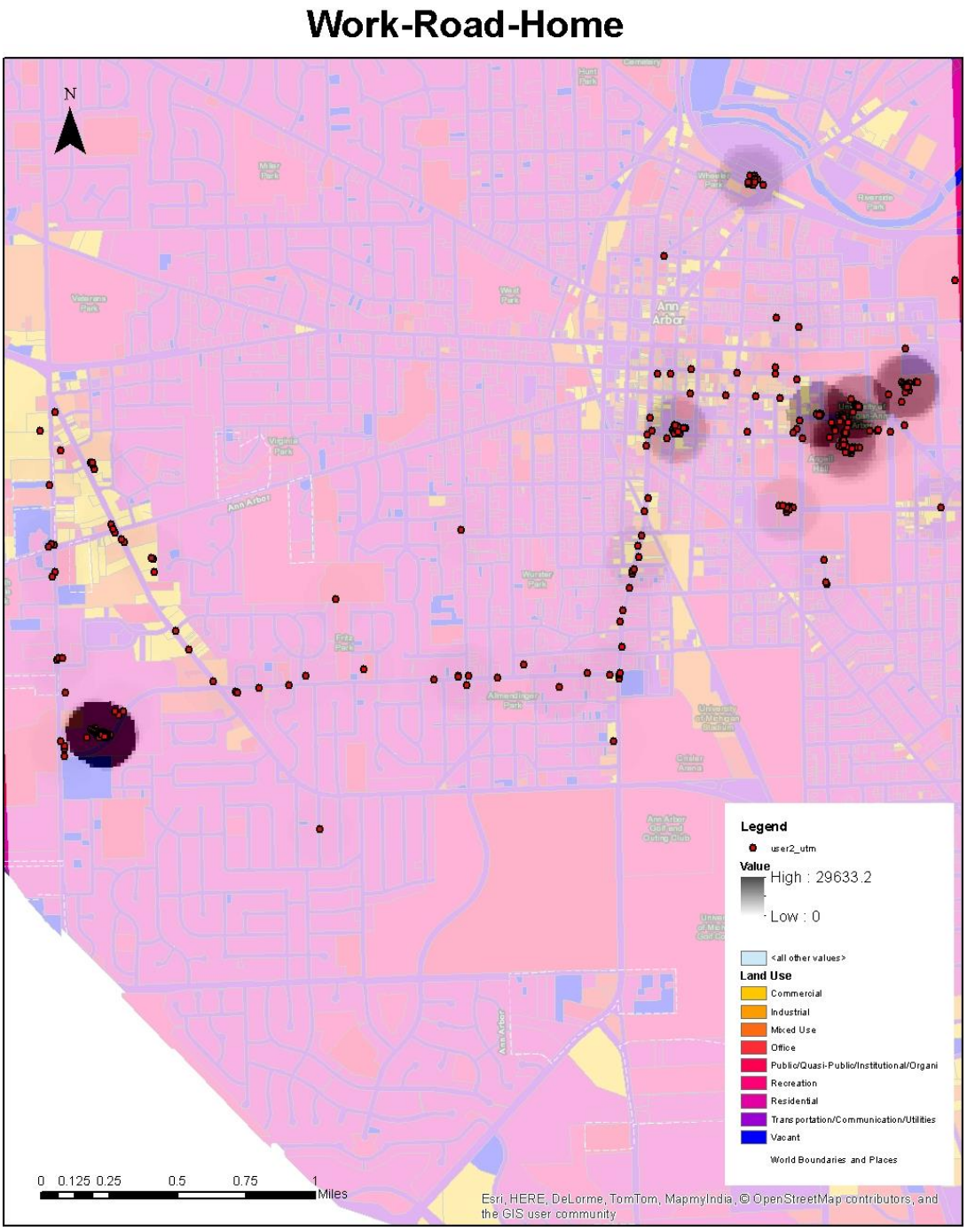


Figure 4.7 Work-Road-Home pattern

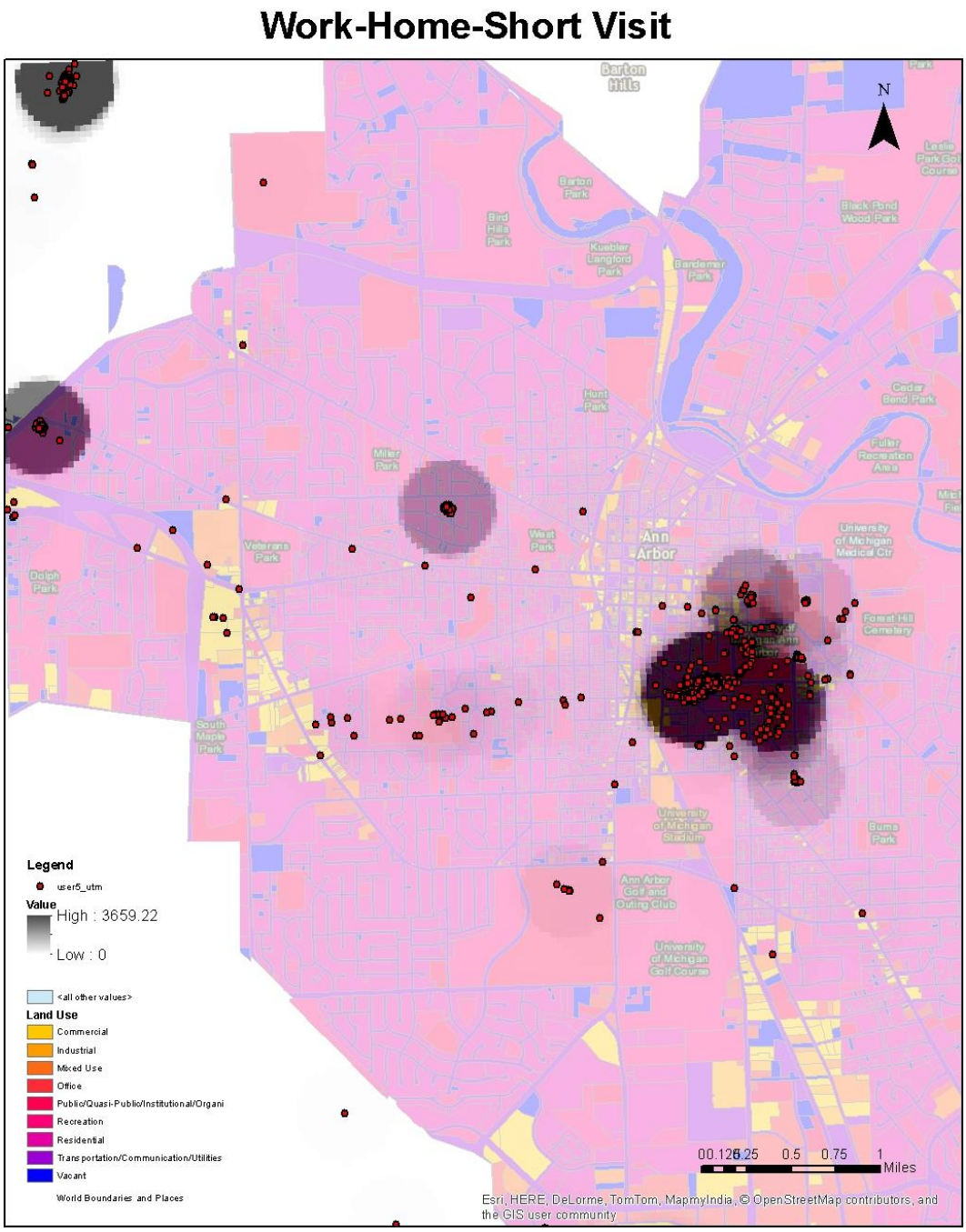


Figure 4.8 Work-Home-Short Visit pattern

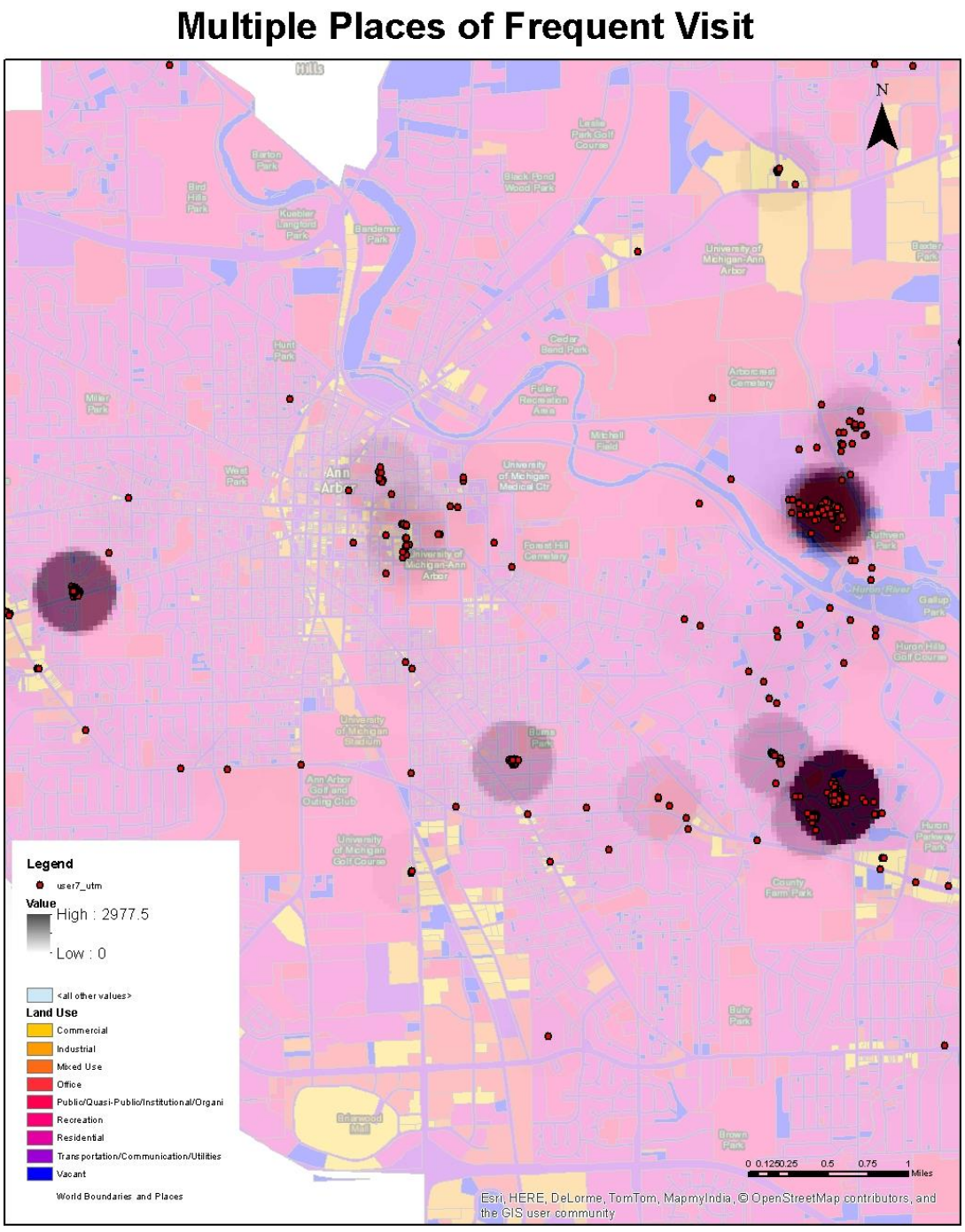


Figure 4.9 Multiple Places of Frequent Visit pattern

With the EM algorithm, the tweets of the individual users were clustered into five groups. However, some groups had very few tweets, so they were not considered as frequently visited places. Therefore, tweet groups with less than 5% of the individual's total tweets were excluded. Most of the users had two, three or four tweet clusters, while very few had one or five clusters (Figure 4.10).

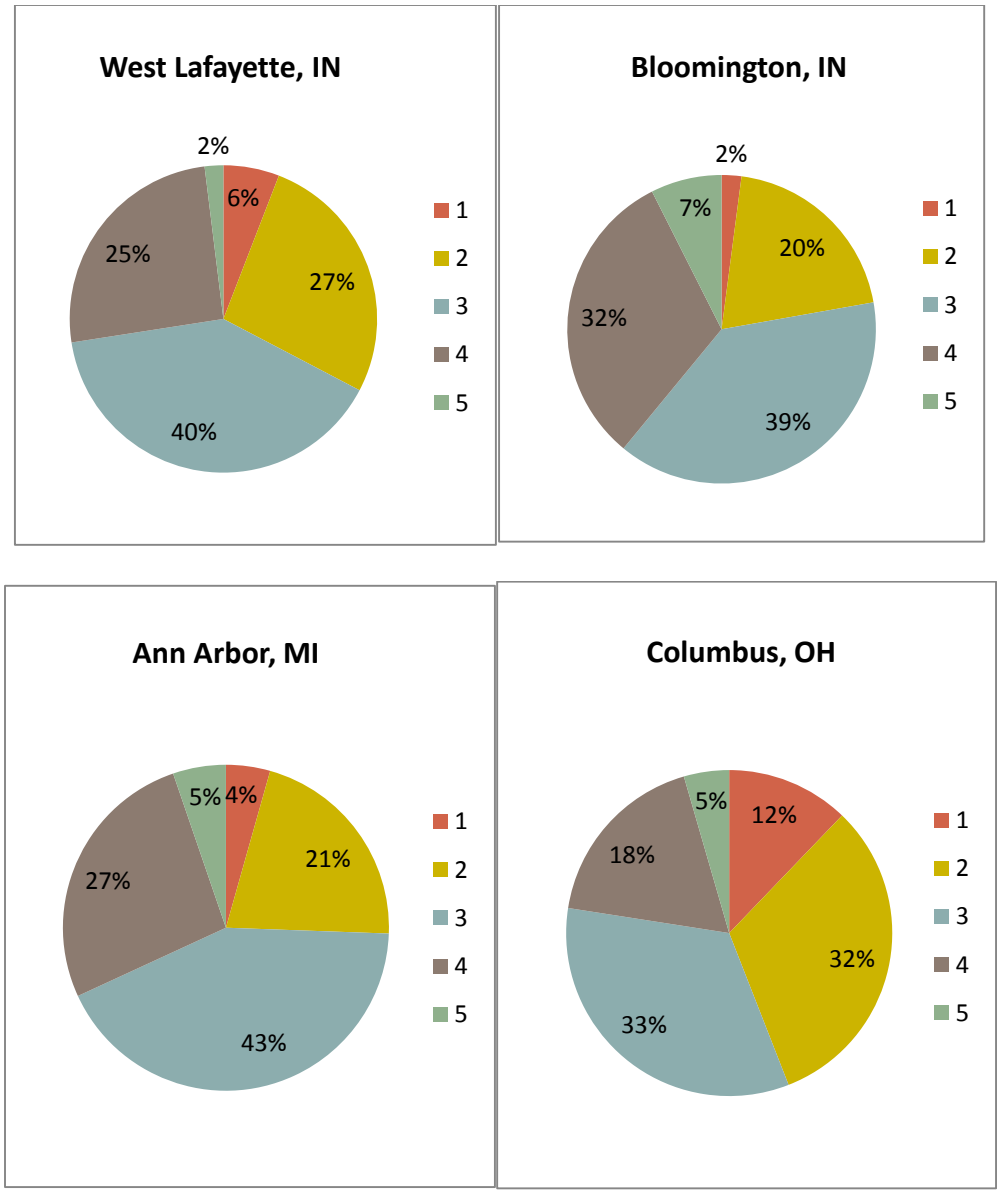


Figure 4.10 Number of users vs. number of spatial clusters

The average commuting distance for the users varied with the city in which they resided, and this analysis determined that the larger the city is, the longer the commute distance is. Users in West Lafayette had the smallest mean and median of the user's average commute distance while users in Columbus had the largest (Table 4.3). For the four cities, the mean values were larger than the median values (Table 4.3), indicating that more than half of the distances was smaller than the average distances. The city radius and median commute distance, and the city radius and mean commute distance are found linearly correlated. The radius was calculated as the squared root of the area divided by π if a city is assumed to be a circle. The coefficients of the two models indicate that the average commute distance is about 40% of the city radius. The R square values of these linear models were around 0.99, indicating that these linear models are likely to be capable to predict the commute distance from the area of the city (Figure 4.11).

Table 4.3 Summary statistics of average commute distance of frequent Twitter users

	Mean (km)	**Median (km)**
West Lafayette, IN	1.342	0.795
Bloomington, IN	1.651	1.260
Ann Arbor, MI	1.892	1.556
Columbus, OH	5.763	4.879

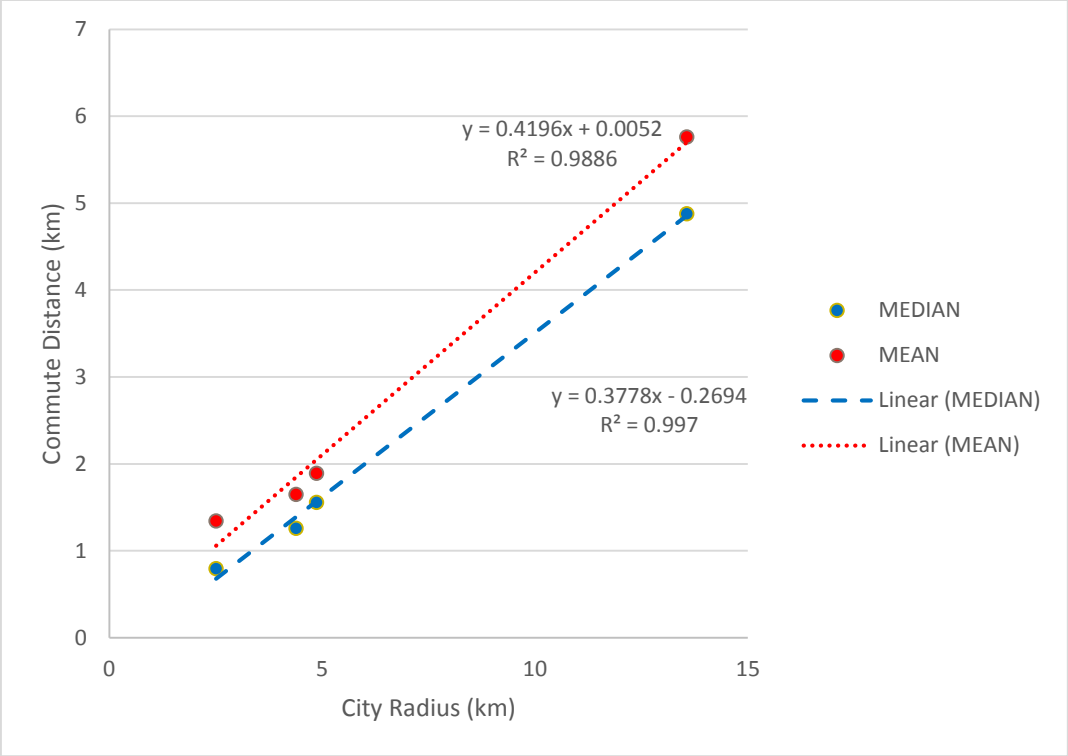


Figure 4.11 Relationship between city radius and median commute distance as well as mean commute distance

The distribution of average commute distances are all skewed sharply to the larger distances, meaning the people living away from campus or workplace drop considerably. In West Lafayette, most users commuted less than 1km (Figure 4.12) with a mean of 1.34km and a median of 0.78km (Table 4.3), inferring that the local residents had short commutes to work or school. In Bloomington, the majority of the users' average commute distance were less than 3km (Figure 4.13), and the mean was 1.65km with a median of 1.26km. The average commute distance of the users in Ann Arbor ranged from around 0.5km to 4km (Figure 4.14) with the mean 1.89km and the median 1.5km (Table 4.3). Although the mean and median commute distances for Bloomington and Ann Arbor were similar, the values aggregated around the median for Bloomington, while the values for Ann Arbor were more evenly distributed (Figure 4.14). The commute distances of users in

Columbus were much longer than that in the other cities (Figure 4.15), with a mean of 5.76km and a median of 4.87km (Table 4.3), likely due to the large size of this metropolitan city, and its zoning characteristics as well as the interstate and highway networks that connect the downtown district with neighborhood areas.

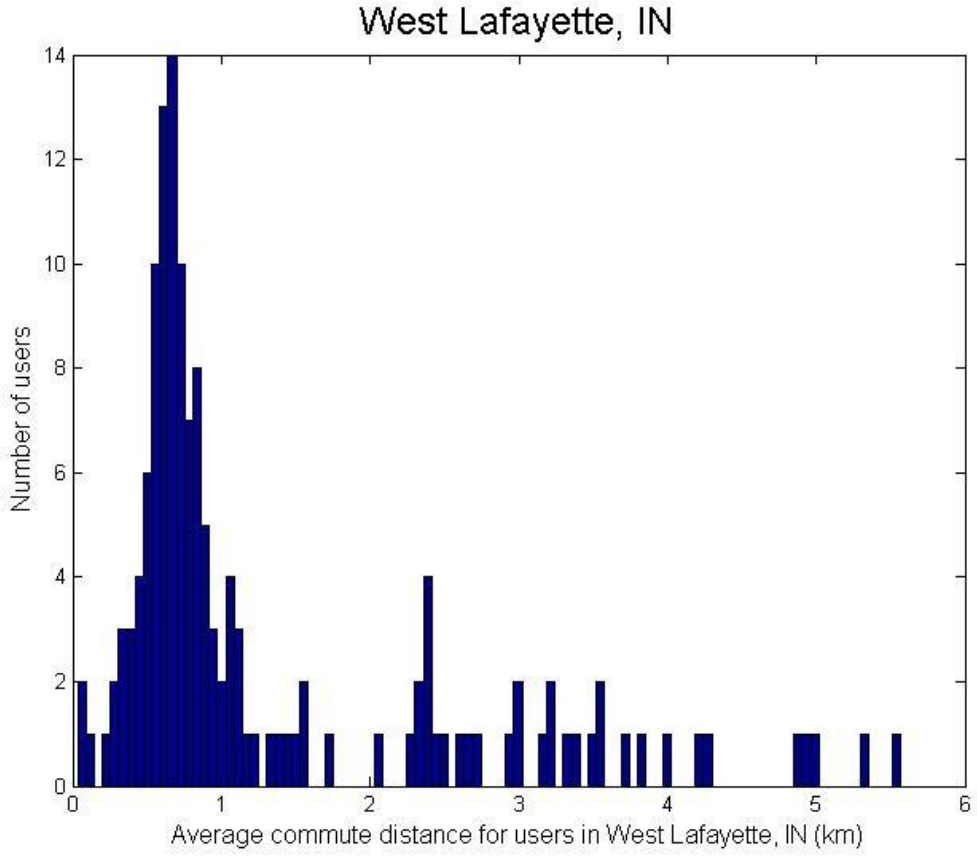


Figure 4.12 Number of users against average commute distance of frequent Twitter users in West Lafayette, IN

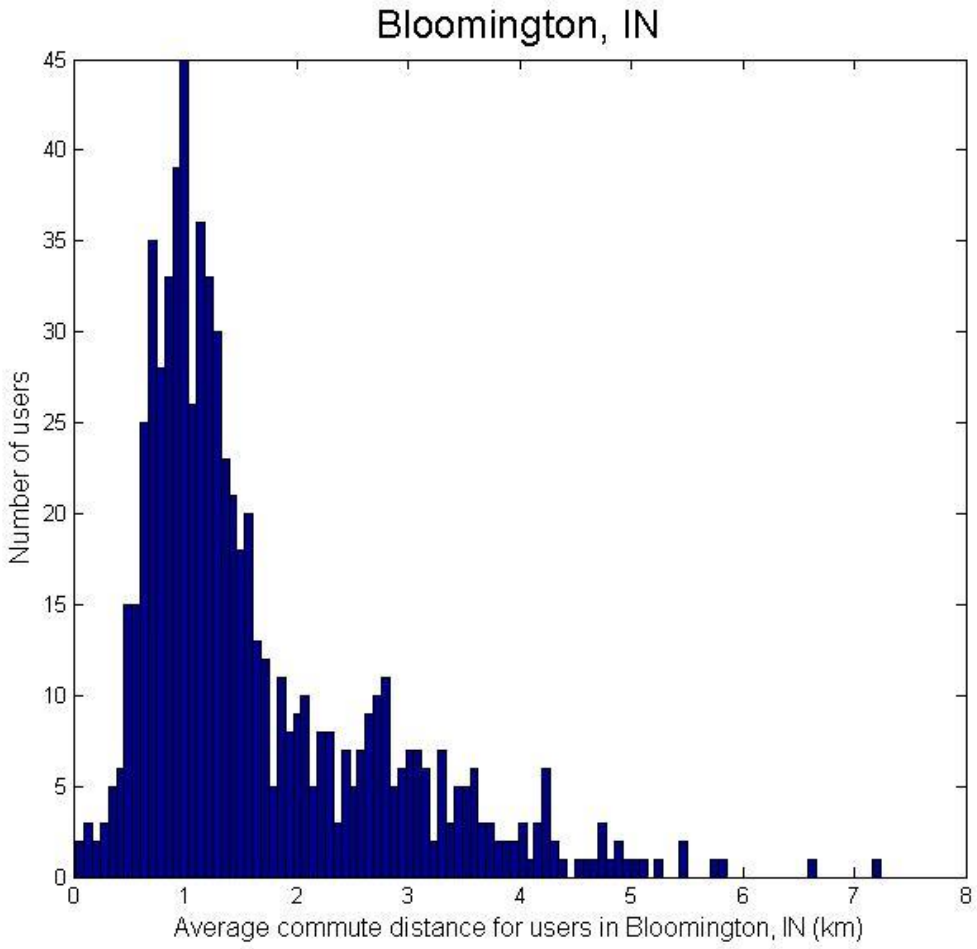


Figure 4.13 Number of users against average commute distance of frequent Twitter users in Bloomington, IN

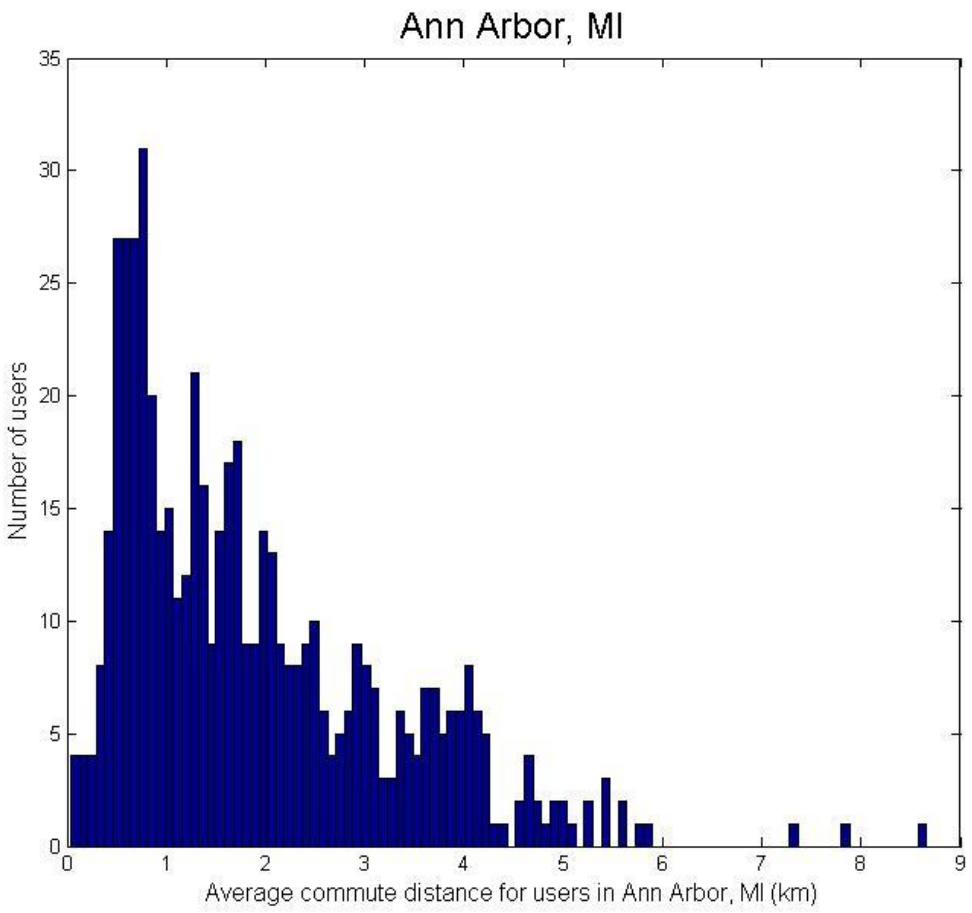


Figure 4.14 Number of users against average commute distance of frequent Twitter users in Ann Arbor, MI

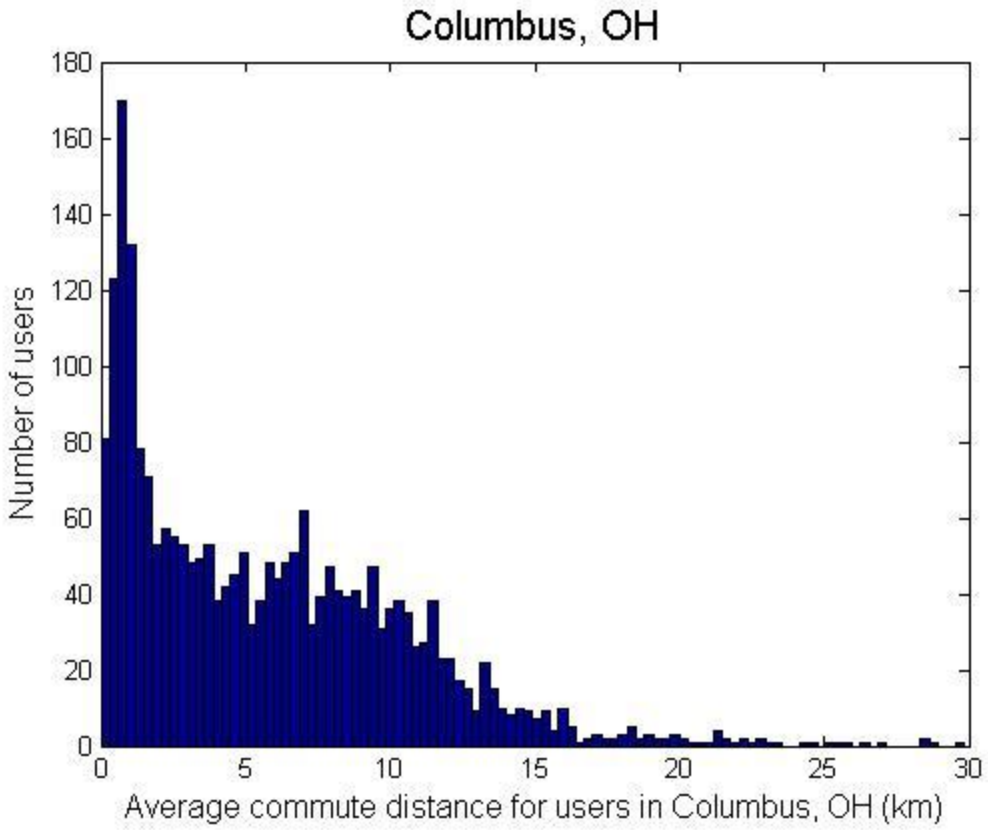


Figure 4.15 Number of users against average commute distance of frequent Twitter users in Columbus, OH

4.4 Temporal Analysis

4.4.1 By hour of a day

Tweets in all four study areas had similar hourly patterns (Figure 4.16 ~ Figure 4.19). The number of tweets, as well as the number of users increased around 6:00 am (Figure 4.16 ~ Figure 4.19) when people were awakening and getting ready for school or work. The tweets continued to grow in all four cities until 12:00 pm (Figure 4.16 ~ Figure 4.19). For West Lafayette, the increase continued until 1:00 pm when it hit at a peak and then began to decrease until 4:00 pm; in the meantime, the number of tweets in the other

three study areas remained stable (Figure 4.16). After 4:00 pm, the tweets began to rise again until around 9:00 pm, when they reached a peak (Figure 4.16 ~ Figure 4.19). This evening time period was likely when people returned from work or study and taking care of the household or relaxing. For West Lafayette, Bloomington, and Ann Arbor, the total tweets around 9:00 pm, the peak time, comprised about 6% of all the tweets (Figure 4.16 ~ Figure 4.18). However, for Columbus, the tweets at the peak time were almost 9% of the total tweets (Figure 4.19), indicating that they may have had more variations in their routines compared to others. It is also noted that, compared to Columbus, the number of users in West Lafayette started to decline at night, while the number of users in Bloomington remained still, implying that the Twitter users in Columbus were more active at night than those in other cities, which was possibly due to the size of Columbus and the variety of activities available there. After 9:00 pm the tweet counts declined again (Figure 4.16 ~ Figure 4.19) until 12:00 am when most people were probably getting ready to go to sleep. The number of tweets continued to decrease until around 4:00~5:00 am, which it reached a valley (Figure 4.16 ~ Figure 4.19). From the above statistics, it was concluded that Twitter user in these four cities were active from 10:00 am to 12:00am.

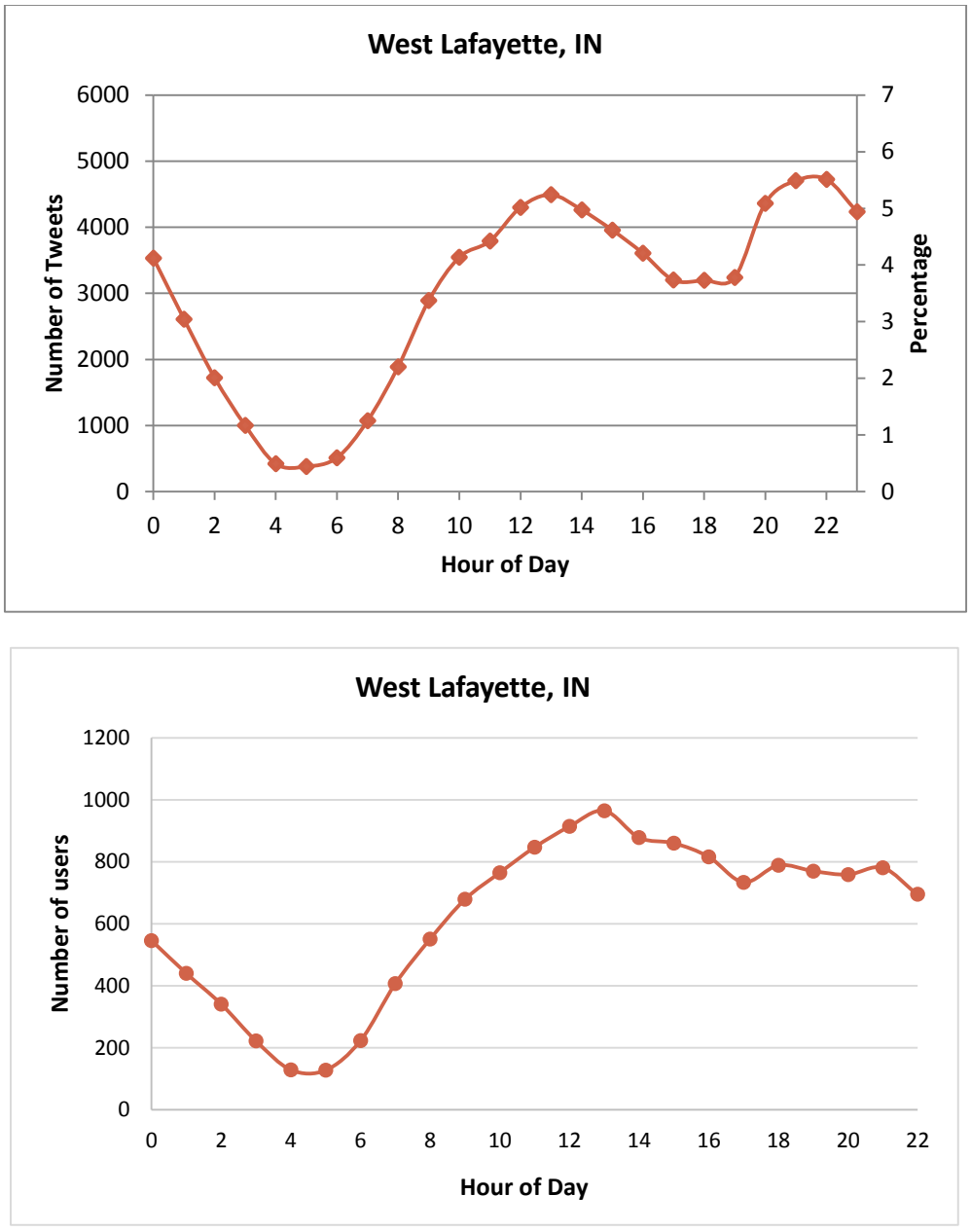
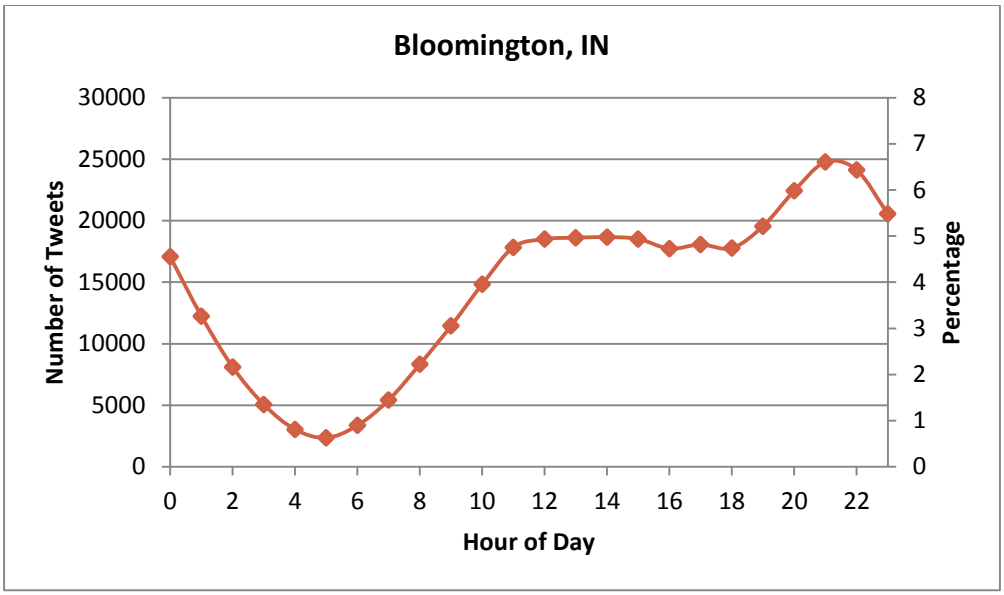


Figure 4.16 Number of tweets and users in each hour of day in West Lafayette, IN



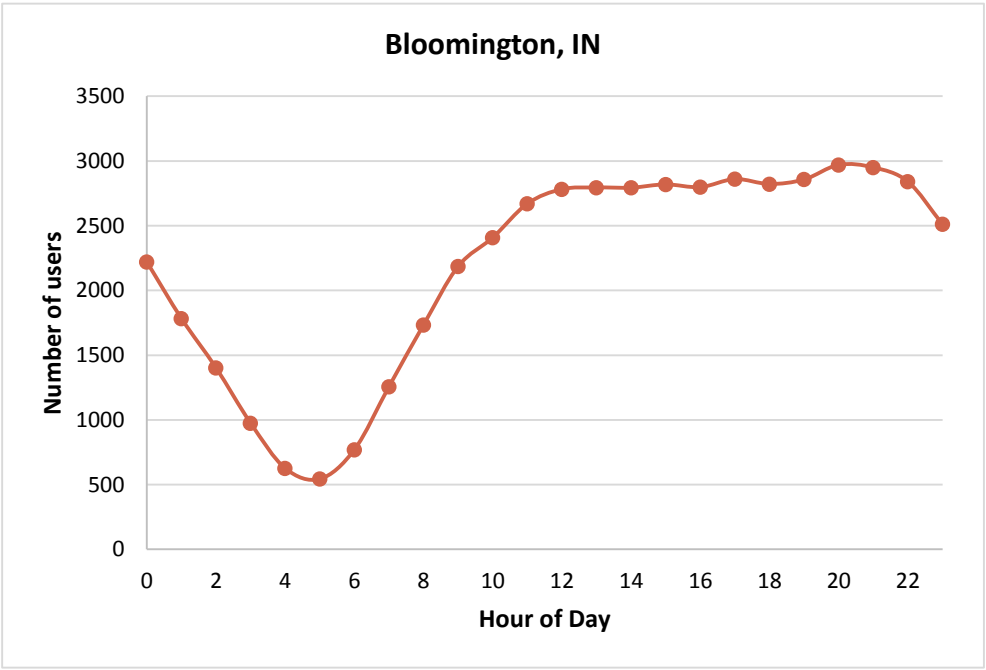
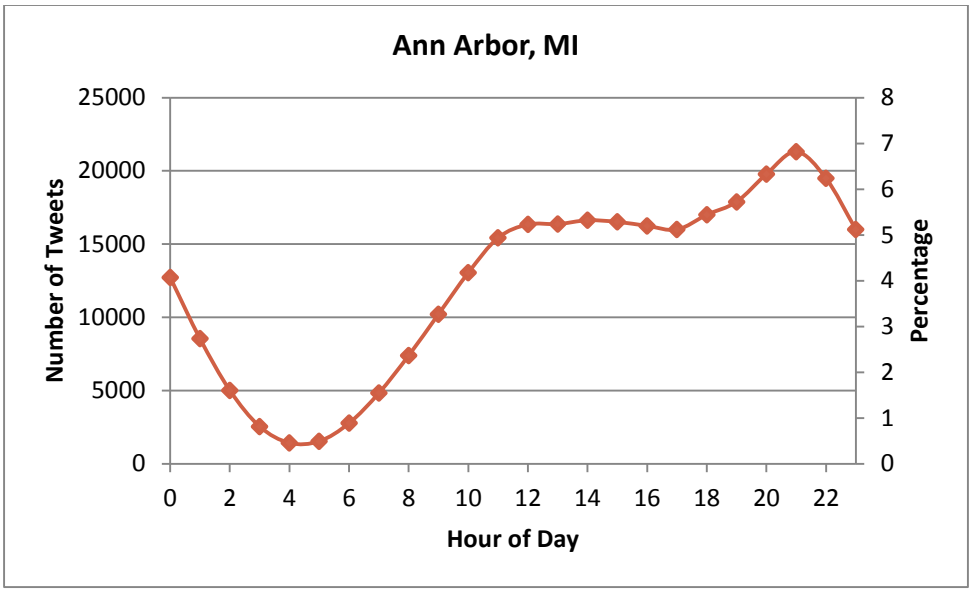


Figure 4.17 Number of tweets and users in each hour of day in Bloomington, IN



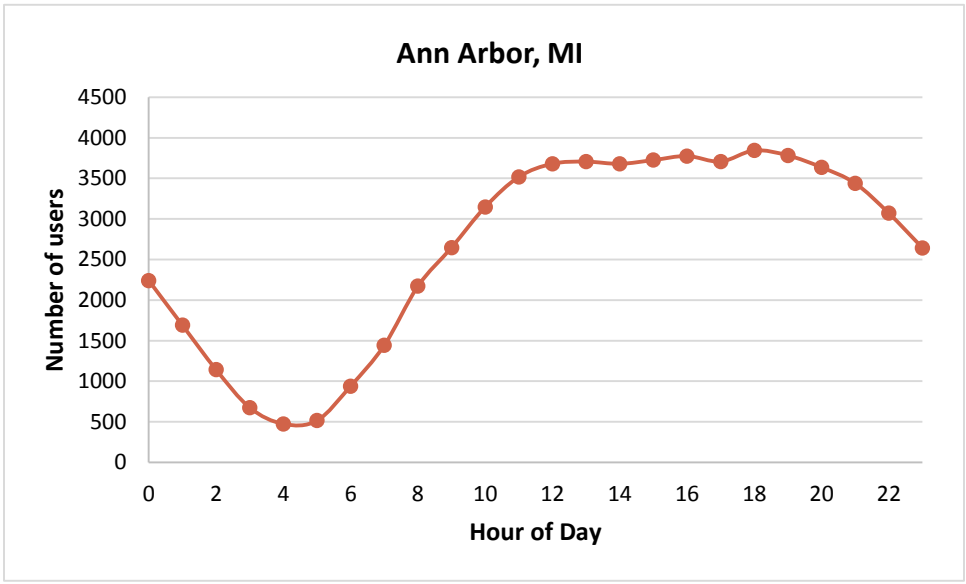


Figure 4.18 Number of tweets and users in each hour of day in Ann Arbor, MI

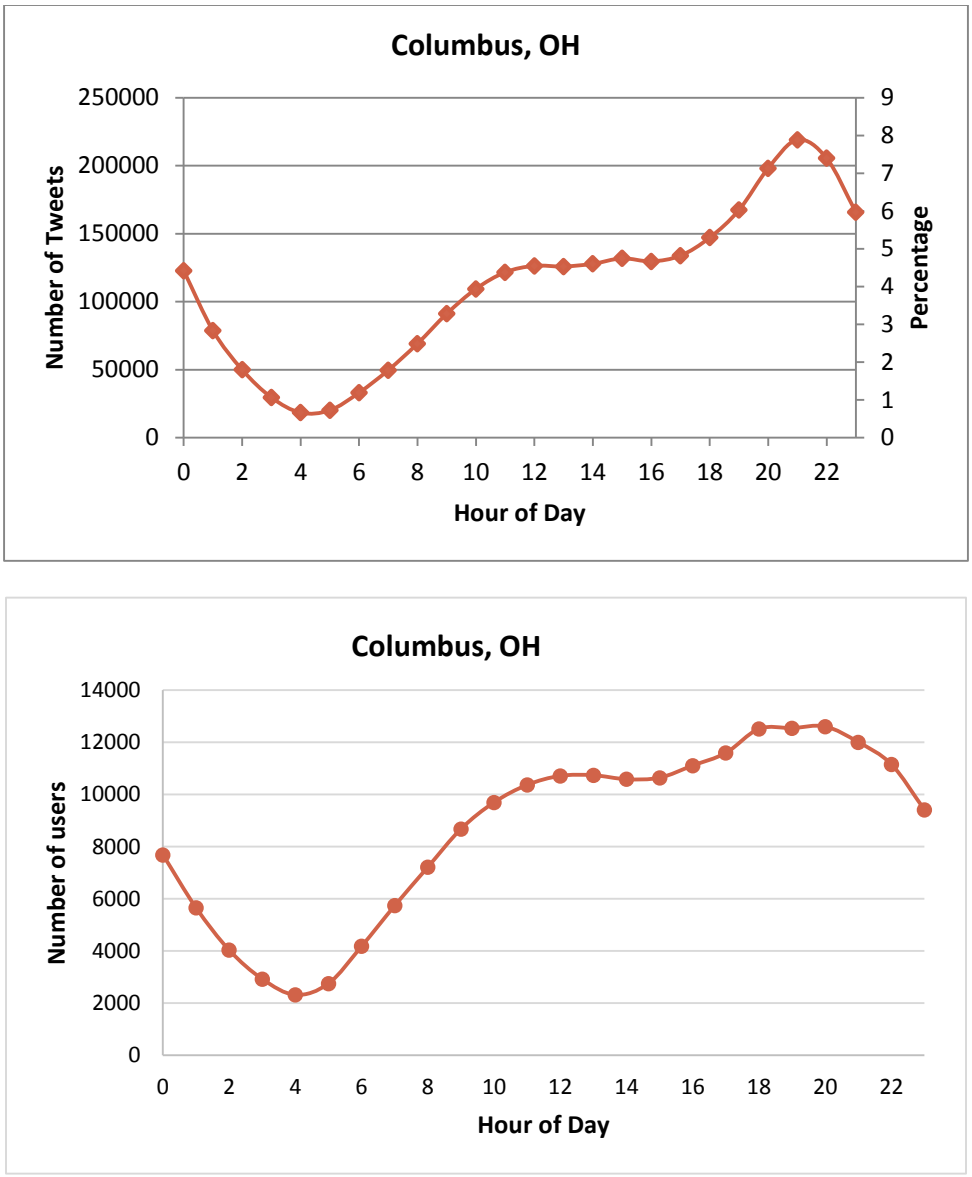


Figure 4.19 Number of tweets and users in each hour of day in Columbus, OH

4.4.2 By day of week

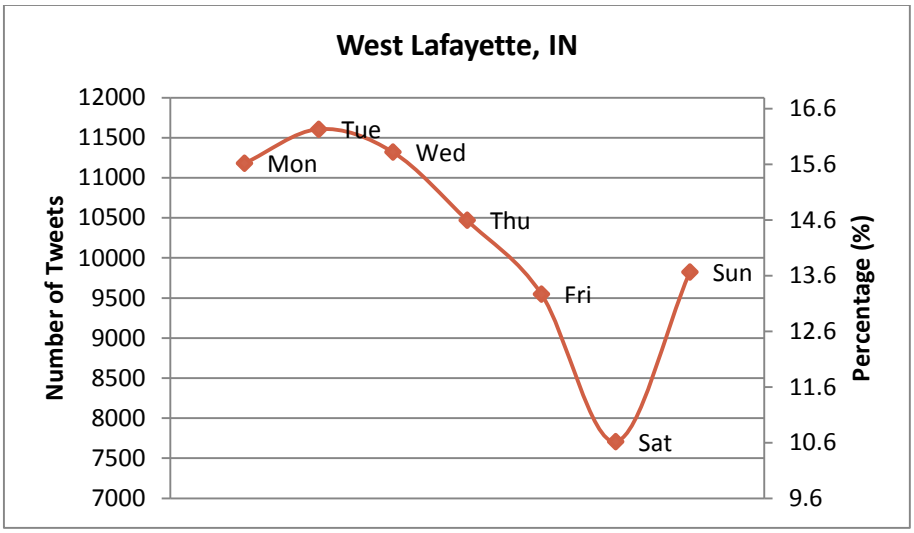
There were more Twitter users during weekends than weekdays. The average number of tweets per weekday was smaller than the average for weekend. Also, the 'weekday' Twitter user group and the 'weekend' Twitter user group had about 15 ~ 20%

overlap (Table 4.4), implying that most people only tweet either on weekdays or weekends. The reason behind this might be the tweeting preference of users, or users leaving or coming to town on weekends.

The daily pattern of tweets varied with the city, which differed from the similar hourly patterns determined for all four study areas (Figure 4.20 ~ Figure 4.23). In West Lafayette, more tweets were posted on weekdays than weekends, and Saturday had the lowest number of tweets (Figure 4.20). However, the number of users on Saturday rose at a peak period (Figure 4.20). Wednesday, Thursday and Sunday had relatively more tweets than the other days of the weeks in Bloomington; and similar to West Lafayette, Saturday had the least tweets and the most users (Figure 4.21). As significantly more users were active in tweeting during weekends than weekdays, the average number of tweets posted during weekend was lower than on weekdays. In Ann Arbor and Columbus, however, contrary to West Lafayette and Bloomington, there were more tweets on the weekends (Figure 4.22 and Figure 4.23). The number of tweets reached a valley on Tuesday and rose to a peak on Sunday (Figure 4.22 and Figure 23), which again may be due to the relatively large size of Ann Arbor and Columbus and their larger offerings of entertainment venues and major events that might keep residents in town during the weekends and attract out of town visitors as well. However, the trends in the number of users in these two cities were similar to West Lafayette and Bloomington.

Table 4.4 Number of users on weekdays and weekends

# users	West Lafayette	Bloomington	Ann Arbor	Columbus
# users who tweet both on weekdays and weekends (d)	391	1456	1613	7401
# users who tweet only on weekdays (A)	2137	6437	11302	41276
d/A (%)	18.3	22.6	14.3	17.9
# users who tweet only on weekends (E)	1841	5766	9219	35184
d/E (%)	21.2	25.2	17.5	21.0



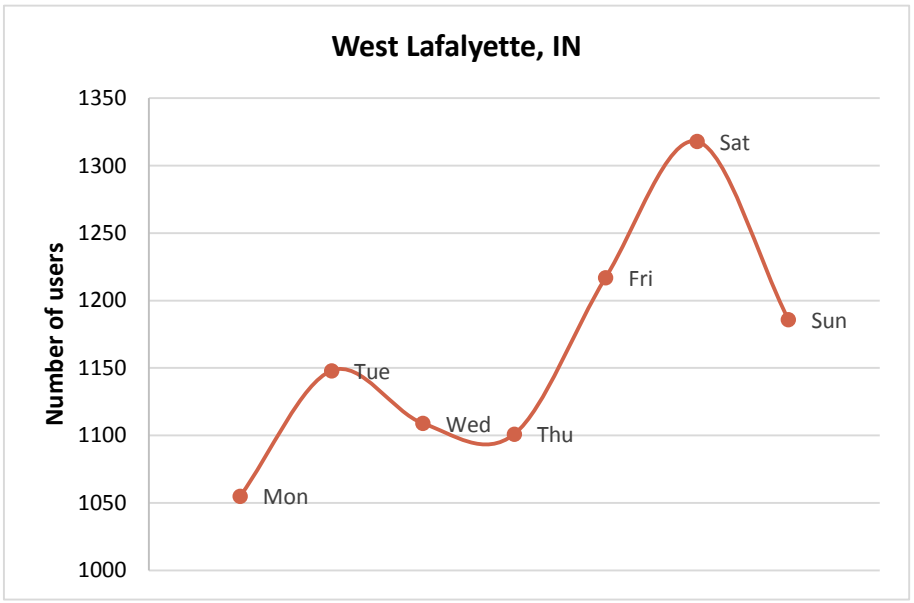
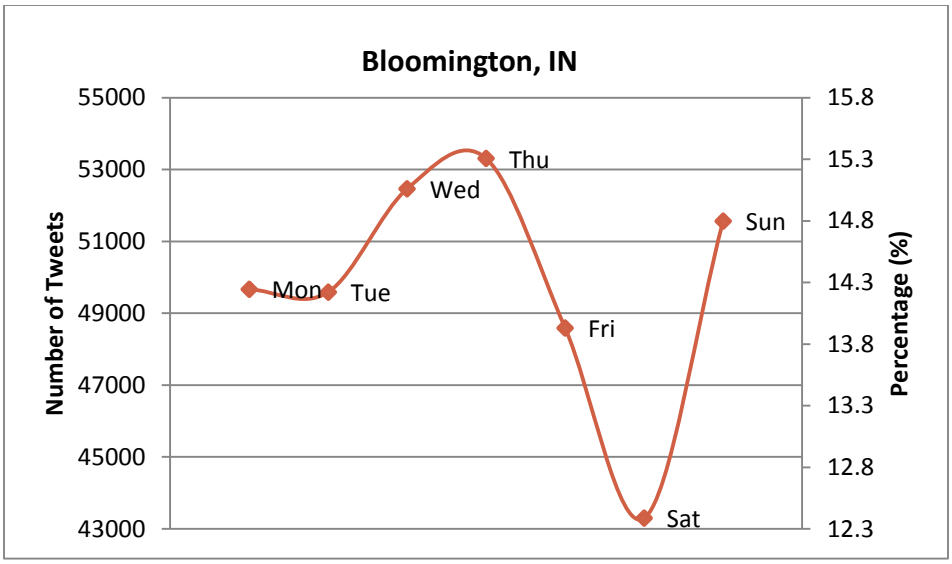


Figure 4.20 Number of tweets and users in each day of week in West Lafayette, IN



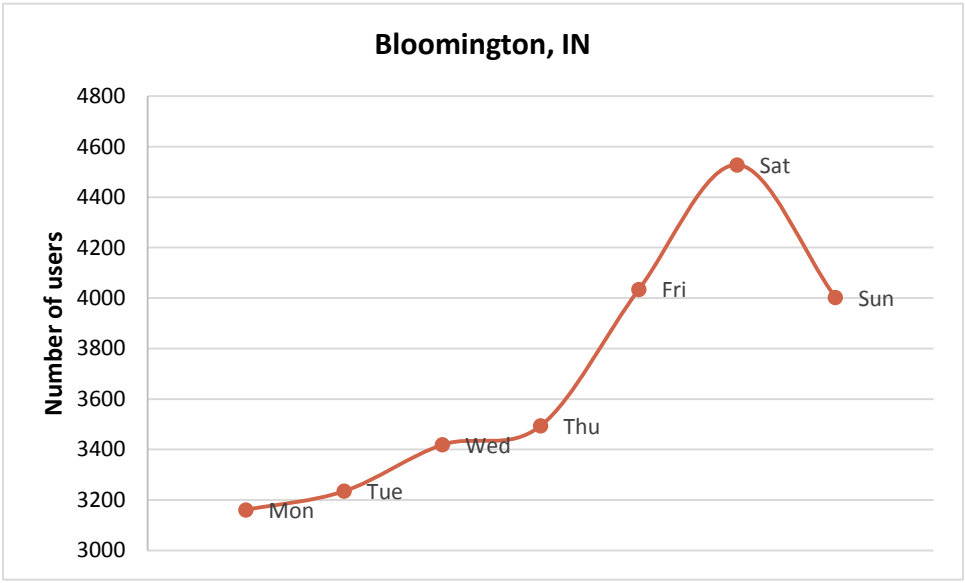


Figure 4.21 Number of tweets and users in each day of week in Bloomington, IN

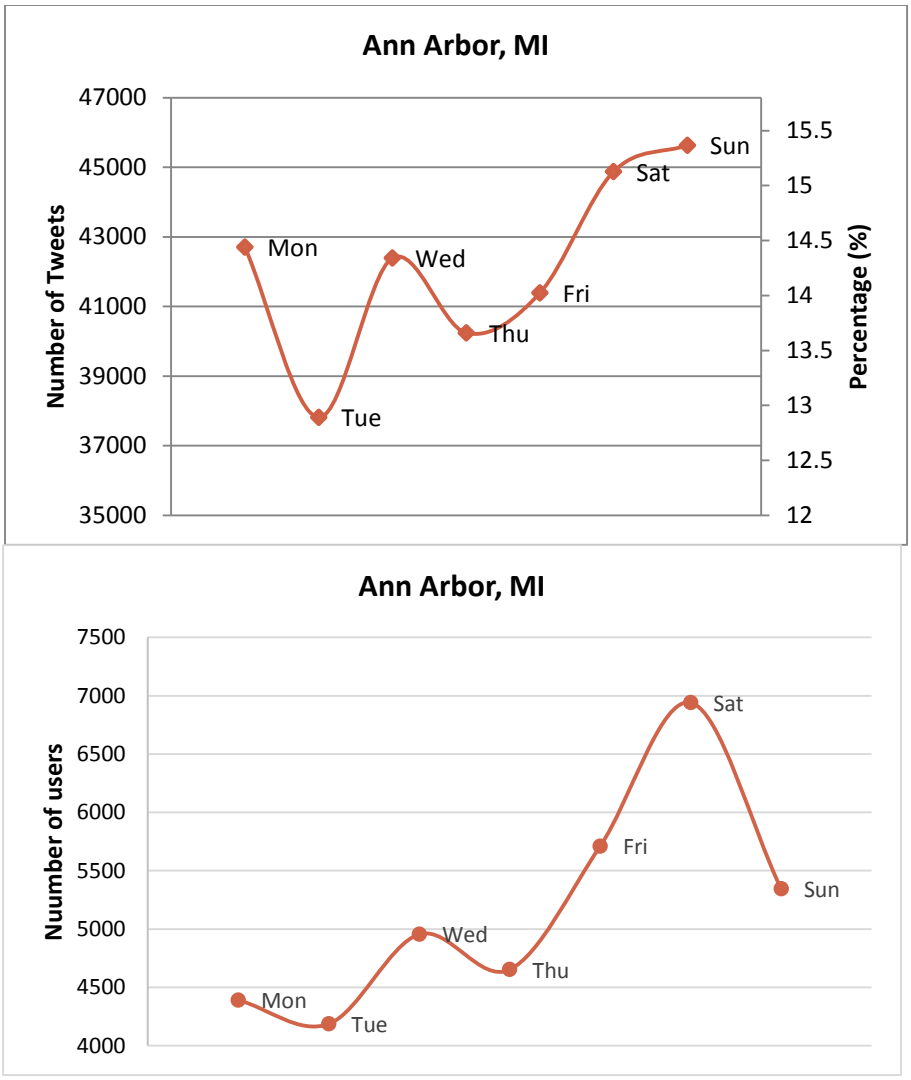


Figure 4.22 Number of tweets and users in each day of week in Ann Arbor, MI

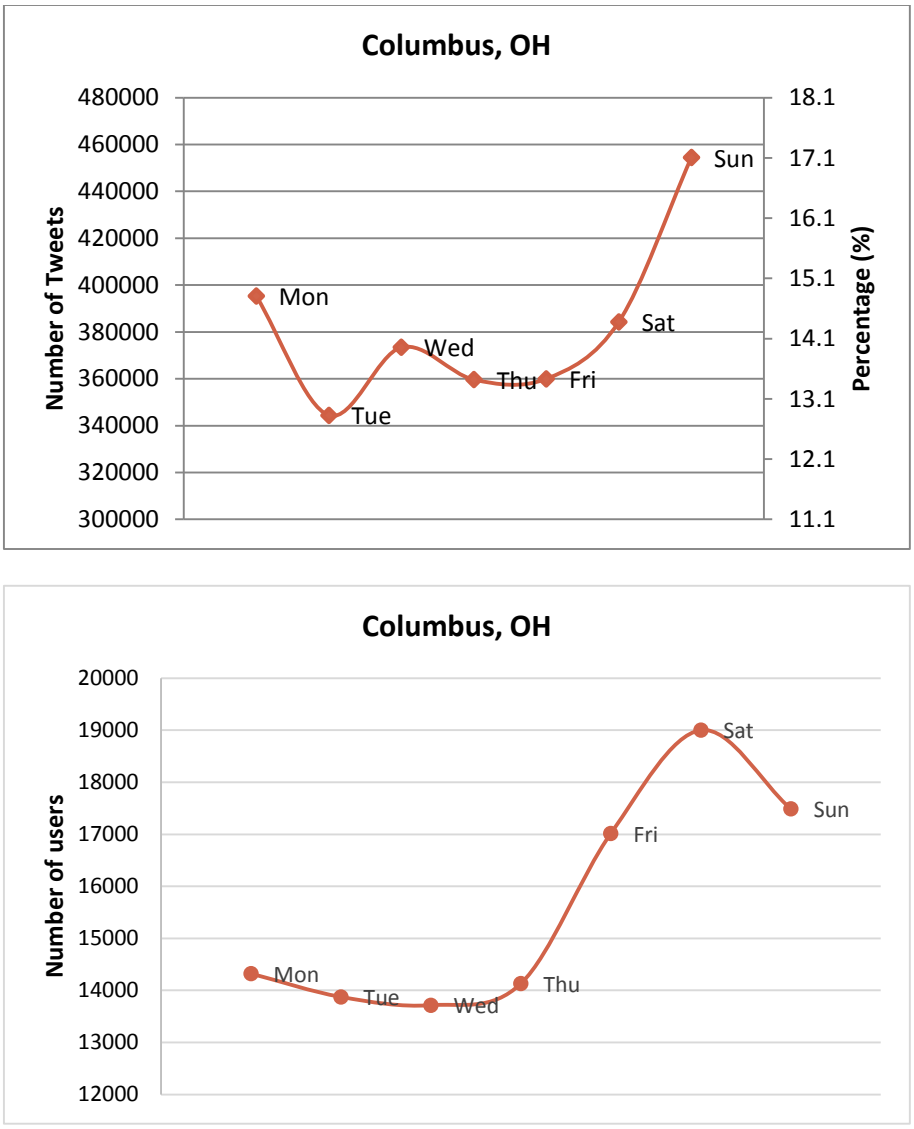


Figure 4.23 Number of tweets in each day of week in Columbus, OH

4.4.3 By the month

Bloomington, Ann Arbor, and Columbus had similar patterns for the number of tweets each month (Figure 4.24 ~ Figure 4.26). More tweets were found in December, and the number began to decline in January and February (Figure 4.24 ~ Figure 4.26) most likely due to the holidays and the semester ending in December as well as the cold weather

in January and February when people tend to stay at home more. Then the number started increasing in March and April (Figure 4.24 ~ Figure 4.26) due to the coming of spring and more activities. Then the tweet count drastically dropped in May (Figure 4.24 ~ Figure 4.26) probably due to the departure of students in mid-May. The difference in number of tweets between April and May was smaller in Columbus than in other cities (Figure 4.24 ~ Figure 4.26), implying that the impact of students' leaving school had the least impact in Columbus on Twitter usage. The trend in the number of users in Columbus, however, was almost identical with the trends in the other three cities (Figure 4.24 ~ Figure 4.26). West Lafayette had a very different pattern of the number of tweets per month from the others (Figure 4.23), namely, there were more tweets in January and February than in December (Figure 4.23). Also, there were relatively fewer Twitter users but more tweets in January, indicating that the users tended to tweet more during the holidays. The tweet counts started to decline and reached a valley in May (Figure 4.23), likely due to the Purdue spring semester starting in mid-January and ending in early May.

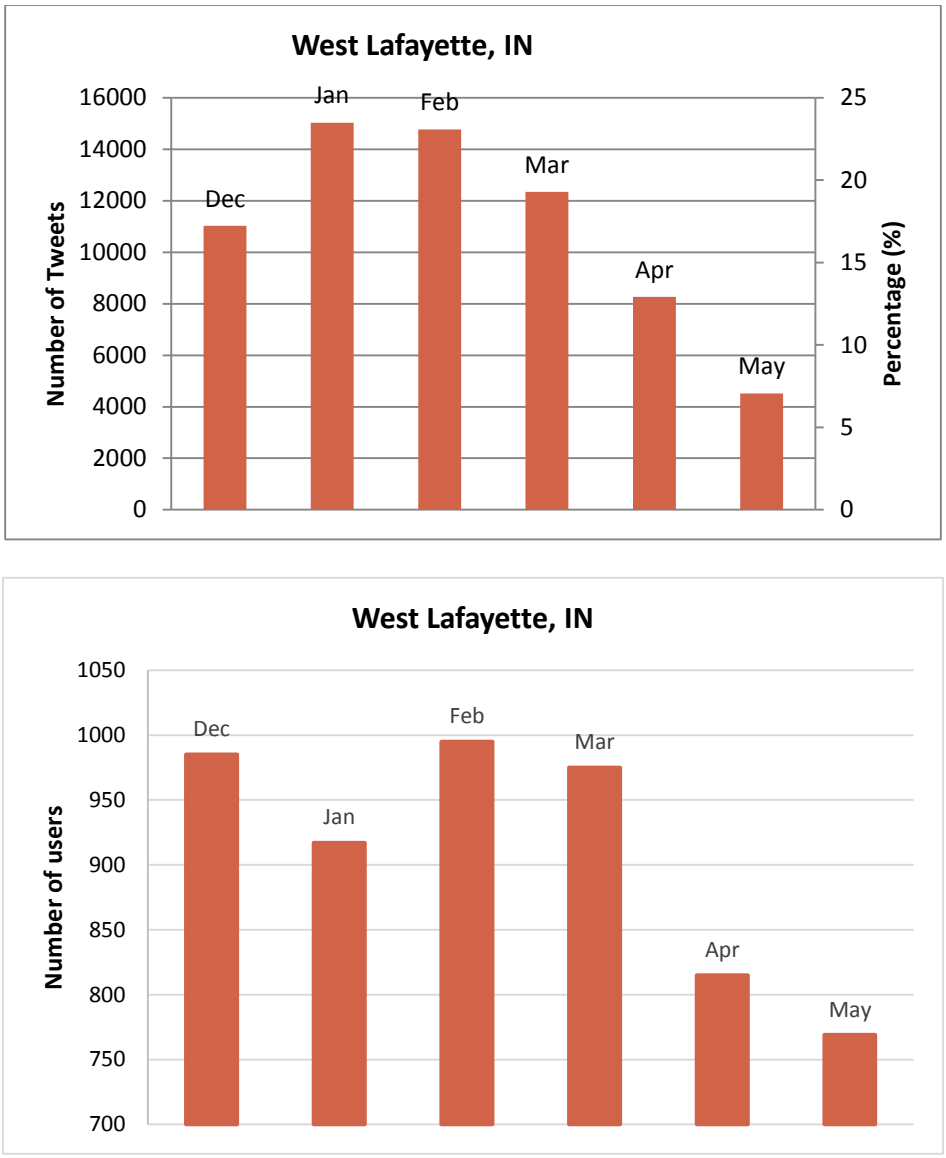


Figure 4.24 Number of tweets and users in each month in West Lafayette, IN

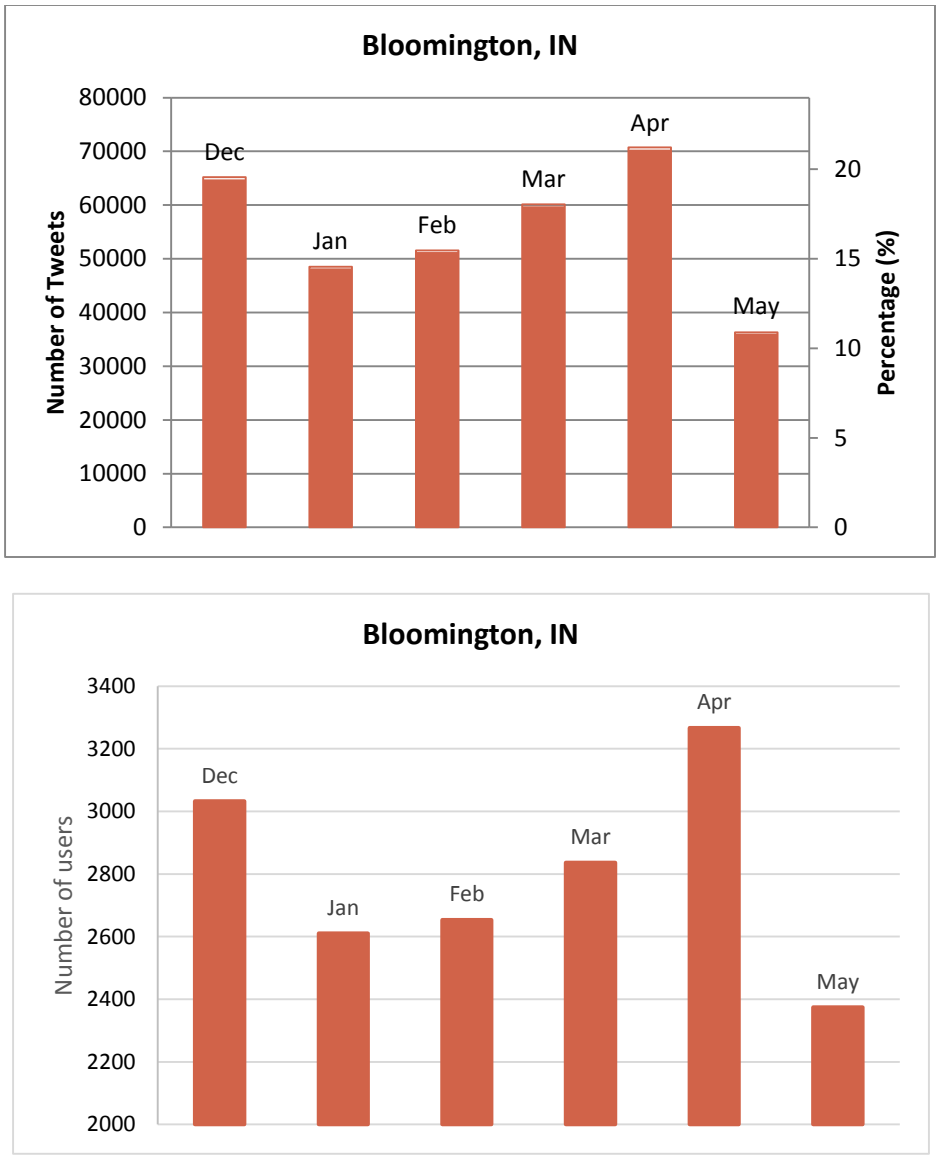


Figure 4.25 Number of tweets and users in each month in Bloomington, IN

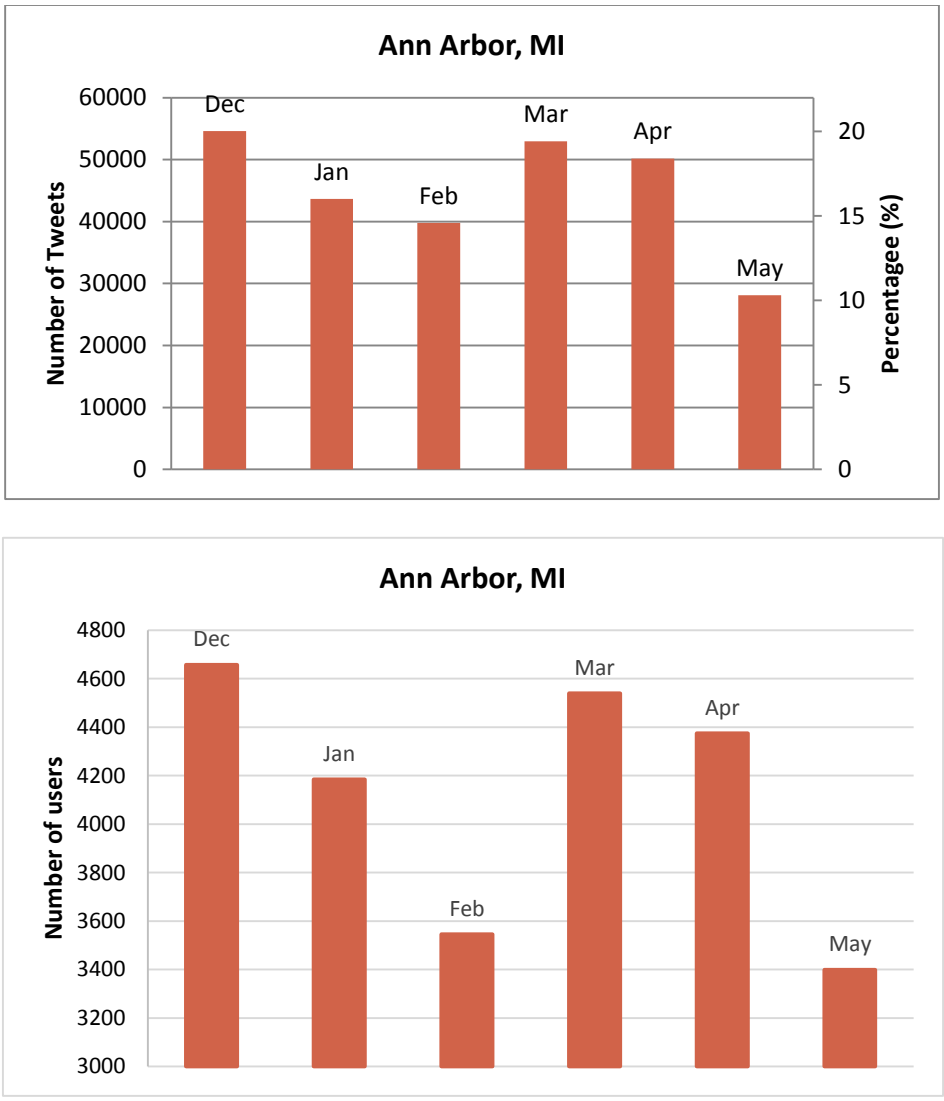


Figure 4.26 Number of tweets and users in each month in Ann Arbor, MI

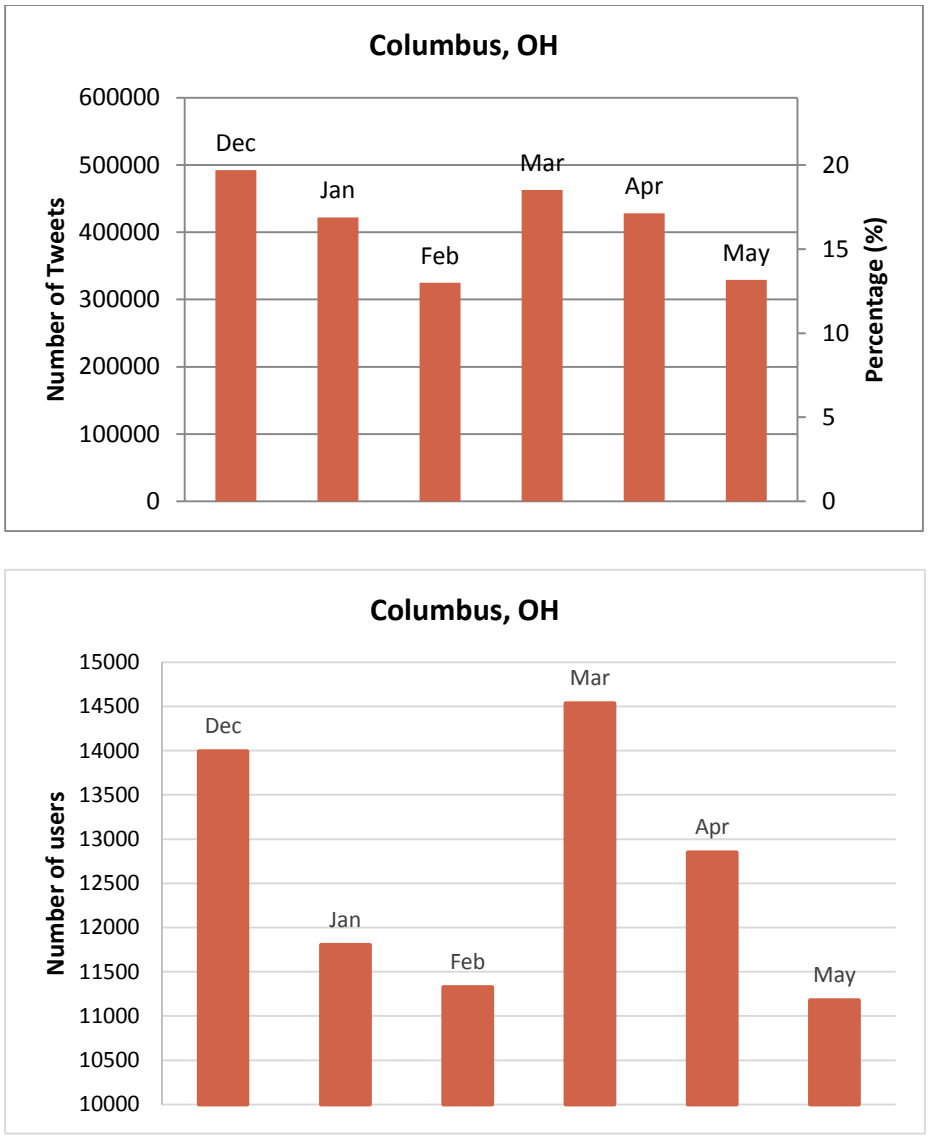


Figure 4.27 Number of tweets and users in each month in Columbus, OH

4.4.4 Tweets in different land uses

Tweets in institutional areas made up the majority of tweets in West Lafayette, and Bloomington, while in Ann Arbor and Columbus, tweets in residential areas accounted for the most than other land uses (Table 4.5). Less than 20% of tweets in West Lafayette were from residential areas, indicating that Twitter users prefer tweeting at school than home.

Over 10% of tweets in Bloomington and Columbus were from commercial areas, indicating that Twitter users were active in these areas, while in West Lafayette and Ann Arbor, very few tweets were from commercial areas (Table 4.5). Further insights on how tweets in land uses change with time are as follows.

Table 4.5 Percentages of Tweets in land uses in four study areas

% in total	West Lafayette	Bloomington	Ann Arbor	Columbus
Institutional	72.60	45.61	17.67	10.39
Residential	18.52	29.39	44.75	68.48
Commercial	1.40	15.64	6.30	11.78

In West Lafayette, most of the tweets were posted from institutional areas, which implied that most of the Twitter users were college students. Different from the temporal pattern of all the tweets in the city, the peak for institutional areas was around 12-1:00 pm. The tweet count began to decrease until around 7:00 pm, when it role to a peak at 10:00 pm. The land use with the second most tweets was residential areas, where the number of tweets drastically increased at 7:00 pm until 10:00 pm, which corresponds to the period of time when people leave from work or school and return home. Very few tweets were found in other land use types such as industrial and business (Figure 4.28).

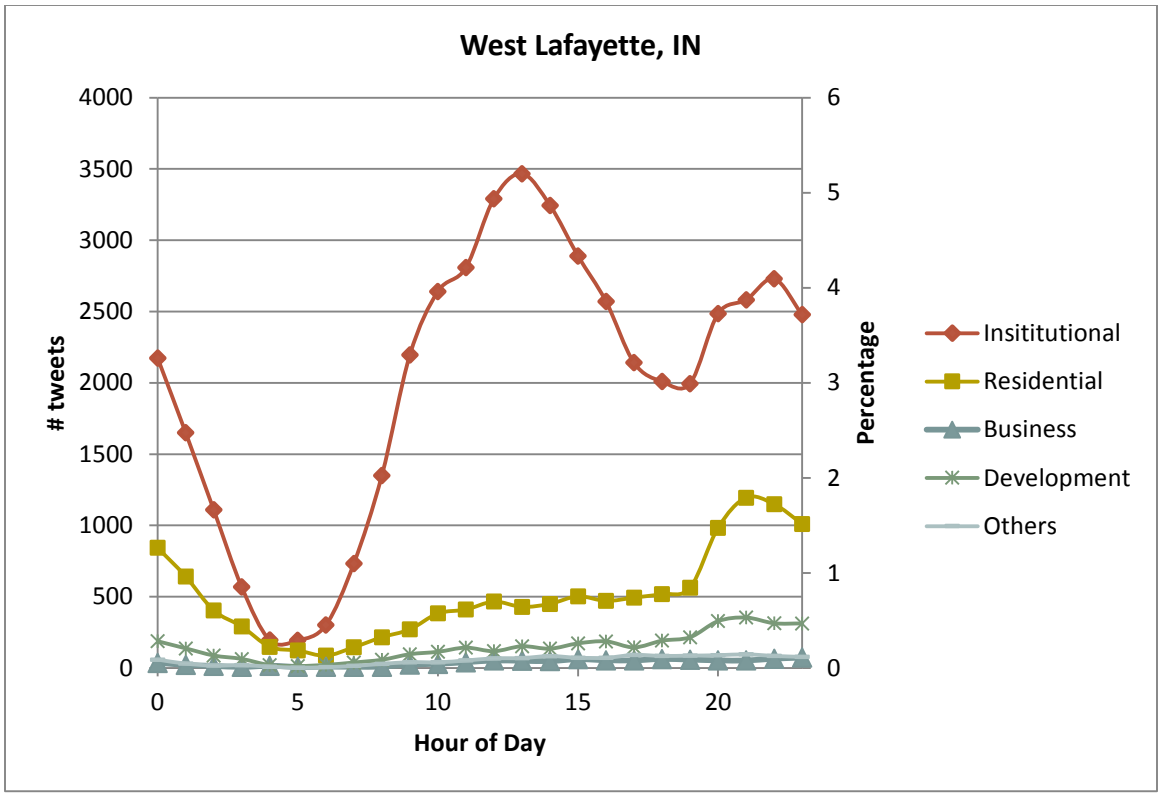


Figure 4.28 Hourly number of tweets in each land use in West Lafayette, IN

Similar to West Lafayette, the land use type with the most tweets was institutional areas in Bloomington. The temporal pattern was also nearly identical with a peak at 12:00 pm followed by a decrease until 6:00 pm and then an increase until a peak at 9~10:00 pm, inferring that many Twitter users are college students. Also, the land use with the second most tweets, similar to West Lafayette, was residential areas. Tweets in residential areas also began to rise from 6:00 pm to 9:00 pm. However, different from West Lafayette, where very fee tweets were posted from other land use areas, commercial area had as up to 1% of the total, indicating Twitter users' were active in these areas. Also, tweets were found in planned unit developments, which may be due to out of date land use data and which did not reflect areas that had been developed (Figure 4.29).

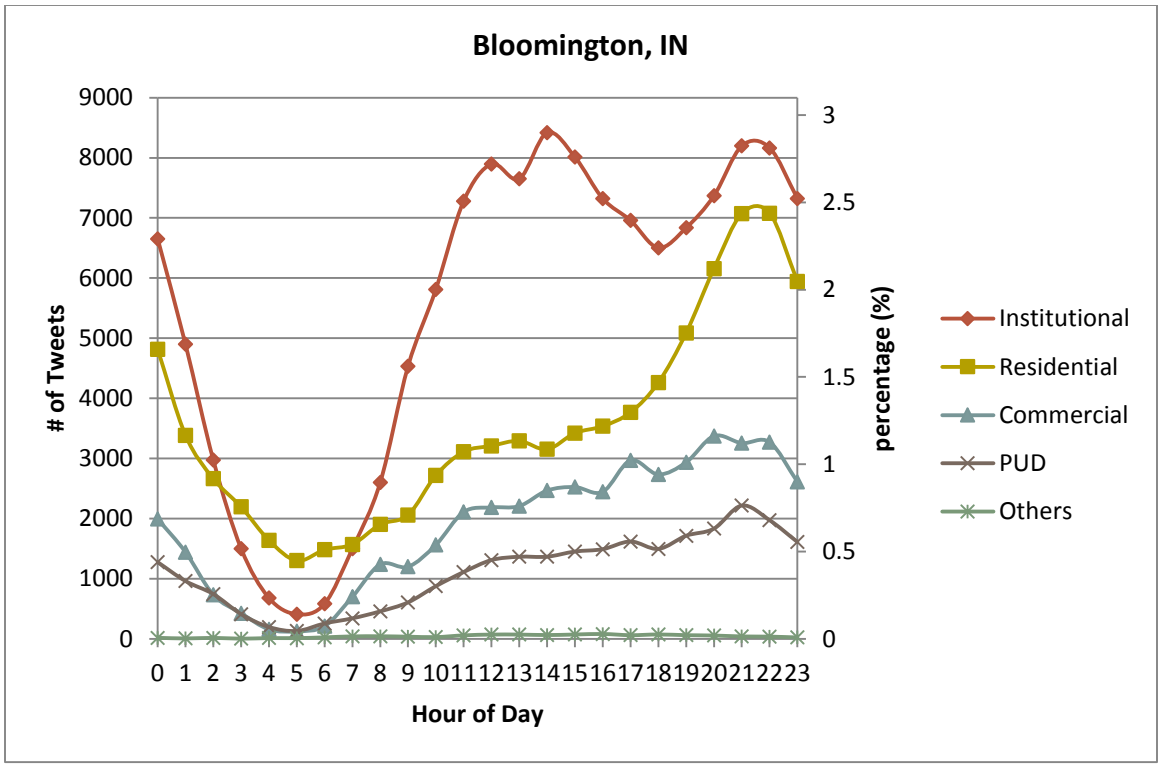


Figure 4.29 Hourly number of tweets in each land use in Bloomington, IN

Differing from West Lafayette and Bloomington, the land use type with the most tweets in Ann Arbor was residential areas, where the tweet counts began to increase from 6:00 am until 10:00 am, remained stable until 6:00 pm, and then continued to increase until 9:00 pm. The land use types with the second most tweets were institutional areas and transportation. In institutional areas, the number of tweets began to decrease at 2:00 pm and did not rise again until evening, which is different from West Lafayette and Bloomington. This implies that fewer students were on campus in Ann Arbor than in West Lafayette and Bloomington. The land use types in Ann Arbor included transportation, which mainly consisted of roads and highways, and it was surprising to discover that a large number of users were tweeting on the roads. When there was a decrease in the tweet

counts in the institution areas and an increase in the transportation and residential areas around 4:00 pm to 5:00 pm, a population flow from the institution areas to the transportation and residential areas was inferred. Finally, knowing that tweet counts in commercial and recreation areas comprised 0.2% ~ 0.5% of the total tweets and that relatively more tweets took place in the daytime, it was concluded that the Twitter users were usually active during the daytime in those areas (Figure 4.30).

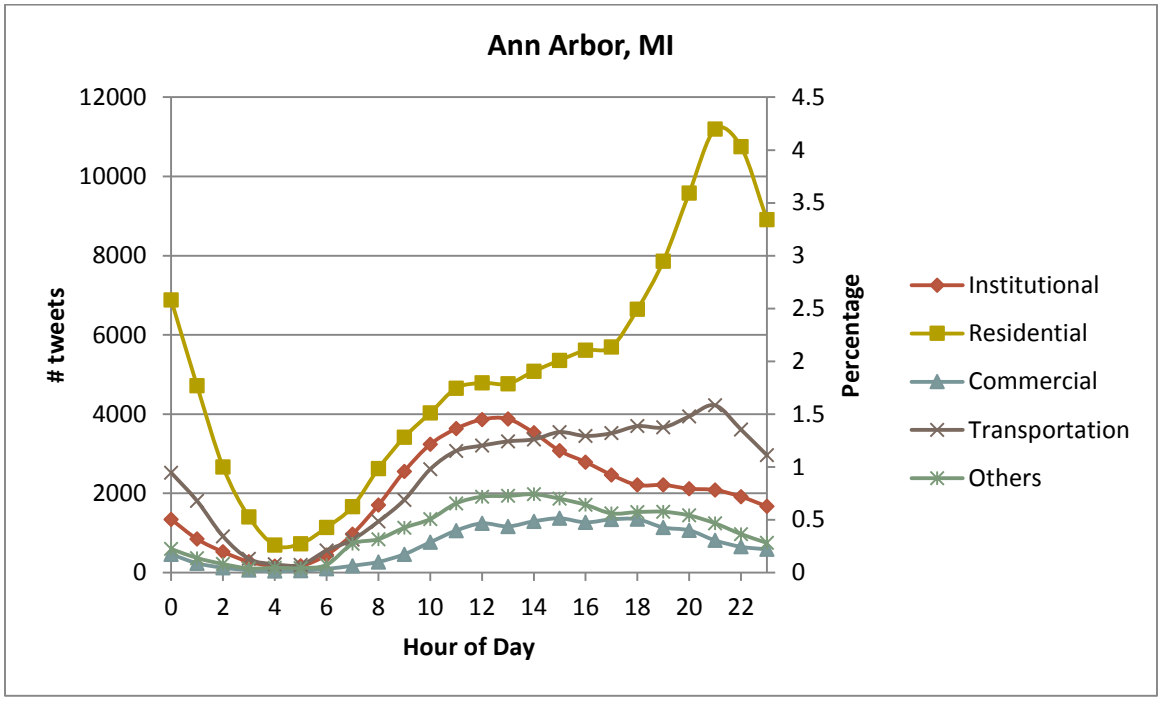


Figure 4.30 Hourly number of tweets in each land use in Ann Arbor, MI

In Columbus, a vast majority of tweets occurred in residential areas followed by commercial and institutional areas. The tweets in institutional areas had patterns similar to West Lafayette and Bloomington, with a peak around 12:00 pm and a small rise around 8:00 pm to 9:00pm. Also, the tweet counts in the commercial areas, with a peak reaching almost 0.8% of all the tweets, were nearly as many as those in the institutional areas. Since

the downtown area, which has several commercial businesses, malls, and restaurants and belongs to a separate land use type, the Downtown District, the tweet counts from the commercial area should be larger than shown here. This percentage was the highest among the other cities. It can be concluded that many Twitter users posted tweets from their homes and were also more active in commercial areas than those in other cities, indicating that Twitter potentially can be utilized for business applications such as market analysis and advertising (Figure 4.31).

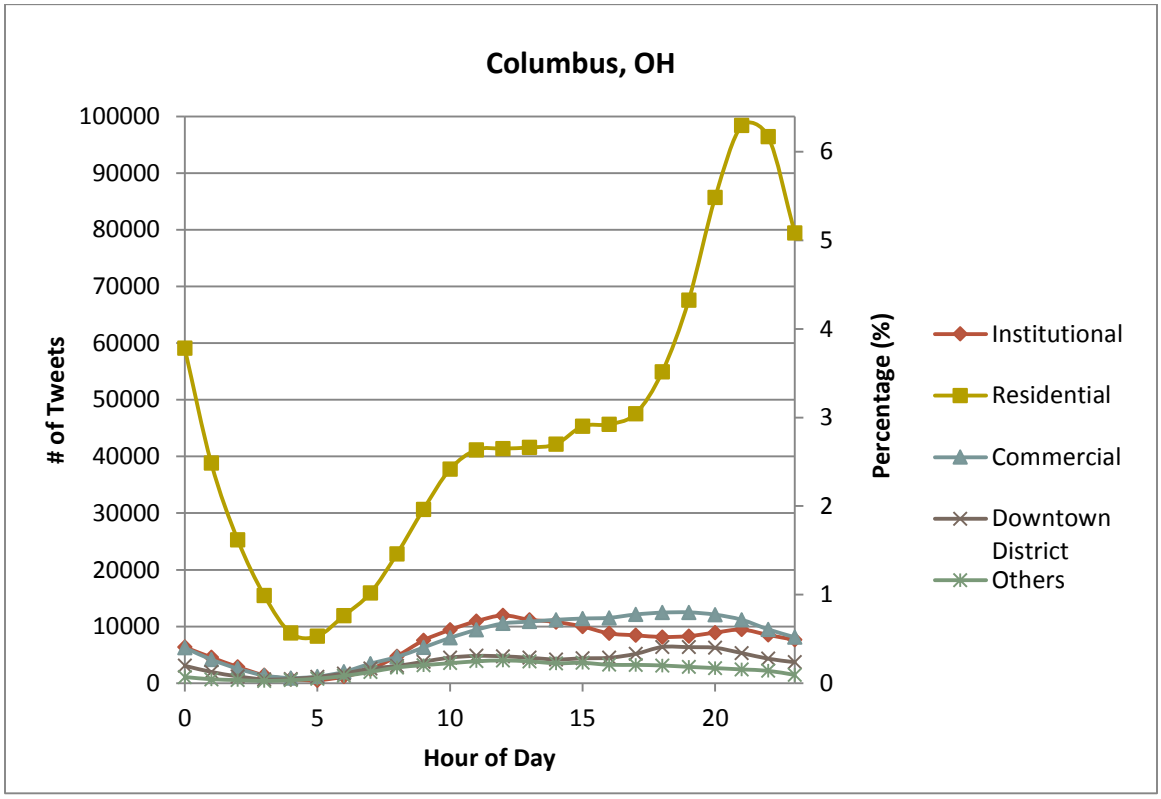
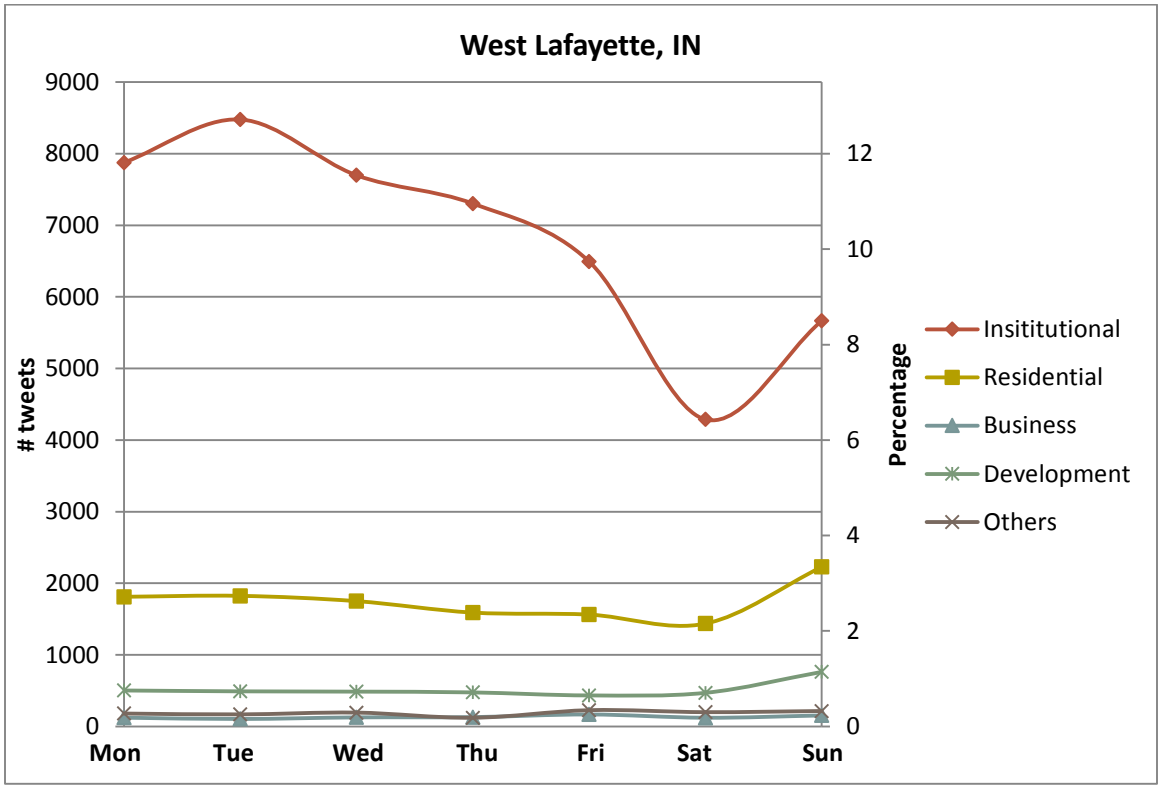
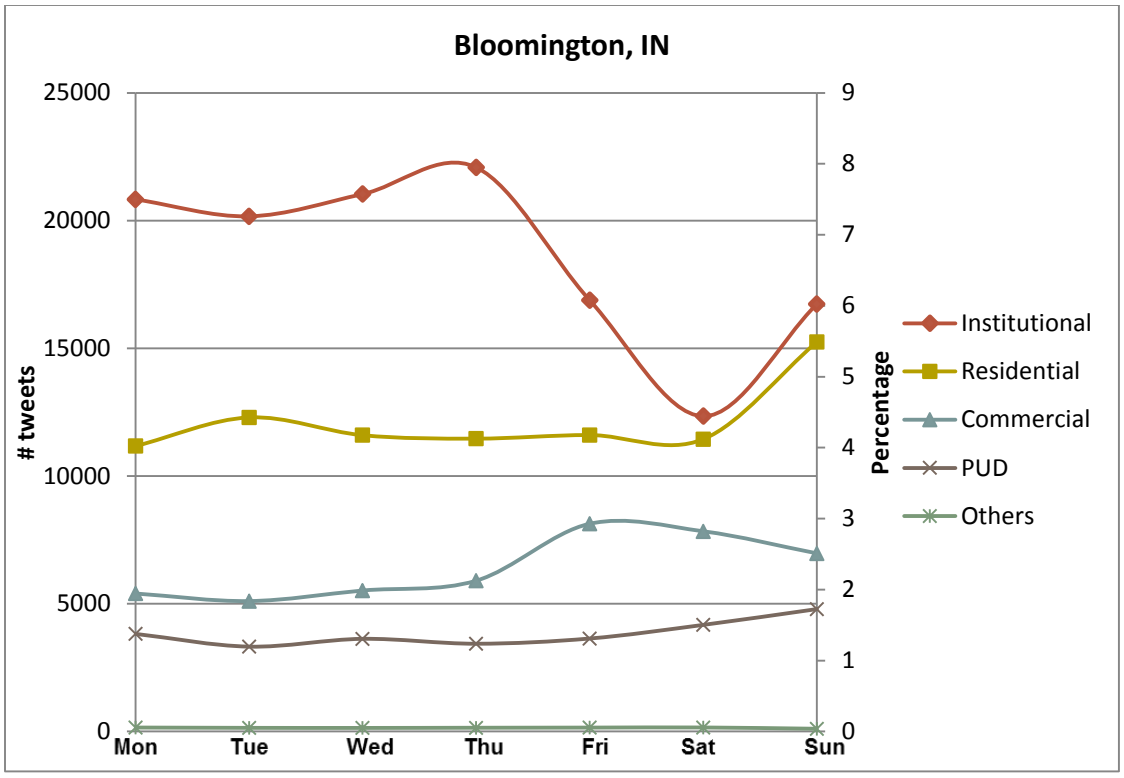


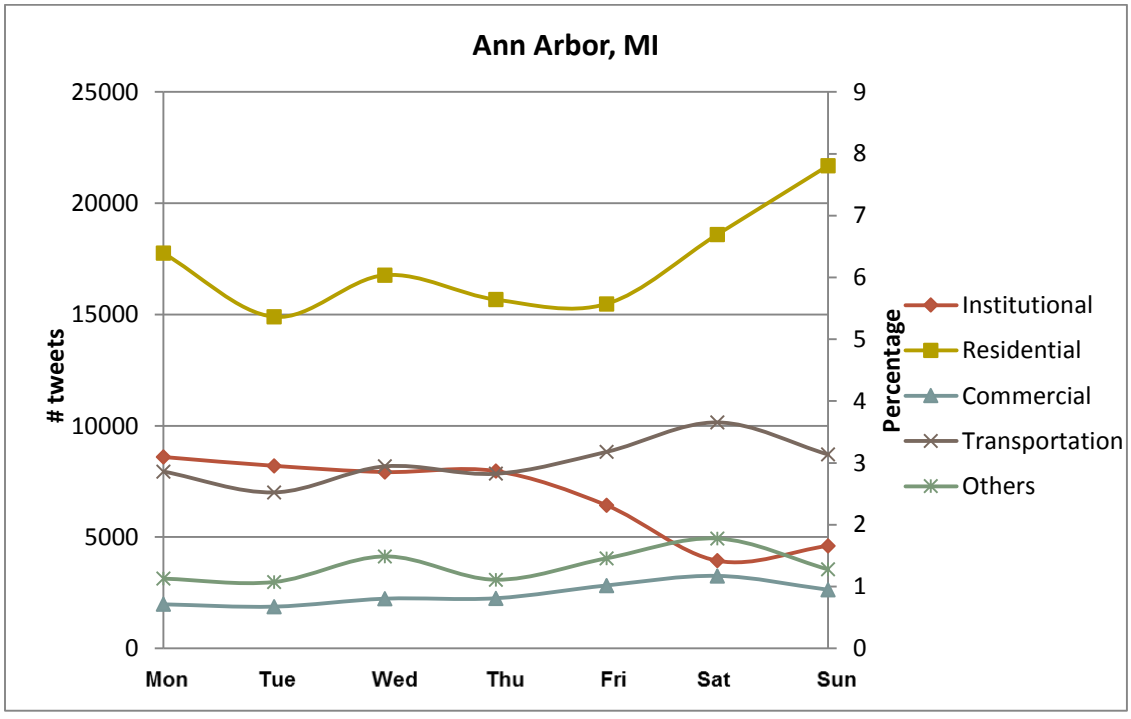
Figure 4.31 Hourly number of tweets in each land use in Columbus, OH

From the tweet counts for the weekdays in the four cities, there was an obvious increase in the number of tweets in residential areas on the weekends and a decrease in the institutional areas. Also, more tweets were posted in the commercial areas on the weekends,

which corresponds to the fact that people are not at work or school and stay at home, go shopping or enjoy entertainment on the weekends. In Ann Arbor, more tweets were found in the transportation patterns of users on weekends, indicating that Ann Arbor users traveled more on the roads. Another interesting result was that the number of tweets on Sunday was larger than on Saturday, especially in West Lafayette and Bloomington, inferring that more students studied on Sundays than on Saturdays (Figure 4.32).




Bloomington, IN
25000
20000
15000
10000
5000
0
tweets
9
8
7
6
5
4
3
2
1
0
Percentage
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Institutional
Residential
Commercial
PUD
Others


Ann Arbor, MI
25000
20000
15000
10000
5000
0
tweets
9
8
7
6
5
4
3
2
1
0
Percentage
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Institutional
Residential
Commercial
Transportation
Others

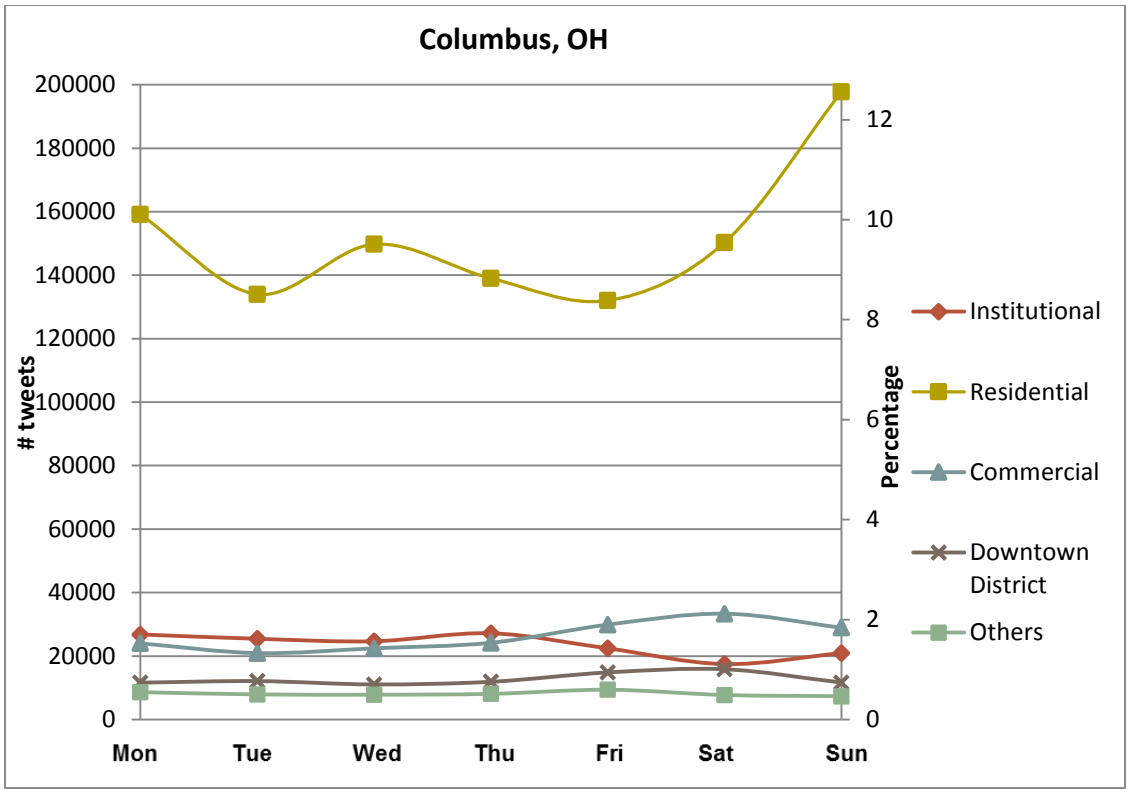
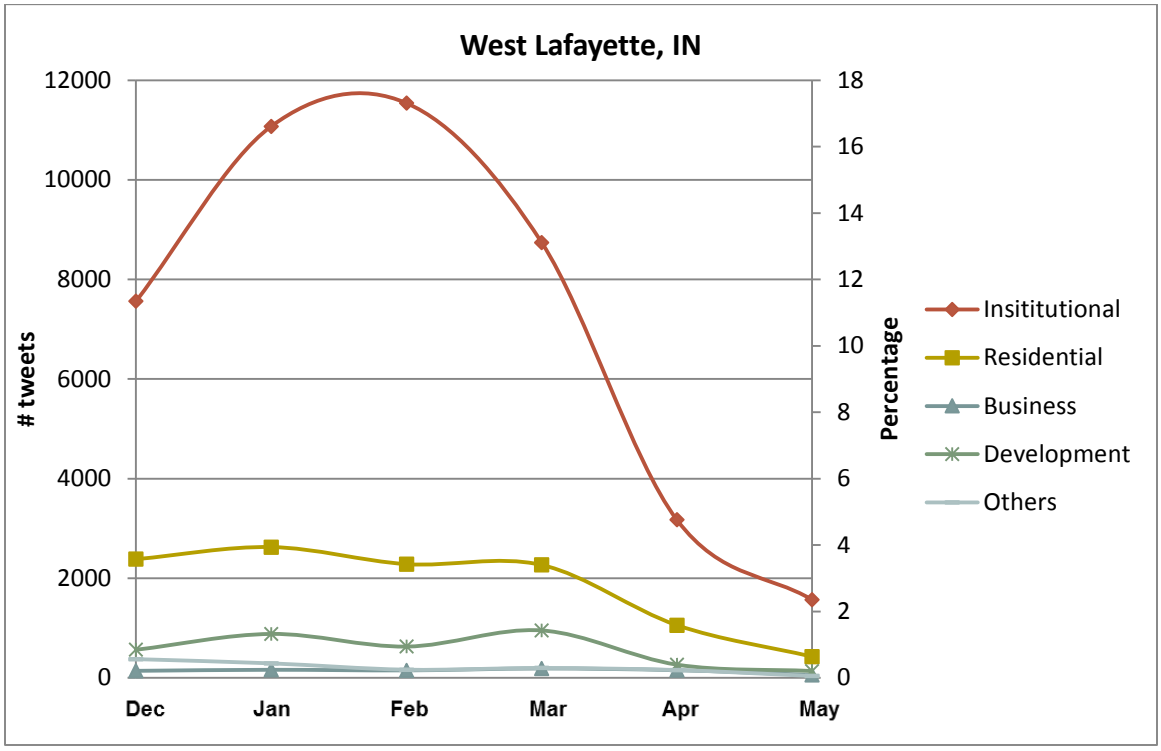
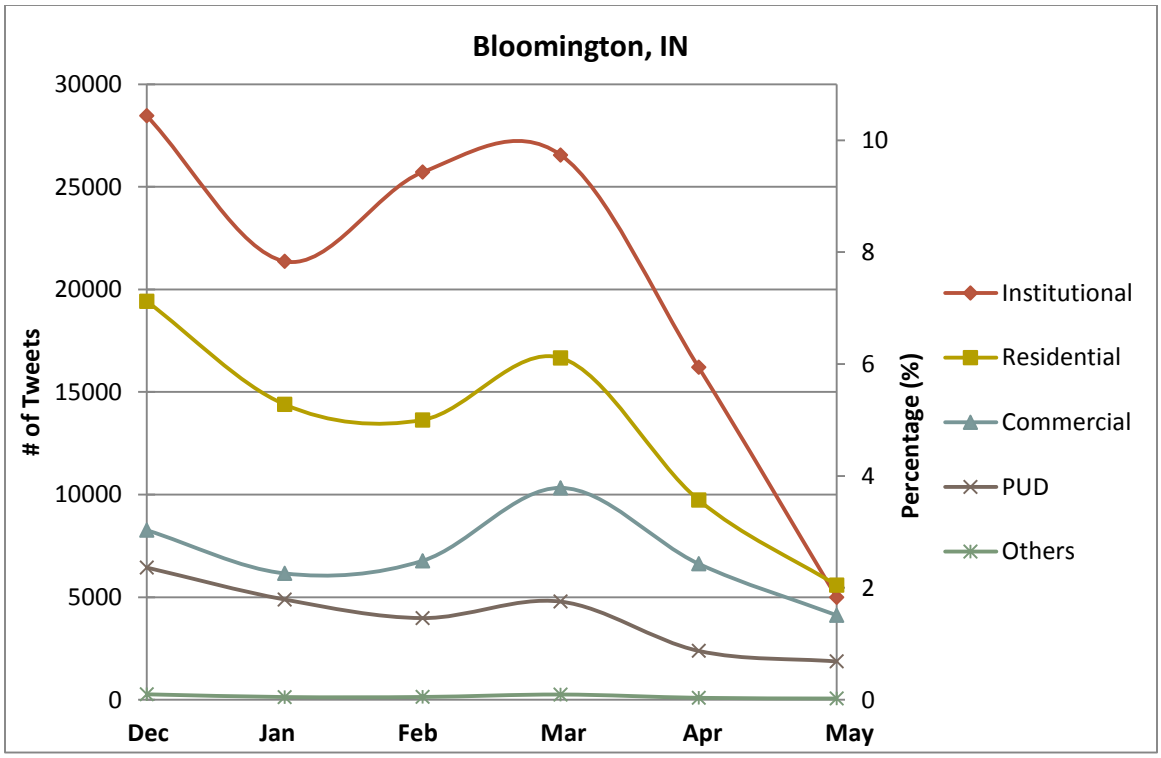


Figure 4.32 Daily number of tweets in each land use changes

For residential areas, the tweet counts began to decrease in April, when Twitter users likely were enjoying outdoor activities instead of staying at home. Also, there was a valley in February, which was probably due to less major events that month. For institutional areas, the tweets in West Lafayette reached a peak in January and February because the Purdue second semester started in early January, and the number of tweets declined after April when spring arrived and the semester was ending. For Bloomington and Ann Arbor, the tweets in January were less than at other times, which was probably because school starts in late January. In Ann Arbor, there was a valley in the number of tweets in transportation in February while the tweets increased in March when warmer weather arrived. For commercial areas, the tweet counts in Bloomington, Ann Arbor, and

Columbus increased in March when the weather improved, making it possible to do more shopping than during the winter months (Figure 4.33).





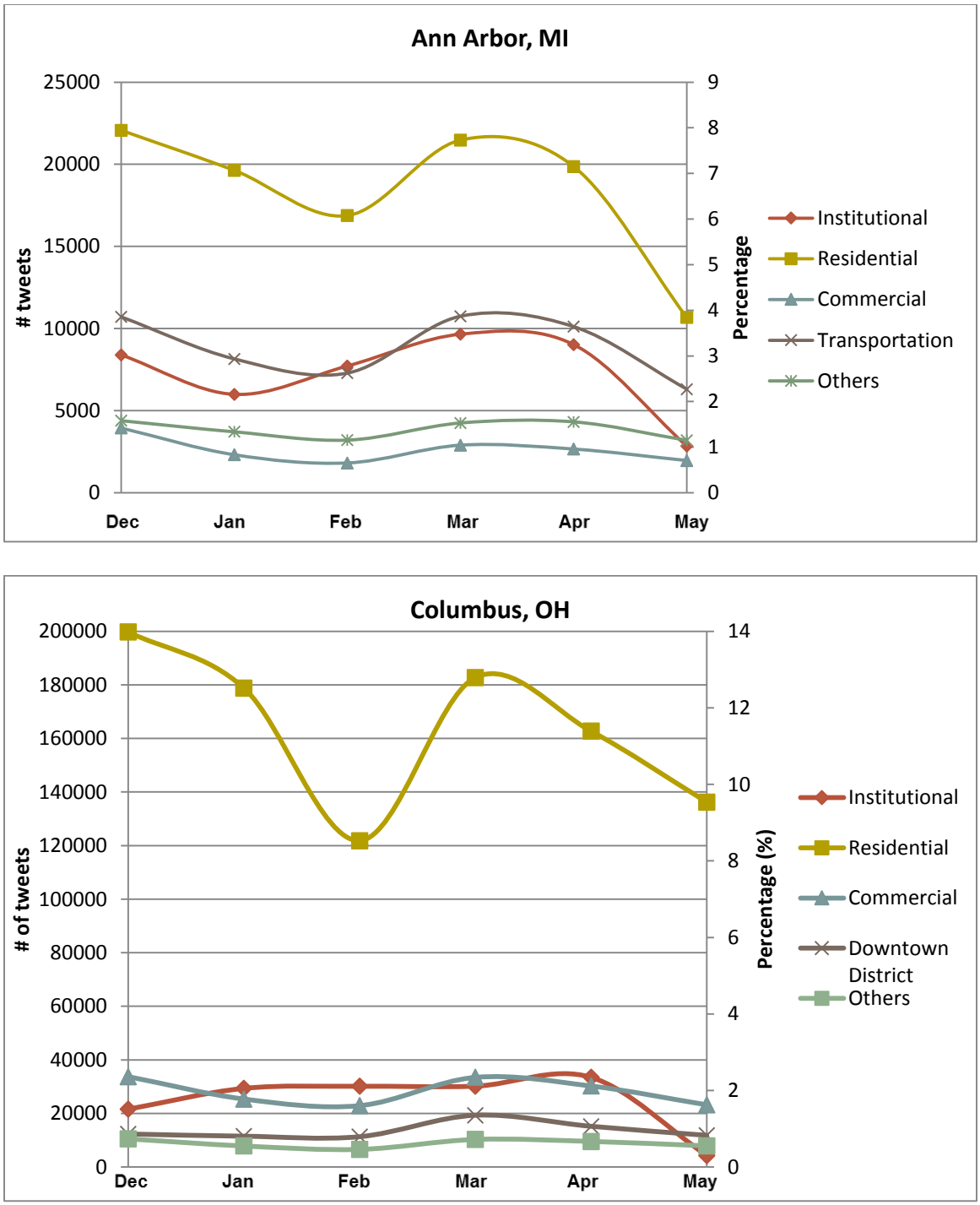


Figure 4.33 Monthly number of tweets in each land use

4.5 Event Detection

The STSS technique was applied to two cases: the football game in the University of Michigan stadium against the Ohio State University on November 30, 2013 beginning at 12:00pm, and the shooting event happened on Purdue University campus on January 21, 2014 around 12:00 pm. The football game received a lot of attention because the two universities are long-time rivals in football. And the game was very exciting; the University of Michigan lost with a final score of 41-42. It is assumed that when people attend the game, they would tweet about the game in the stadium, which lead to a space-time cluster. Thus, STSS method is used to identify space-time clusters. Due to the limited time for the analysis as well as the performance of the computer, only tweets on University of Michigan campus on that day were included in this analysis.

The shooting happened on the Electrical Engineering building around noon, and then all students on campus sheltered-in-place until around 1:30pm. As this is a sudden and shocking event, word spread very quickly and people all over West Lafayette, especially students on campus talked about this on Twitter. Particularly, during the lockdown period, students went on Twitter for latest updates from Purdue official accounts as well as their friends, and they tweeted or retweeted about the event. Therefore, tweets about this event are assumed to be clustered in time, but not necessarily in space. In this analysis, all tweets in West Lafayette on January 20 - 22, 2014, before, on and after the day of shooting, are used. The maximum temporal window is set to 3 hours.

4.5.1 University of Michigan football game

Within the dataset, five tweet clusters were found. Most of the points (red) in Cluster 1 were in or around the University of Michigan football stadium (Figure 4.34); and

the time period of the cluster, 10:00 am ~ 3:00 pm, is when the football game took place. Also, Cluster 1 has the highest test statistics, 77.932452 (Table 4.6), which indicates a strong clustering of points. Therefore, the tweets in Cluster 1 were very likely about the game, but the other clusters were uncertain. However, it was speculated that Cluster 4, which was located around the university campus and the downtown area, appeared right after the game might have been people gathering after the game. As for Clusters 2, 3, and 5, based on their sizes and the number of tweets in the cluster, as well as their short duration perhaps indicated home parties or friends gathering. It can be concluded that this analysis successfully detected the event, which was the football game between the University of Michigan and the Ohio State University. The time and location of the event was inferred with the utilized method without any prior knowledge of it.

Table 4.6 Statistics for tweet clusters found for University of Michigan football game

Cluster ID	**1**	**2**	**3**	**4**	**5**
Time Frame	10 ~ 15	0 ~ 1	23 ~ 23	16 ~ 19	12 ~ 13
Longitude	-83.7505	-83.7386	-83.7302	-83.7485	-83.7403
Latitude	42.1643	42.2667	42.2767	42.2772	42.2711
Radius (km)	0.51	0.043	0.52	0.60	0.01
Number of Cases	439	44	48	160	22
Expected Cases	240.65	5.53	8.75	87.81	3.17
Observed/Expected	1.82	7.95	5.28	1.82	6.93
Test Statistics	77.93245	53.16484	42.84346	25.26754	23.85545
P-value	10^{-17}	10^{-17}	10^{-16}	10^{-8}	$5*10^{-8}$

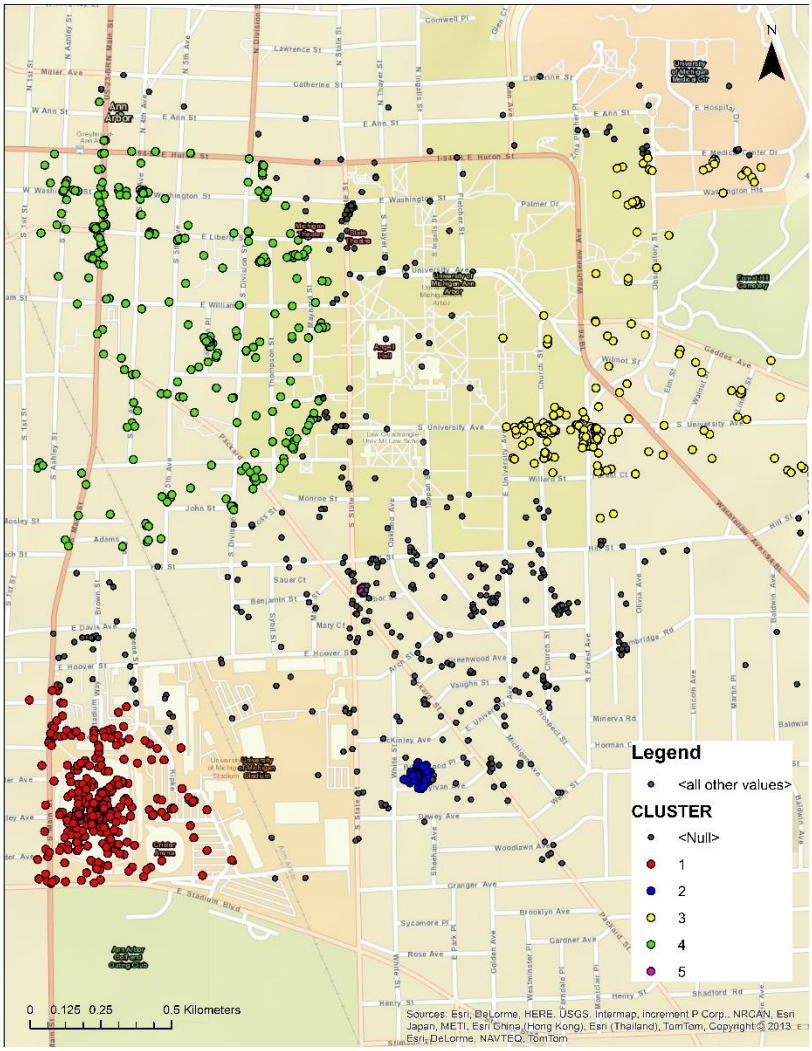


Figure 4.34 Tweet clusters found on University of Michigan campus on November 30, 2013

4.5.2 Shooting on Purdue University campus

In West Lafayette dataset, only one tweet cluster was found, which includes tweets from 12:00 pm to 2:00 pm (Table 4.7). This coincides with the occurrence of the shooting around noon and the shelter-in-place until 1:30pm. The high relative risk, large likelihood ratio and small p-value indicate a significant cluster (Table 4.7). Therefore, this confirms the ability of using Twitter in detecting events. A sharp rise in number of tweets is observed

around noon on January 20 (Figure 4.35). The number is much larger than the one on the day before and after the shooting. The results from the STSS method successfully reflected and detected this rise and temporal cluster of tweets. Also, the cluster only lasts for two hour (Table 4.7), and the number of tweets decreased around 3:00 pm (Figure 4.35). This implies that local discussion about the event on Twitter diminish very quickly.

Table 4.7 Statistics for tweet clusters in time for shooting of Jan 21, 2014 on Purdue campus

Time Frame	Number of Cases	Expected Cases	Observed/ expected	Relative Risk	Log Likelihood Ratio	P-value
12-14	941	46.14	20.40	27.68	2072.30	0.001

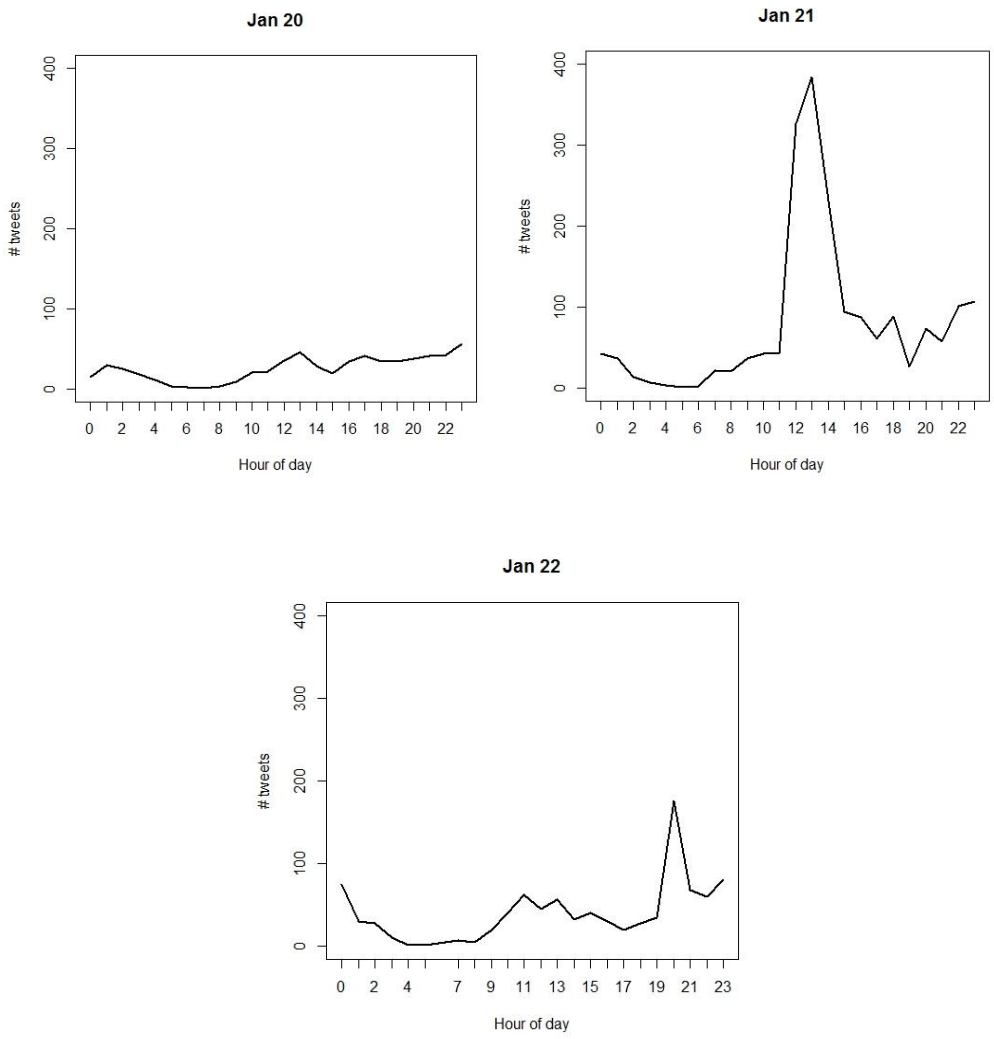


Figure 4.35 Number of tweets in West Lafayette on January 20 - 22, 2014

CHAPTER 5. CONCLUSION

This thesis explored the spatial and temporal patterns of geo-tagged tweets from Midwestern college cities/towns, and revealed the human mobility patterns of the Twitter users. The results generally reflected everyday human activity patterns and urban characteristics. It is discovered that the majority of tweets were posted from a small portion of Twitter users. A long tail was discovered in the distribution of the number of users vs. the number of tweets. The long tails included a small number of users who each had posted relatively more tweets, which made up the majority of the distribution.

This thesis also discovered a positive linear correlation between the radius of city and the median or mean commute distance. The larger the city is, the longer the median or mean commute distance is. The average commute distance is about 40% of the city radius. The model might be used for other cities. This thesis also developed a methodology to find the places of frequent visits of the Twitter users and calculate commute distances from geo-tagged tweets. With this methodology, majority of Twitter users had two to four places of frequent visits.

Moreover, Twitter users in these four cities were active from 10:00 am to 12:00 am at midnight. The tweet count rose at a peak at 9:00 pm. Also, there were more Twitter users during weekends than weekdays. The “weekday” Twitter user group and the “weekend” Twitter user group had only a small overlap, about 20%, implying that most people only

tweet either on weekdays or weekends. There were more tweets on weekdays than weekends in smaller cities; however, in bigger cities, there were more tweets on weekends than weekdays likely due to the relatively large size and their larger offerings of entertainment venues and major events that might keep residents in town during the weekends and attract out of town visitors as well. Plus, the tweet counts started to decline and reached a valley in May due to end of school and departure of students.

Moreover, in smaller cities, tweets in institutional areas made up the majority of tweets; and in bigger cities, tweets in residential areas accounted for most. For institutional areas, number of tweets began to rise around 7:00 am when the classes began, and it continued to rise at a peak at lunchtime around 12:00 pm. Then the number began to decrease until 6:00 pm and then an increase until a peak at 9:00 pm, implying that students work hard at night at school. Also, the number of tweets from institutional areas on Sunday is larger than that on Saturday, inferring that students return to school to study on Sundays. Plus, there was a drastic drop in number of tweets in institutional areas in May due the departure of students. For residential areas, tweets began to rise from 6:00 pm to 9:00 pm when people return from work and relax at home. There was an obvious increase in the number of tweets in residential areas on the weekends. The tweet counts in residential areas began to decrease in April, when Twitter users likely were enjoying outdoor activities instead of staying at home. For commercial areas, more tweets were posted on weekends than weekdays. The tweet counts in commercial areas increased in March when the weather got warmer, making it possible to do more shopping than during the winter months.

Furthermore, tweet clusters usually emerged on university campuses and apartment complexes. In big cities such as Ann Arbor and Columbus, tweet clusters were found at shopping malls.

Finally, tweets were shown to be capable of not only successfully illustrating general human activity patterns, but pinpointing the occurrence of anomalies or events as well. This thesis also found that discussion on Twitter about events diminished quickly in local areas. Thus, this thesis demonstrated the potential for using tweets in human behavior research and suggests the possibility of applying this method to other geo-social research.

However, there are limitations in using tweets in social research since the data may be biased for various reasons. There is no current quantitative information available on the socioeconomic structure of Twitter users due to privacy restrictions. Also, since Twitter requires users to opt-in to enable the geo-tag function, the motivation to do this varies with their social behaviors and personalities, or even the rewards of doing so. Thus, Twitter users may not be well representative of the general public. The captured information may cover only a portion of the total human activity and mobility patterns of its users. Also, for the event detection analysis, the methodology did not consider the number of users in one cluster, which means, one user tweeting multiple times at one location may result in a cluster in this scenario.

One possible future direction of this research can be taking advantage of the content of tweets, and combined with text mining, topic modeling, and natural language processing, to discover more information and patterns. This can facilitate the interpretation of users' activity type, the function of the tweet clusters of frequent Twitter users as well as the detection of space-time tweet clusters and types of gatherings or events. The other future

direction is to investigate the possibilities of applying the spatial and temporal patterns into more fields such as traffic planning, market analysis, business, urban study, politics, and social media research.

REFERENCES

2010 Population Finder. (2010). *United States Census Bureau*, retrieved July 31st, 2013 from http://www.census.gov/popfinder/?fl=18

Adebisi, A. A., Olusayo, O. E., Stephen, O., Olatunde, A. A. B., & Titilayo, A. O. (2012). Development of a hybrid K-means-expectation maximization clustering algorithm. *Journal of Computations & Modelling*, 2(4), 1-23.

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.

Ann Arbor, Michigan. (n.d.). In *Wikipedia*, Retrieved November 6, 2014, from http://en.wikipedia.org/wiki/Ann_Arbor,_Michigan

Bayir, M. A., Demirbas, M., Eagle, N. (2009). Discovering spatiotemporal mobility profiles of cellphone users. In *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a* (pp. 1-9). IEEE.

Block, R. (2007). Scanning for Clusters in Space and Time: A Tutorial Review of SaTScan. *Social Science Computer Review*.

Bloomington, Indiana (n.d.). In *Wikipedia*. Retrieved November 5, 2014, from http://en.wikipedia.org/wiki/Bloomington,_Indiana

Cheng, Z., Caverlee, J., Lee, K., Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. *ICWSM*, *2011*, 81-88.

Cheng, T., & Wicks, T. (2014). Event Detection using Twitter: A Spatio-Temporal Approach. *PloS one*, *9*(6), e97807.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., Zook, M. (2013). Beyond the geotag: situating 'big data'and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130-139.

Columbus, Ohio (n.d.) In *Wikipedia*, Retrieved November 6, 2014, from http://en.wikipedia.org/wiki/Columbus,_Ohio

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Elwood, S. (2008). Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS, *GeoJournal*, 72, 173-183.

Elwood, S., & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, *42*(1), 6-15

Fujisaka, T., Lee, R., Sumiya, K. (2010, January). Discovery of user behavior patterns from geo-tagged micro-blogs. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication* (p. 36). ACM.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90-102.

Gonzalez, M. C., Hidalgo, C. A., Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779-782.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, *3*(3), 231-241.

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, *7*(4), 36-43

Hasan, S., Schneider, C. M., Ukkusuri, S. V., González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*,*151*(1-2), 304-318.

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

Housing (n.d.). Retrieved November 5, 2014, from http://www.iub.edu/student/housing.shtml

Housing Options (n.d.). Retrieved November 6, 2014, from http://housing.umich.edu/options

Hiruta, S., Yonezawa, T., Jurmu, M., Tokuda, H. (2012). Detection, classification and visualization of place-triggered geotagged tweets. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 956-963). ACM.

Kang, C., Gao, S., Lin, X., Xiao, Y., Yuan, Y., Liu, Y., Ma, X. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Geoinformatics, 2010 18th International Conference on* (pp. 1-7). IEEE.

Kwan, M. P. (1999a). Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities*. *Economic geography*, *75*(4), 370-394.

Kwan, M. P. (1999b). Gender and individual access to urban opportunities: a study using space–time measures. *The Professional Geographer*, *51*(2), 210-227.

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

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine*, *2*(3), e59.

Kulldorff, M., Information Management Systems. (2009). *SaTScan v8. 0: software for the spatial and space-time scan statistics*.

Kulldorff, M. (2014). SaTScan™ User Guide.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*,*18*(5).

Li, L., M. F. Goodchild, B. Xu, (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr, *Cartography and Geographic Information Science*, 40(2), 61–77.

Li, Y., & Shan, J. (2013). Understanding the Spatio-temporal Pattern of Tweets. *Photogrammetric engineering and remote sensing*, *79*(9), 769-773.

Lunden, I. (2012). Analyst: Twitter passed 500m users in June 2012, 140m of them in US; Jakarta 'Biggest Tweeting'city. *TechCrunch RSS*, *30*.

Malleson, N., & Andresen, M. A. (2014). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, (ahead-of-print), 1-10.

Milstein, S., Lorica, B., Magoulas, R., Hochmuth, G., Chowdhury, A., O'Reilly, T. (2008). *Twitter and the micro-messaging revolution: Communication, connections, and immediacy--140 characters at a time*. O'Reilly Media, Incorporated.

Miller, G. (2011). Social scientists wade into the tweet stream. *Science*, *333*(6051), 1814-1815.

Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS*, *14*(3), 223-239.

Nakaji, Y., & Yanai, K. (2012, July). Visualization of real-world events with geotagged tweet photos. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on* (pp. 272-277). IEEE.

Nasser, S., Alkhaldi, R., Vert, G. (2006, July). A modified fuzzy k-means clustering using expectation maximization. In *Fuzzy Systems, 2006 IEEE International Conference* on (pp. 231-235). IEEE

Purdue University. (n.d.). In *Wikipedia*. Retrieved August 1, 2013 from http://en.wikipedia.org/wiki/Purdue_University

Popescu, A., Grefenstette, G., Moëllic, P. A. (2009, November). Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1713-1716). ACM.

Point Density (Spatial Analyst). (n.d.). Retrieved November 24, 2014, from http://resources.arcgis.com/en/help/main/10.2/index.html#//009z0000000v000000

Sakaki, T., Okazaki, M., Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). ACM.

Stevenson, J. R., Emrich, C. T., Mitchell, J. T., Cutter, S. L. (2010). Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following Hurricane Katrina. *Cartography and Geographic Information Science*, *37*(1), 57-68.

Stone, B. (2012). Twitter, the startup that wouldn't die. *Bloomberg BusinessWeek*, *1*, 62-67.

The Ohio State University (n.d.). In *Wikipedia*, Retrieved November 6, 2014, from http://en.wikipedia.org/wiki/Ohio_State_University

The Streaming APIs Overview (2014). Retrieved November 6, 2014, from https://dev.twitter.com/streaming/overview

Tsou, M. H., & Leitner, M. (2013). Visualization of social media: seeing a mirage or a message?. *Cartography and Geographic Information Science*,*40*(2), 55-60.

Tsou, M. H., Yang, J. A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, *40*(4), 337-348.

Twitter (n.d.). In *Wikipedia*. Retrieved November 12, 2014 from http://en.wikipedia.org/wiki/Twitter

Unix Time (n.d.). In *Wikipedia*. Retrieved November 6, 2014, from http://en.wikipedia.org/wiki/Unix_time

Vadrevu, K. P. (2008). Analysis of fire events and controlling factors in eastern India using spatial scan and multivariate statistics. *Geografiska Annaler: Series A, Physical Geography*, *90*(4), 315-328.

West Lafayette, Indiana (n.d.). In *Wikipedia*. Retrieved November 5, 2014, from http://en.wikipedia.org/wiki/West_Lafayette,_Indiana

Zook, M., Graham, M., Shelton, T., Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy*, *2*(2), 7-33.