

PATTERN EXPLORATION FROM CITIZEN GEOSPATIAL DATA

A Thesis

Submitted to the Faculty

of

Purdue University

by

Ke Liu

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

December 2018

Purdue University

West Lafayette, Indiana

**THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF THESIS APPROVAL**

Dr. Jie Shan, Chair

Lyles School of Civil Engineering

Dr. James S. Bethel

Lyles School of Civil Engineering

Dr. Mary L. Comer

School of Electrical and Computer Engineering

Approved by:

Dr. Dulcy Abraham

Head of the School Graduate Program

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Prof. Jie Shan for the continuous support of my study and research, and for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. James Bethel and Prof. Mary Comer for their kindness, encouragement, and insightful comments.

My sincere thanks also goes to Dr. Fenggang Yang, who provided me a research assistant position in his center, and who gave me access to projects and research facilities. Without his precious support it would not be possible to finish my degree.

I am extremely thankful for Yue Li, Qinghua Li, and Yuqian Huang for sharing expertise and valuable guidance. I am extending my thanks to Joanne Yang, Yunping Tong, Yun Lu, Jonathan Pettit, and everyone in Center on Religion and Chinese Society for their constant encouragement.

A very special gratitude goes out to SafeGraph for providing the data for this work. I am also grateful to HERE Technologies and Gistic Research for providing me with internship opportunities. The work experience I have gained and skills I have developed from these internships are valuable. I would like to show my special thanks to Justin Eylander, my manager at HERE Technologies. He not only mentored me technically, but also helped me to build up confidence and kept looking for job openings for me.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and for their unceasing encouragement and attention.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
ABSTRACT	ix
1 INTRODUCTION	1
1.1 Background	1
1.2 Related Work	4
1.3 Overview of the Thesis	7
2 METHODOLOGY	8
2.1 Kernel Density Estimation	8
2.2 Flow Extraction Methods	11
2.2.1 Gravity Based Flow Extraction Model	11
2.2.2 Flow Extraction from Density Difference	12
2.3 Spatial Autocorrelation	15
2.3.1 Moran's I	15
2.3.2 Getis-Ord G_i^*	16
2.4 Sentiment Analysis	16
3 TEST AREA AND DATA	22
3.1 GPS Trajectory Data	22
3.2 Twitter Data	24
4 PATTERNS OF HUMAN MOBILITY	26
4.1 Density Maps	26
4.2 Flow Maps	34
4.3 Summary	43
5 PATTERNS OF HUMAN SENTIMENTS	47

	Page
5.1 Citizen Sentiment Patterns	47
5.2 Temporal Patterns	48
5.3 Spatial Patterns	59
5.4 Summary	60
6 CONCLUSION	64
REFERENCES	66

LIST OF TABLES

Table	Page
2.1 Kernels of $W = 2$	14
2.2 Kernel (p, q)	14
2.3 Kernel d	14
2.4 Kernel p	14
2.5 Kernel q	14
2.6 Sentiment lexicon	19
2.7 Sentence parser annotation	20
2.8 Polarity of tweets	21
3.1 GPS TRAJECTORY DATA SETS	22
3.2 Twitter data sets	25
4.1 Emerging hot spot category	33
5.1 ANOVA table	52
5.2 Tukey's test for polarity mean in West Lafayette	53
5.3 Tukey's test for percentage of positive polarity in West Lafayette	55
5.4 Tukey's test result for polarity mean of Bloomington, Ann Arbor, and Columbus	56
5.5 Global Moran's I result	59

LIST OF FIGURES

Figure	Page
2.1 Flow extraction from density difference	13
2.2 Sentence parser	19
3.1 Number of points against number of users	23
4.1 Map of West Lafayette	26
4.2 Heat maps of one hour before the volleyball game	28
4.3 Heat maps of one hour after the volleyball game	29
4.4 Emerging hot spot analysis before and after the volleyball game	32
4.5 Probability density before and after volleyball game	35
4.6 Density difference-based flow map before volleyball game	36
4.7 Density difference-based vector map and direction map before volleyball game	38
4.8 Convergent area and divergent area identified by direction map	39
4.9 Density difference-based flow maps before volleyball game with different window sizes	40
4.10 Density difference-based flow map and direction map after volleyball game	41
4.11 Flow maps with gravity-based flow extraction model	42
4.12 Density difference-based flow map of Puerto Rico, 7 AM to 10 AM, Aug 28, 2017	44
4.13 Density difference-based flow map of Puerto Rico, 4 PM to 7 PM, Aug 28, 2017	44
4.14 Density difference-based mobility map of Puerto Rico, 7 AM to 10 AM, Aug 28, 2017	45
4.15 Density difference-based mobility map of Puerto Rico, 4 PM to 7 PM, Aug 28, 2017	45
5.1 Histograms of average sentiment polarity per user in West Lafayette	48

Figure	Page
5.2 Histogram of positive and negative polarity percentage per user in West Lafayette, 2016	49
5.3 Histograms of average sentiment polarity per user in 2016 in Bloomington, and Ann Arbor, and Columbus	50
5.4 Temporal distribution of polarity in West Lafayette	51
5.5 Weekday pattern of positive polarity percentage in West Lafayette	54
5.6 Weekday pattern of polarity in Bloomington, Ann Arbor, and Columbus	57
5.7 Sentiment polarity distribution over time	59
5.8 Hot spot analysis of West Lafayette in 2016	61
5.9 Hot and cold clusters of West Lafayette in 2016	62
5.10 Hot and cold clusters of Columbus in 2016	63

ABSTRACT

Liu, Ke M.S., Purdue University, December 2018. Pattern Exploration from Citizen Geospatial Data. Major Professor: Jie Shan.

Due to the advances in location-acquisition techniques, citizen geospatial data has emerged with opportunity for research, development, innovation, and business. A variety of research has been developed to study society and citizens through exploring patterns from geospatial data. In this thesis, we investigate patterns of population and human sentiments using GPS trajectory data and geo-tagged tweets. Kernel density estimation and emerging hot spot analysis are first used to demonstrate population distribution across space and time. Then a flow extraction model is proposed based on density difference for human movement detection and visualization. Case studies with volleyball game in West Lafayette and traffics in Puerto Rico verify the effectiveness of this method. Flow maps are capable of tracking clustering behaviors and direction maps drawn upon the orientation of vectors can precisely identify location of events. This thesis also analyzes patterns of human sentiments. Polarity of tweets is represented by a numeric value based on linguistics rules. Sentiments of four US college cities are analyzed according to its distribution on citizen, time, and space. The research result suggests that social media can be used to understand patterns of public sentiment and well-being.

1. INTRODUCTION

1.1 Background

Advances in location-acquisition and mobile computing techniques have generated a massive amount of geospatial data. More now than ever before, exploring patterns within geospatial data is a priority of geographers, economists, and regional scientists. Geospatial data contains information that identifies geographical location and characteristics of natural or constructed features and boundaries on the surface of the earth [1]. Examples of geospatial data include satellite images, site addresses, and Global Positioning System (GPS) coordinates of a smart phone. Geospatial data is distinguished from traditional data in several ways. First, geospatial data is typically multidimensional [2]. The coordinates are often selected such that one of the numbers represents a vertical position and two or three numbers represent a horizontal position. For example, any place on the surface of the earth can be identified by a unique set of longitude and latitude values. Sometimes, time is considered to be a dimension of geospatial data to emphasize features that vary over time, such as time-dependent changes of vegetation cover. The second distinguishing factor of geospatial data is that the data are autocorrelated. Tobler's first law of geography states that everything in space is related but nearby things are more related than distant things [3]. For example, measurements made at locations that are nearby (i.e., in close physical proximity) tend to be closer in value as compared to measurements made at locations farther apart. Consequently, certain standard statistical methods cannot be applied to geospatial data since they do not account for the independence assumption. Another distinguishing factor of geospatial data is spatial heterogeneity. Geospatial data contain distance and topological information associated with Euclidean space. Every

location is unique due to its situation with respect to the rest of the spatial system, which implies that variation in spatial data is a function of the associated location [4].

There are many different forms and formats of geospatial data. Geospatial data can be discrete, continuous, accurate, fuzzy, and of various shapes. Sources of geospatial data include GPS, aerial photographs, land surveys, and location-aware technology sources. The increased use of embedded devices and advances in capturing techniques have led to an increased volume and density of geospatial data related to individual position and human activity. Human related data is called citizen geospatial data. For example, ride sharing systems such as Uber gather vast amounts of taxi trajectories. By embracing new approaches, citizen geospatial data now represents an opportunity for exploring the natural world and human society. Since information available on the Internet is constantly growing, a respectable amount of geospatial data are derived from review sites, forums, blogs, and social media. These geospatial data contain abundant semantic information such as public opinion about products, services, brands, or politics.

Geospatial social media data have three main features different than normal geospatial data. First, social media data are individual specific, which means that the contribution of different people to the collective data varies vastly based on the willingness of people to share information with the public. People have different habits regarding the use of social media. Some people frequently update their homepage while other people prefer to view content posted by others. Some people treat social media as a place for sharing opinions, while some people look for business opportunities via expanding their network. The second feature specific to geospatial social media data is the use of event-oriented social media data. Unlike methods that actively collect data, the volume and distribution of social media data is not regular. The frequency and explosiveness of posts not only depend on headcount, but also on the event, the time, and location. This feature emphasizes its association with human life and society. Third, social media is associated with an abundant amount of information that other sources cannot provide. Nowadays, most social me-

dia platforms allow people to post words, articles, pictures, and videos. These types of information are embedded in social media data and may reveal hidden patterns of human behavior. Twitter is one of the most popular social media software, as many celebrities, politicians, and bloggers use this platform to make announcements and share opinions. Twitter tightly follows social topics because people join discussions and exchange information via this app. People also use Twitter to initiate slogans, commercials, and social activities. As a result, Twitter contains immense amounts of energy emanating from human sentiment, social events, and public attention. This information can be very useful for commercial applications such as marketing analysis, public relations, product reviews, product feedback, and customer service.

Thanks to the rapid growth in the size of geospatial data, geospatial data mining has emerged as one of the most active areas of research over the past several decades. Distribution of a variety of objects such as people, vehicles, and animals has been a well-studied topic. A spatial distribution is defined as a perceptual structure, placement, or arrangement of objects on Earth. Spatial distribution is typically expressed by selecting a variable and plotting incidents of that variable on a map. A good example is Dr. John Snow's map of the 1854 Broad Street cholera outbreak in London. This map provides convincing evidence of the miasma theory of virus transmission. However, dot maps have disadvantages. For example, inappropriate size and spacing of dots can transmit biased information and mislead readers. Maps may be ineffective for communicating the message of interest if there are too many dots. As a result, these problems were addressed by the development of advanced visualization techniques such as heat maps and cluster maps. In addition, a set of spatial analysis theories and tools were developed for better describing and quantifying spatial characteristics. The goal of geospatial data mining is to reveal non-trivial patterns that were previously unknown. A spatial pattern can be a frequent arrangement, regularity, major direction, prediction, or composition that opposes randomness and causality. The tasks of geospatial data mining are diverse. For example, the goal of clustering is to assign objects into groups based on their attributes. Some tasks find

association rules that are used to correlate events that are likely to occur together or correlate events ordered in time. There are also some tasks that focus on the temporal pattern of geospatial data such as fitting regression models for a time series.

1.2 Related Work

Geospatial data mining is usually performed with two types of approaches [2]. One common approach uses statistics to test a hypothesis, such as Moran’s I, Geary’s C, Getis’s G, the standard deviational ellipse for autocorrelation [5] [6], a hidden Markov model, or a belief network. With such an approach, some forms of a statistical model are fitted to the data and then the resulting values will indicate if it is sufficiently reasonable or not to believe the expected patterns exist in this data. Another approach uses computational models to explore frequently occurring phenomena or anomalous patterns. This approach is usually associated with a specific topic and a descriptive and exploratory analysis of the results. For example, geographically weighted regression is a regression method designed for capturing spatial dependency in regression analysis [7]. Geospatial big data has become an important asset for analysis, decision-making, and resource management, but has also increased the complexity of creating responsive and scalable geospatial applications. Thus, research contributions lie in improved data simplification methods and spatial database organization strategies for advancing the query, analysis, and computation capacity of geospatial databases. The increased amounts of geospatial information being generated also highlights the need for techniques that will facilitate the discovery and visualization of patterns, anomalies, events, and interactions over space and time.

In recent years, many statistical techniques and machine learning tools have been developed for conducting geospatial data analysis. However, integrating the dimension of time into spatial analysis is challenging. Previous research focused on statistical representation of spatio-temporal patterns. The Mann-Kendall trend test was proposed by Mann and Kendall to analyze time series trends. The power and sig-

nificance of this test are not subject to the actual distribution of the data [8] [9]. Hamed investigated the Mann-Kendall trend statistic for persistent data with exact distribution. The Mann-Kendall test is applied to a group of river flow time series to confirm the effectiveness of scaling small samples [10]. Fuentes-Vallejo et al. studied cases of Dengue Fever in Girardot that included space (Getis-Ord index) and space-time (i.e., Kulldorffs scan statistics) analyses [11]. Li et al. reviewed spatial and temporal distribution of Twitter and Flickr use with socioeconomic factors [12]. Mitra et al. developed a data-driven discrete model based on a Markov random field. This method characterizes seasonal evolution of India monsoon rainfall and showed robustness with real data [13]. Yang et al. provided a two-phase algorithm for detection of asynchronous periodic patterns in time series data. This algorithm solves the problem patterns being only present within a subsequence or occurrences are shifted due to disturbance [14]. Li et al. examined using a Gaussian mixture model and kernel density estimation (KDE) to annotate mobility data. The results show that KDE is more capable of capturing the locality of word distribution [15].

Multivariate data visualization is a challenging research problem of great importance. An example of recent work in spatio-temporal interactions is VIST-STAMP, which performs multivariate clustering and abstraction with Self Organizing Map (SOM) [16]. Maciejewski et al. presented a tool to detect hot spots using kernel density estimated heat maps linked with temporal analysis views [17]. In the early 1970s, Hgerstrand introduced space-time cube by developing a graphic view of time as an additional spatial dimension [18]. This space-time model was first proposed for the visualization of movement in geographical space. Gatalsky et al. described an implementation of the space-time cube technique and showed its usefulness in detecting spatio-temporal clusters of earthquake series in Marmara, Turkey [19]. Bogucka et al. extended the cube concept to visualize cultural landscapes [20]. Kraak et al. and Huisman et al. explored the storytelling capability of space-time cube with historical events [21] [22].

Movement and mobility are important aspects of human behavior. Kim et al. proposed a gravity-based flow extraction model, which can effectively identify human movement from spatio-temporal data without using trajectory information [23]. Then, the spatio-temporal patterns are visualized by employing flow visualization techniques. Liu et al. showed how to estimate population-based vector fields using vector kernel density [24]. Transport systems are represented as vector fields for visualizing relationships between population demand and transport systems. Shen et al. represented mobility events in an area over time as an event video and built a deep neural network for mobility event prediction [25]. This model simultaneously takes into account all correlated spatial and temporal mobility patterns. Zheng et al. used machine learning to infer transportation mode. Their method first detects change points, then employs an inference model and finally implements a post-processing algorithm based on conditional probability [26]. Li et al. proposed a two-stage algorithm, termed Periodica, for mining periodic behaviors of the movement of objects [27]. The first stage combines Fourier transform and auto-correlation to capture reference locations. In the second stage, a probabilistic model is used to characterize the periodic behaviors.

In the past few years, there has been immense growth in the use of micro-blogging platforms such as Twitter. Sentiment analysis is a growing area of natural language processing for learning the polarity of words and phrases [28] [29] [30]. Wilson et al. illustrated a new phrase-level sentiment analysis approach that conducts subjectivity classification followed by polarity classification [31]. Kouloumpis et al. evaluated the feasibility of using informal and creative language, such as hashtag and emoticon, as training data in microblogging [32]. Nasukawa et al. focused on detecting favorable or unfavorable attitudes toward specific subjects [33]. Another genre of sentiment analysis focuses on societal meaning of opinions from social media. Ozturk et al. performed a comparative sentiment analysis of public attitudes toward the Syrian refugee crisis. The results indicated a different sentiment attitude between Turkish tweets and English tweets [34]. Bertrand et al. generated a sentiment map of New

York City, showing that public sentiment is highest in public parks and is lowest at transportation hubs [35]. Froehlich used a newspaper headline data set to study the public attitude toward aquaculture [36]. These headlines were manually assigned either a 1, -1, or 0 to represent positive, negative, and neutral sentiment, respectively. They found an expanding positive trend of general aquaculture coverage, while marine and offshore appeared more negative.

1.3 Overview of the Thesis

This thesis is divided into six chapters. Chapter 1 provides an introduction of this thesis and reviews previous work. Chapter 2 introduces the main methods used in this research. KDE is first introduced as an effective method for exploring spatial distribution. Subsequent sections propose two flow extraction models based on results from the KDE. Next, Moran's I and Getis Ord G_i^* statistics are illustrated to study spatial autocorrelation. The final section of Chapter 2 presents the theory of a sentiment analyzer, which is used to quantify sentiment polarity of text.

In subsequent chapters, the application of the above methods to explore patterns of human mobility and sentiment are described. In Chapter 3, two citizen geospatial data sets and the test area are introduced. Chapter 4 discusses spatio-temporal patterns of human mobility. A two-step statistical test is employed on GPS trajectory data to identify emerging hot spots and cold spots. Multiple visualization techniques are developed with flow extraction models to explore population movement related to events and traffic patterns. Chapter 5 surveys the sentiment of tweets. Each tweet is scored with a polarity value by computer. This chapter explores patterns of polarity that reveal fluctuations in human emotion across citizens, time, and space.

In Chapter 6, we review the contributions of this thesis and list possible directions of future research.

2. METHODOLOGY

This chapter introduces methods used in this thesis, including KDE, flow extraction methods based on density difference and a gravity model, and Getis-Ord G_i^* and Moran's I for spatial autocorrelation. The concept of a sentiment analyzer is illustrated at the end of this chapter.

2.1 Kernel Density Estimation

Heat map is one of the most effective technique to visualize spatial distribution of sparse data. Instead of mapping out the location of individual geographic incident, heat map highlights area with high occurrence rate. Heat maps are usually represented as a raster grid with color ramp, in which the hue encodes count, probability, density, etc. Bivariate kernel density estimation is one of the most frequent methods to accomplish the conversion from sparse point data to heat map.

Kernel density estimation (KDE), also termed as the Parzen-Rosenblatt window method, is a non-parametric approach for estimating probability density function of a dataset [37]. Intuitively, KDE has the effect of smoothing out each data point into a smooth bump, whose shape is determined by the kernel function $K(x)$. KDE sums over all these bumps to obtain a density estimator. At regions with many observations, because there are many bumps around, KDE yields a large value. Otherwise, when only a few bumps contribute to the density estimate, the density value from summing over the bumps will be low [38]. Because non-parametric estimator does not assume any underlying distribution, KDE does not calculate parameters for fixed functional form based on the data sample.

Suppose X_1, \dots, X_n are random samples from an unknown continuous distribution. The frequency density of a histogram is the number of cases per unit of the variable

on the horizontal axis. Let the bin center be x and bandwidth be h , the frequency density is

$$\begin{aligned}\hat{p}(x) &= \frac{F_n(x + h/2) - F_n(x - h/2)}{h} \\ F_n(x) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i \leq x)\end{aligned}\tag{2.1}$$

where $\mathbb{1}_A = \begin{cases} 1 & X \in A \\ 0 & X \notin A \end{cases}$ is an indicator function. Equation 2.1 can also be written as

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(x - \frac{h}{2} \leq X_i \leq x + \frac{h}{2})}{h}\tag{2.2}$$

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)\tag{2.3}$$

where K is the uniform distribution $U(-0.5, 0.5)$.

KDE smooths frequencies over the bins by replacing the above uniform distribution with a kernel function. Formally, a univariate KDE can be expressed as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)\tag{2.4}$$

where $X_1, X_2, \dots, X_n \in \mathbb{R}$ are independent, identically distributed random samples with density function p . Positive number h is the smoothing bandwidth that controls the amount of smoothing. $K : \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel function. A kernel is a special type of probability density function (PDF) with the following properties [39]:

- (a) $K(u)$ is symmetric about the origin, $\int uK(u)du = 0$
- (b) $K(x) \geq 0$ and $\int K(u)du = 1$
- (c) $\int u^2K(u)du > 0$ and $K(u)$ has finite second moment.

The main role of kernel function is to confer differentiability and smoothness properties on the resulting estimate [40]. The choice of the kernel function is not crucial to the accuracy of kernel density estimators [41]. There are a range of kernel

functions: uniform, triangular, biweight, triweight, normal, and others. Gaussian kernel is the most commonly used one and we use this kernel function throughout:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad (2.5)$$

An analogous estimator for multi-dimensional data is the multivariate kernel density estimator. For a random sample $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$,

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i) \quad (2.6)$$

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

where $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id})^T$, $i = 1, 2, \dots, n$. \mathbf{H} is the symmetric and positive definite $d \times d$ bandwidth matrix. Kernel function K is a symmetric multivariate probability density function. A multivariate normal kernel is expressed as:

$$K_{\mathbf{H}}(\mathbf{x}) = (2\pi)^{-d/2} |\mathbf{H}|^{-1/2} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{H}^{-1} \mathbf{x}} \quad (2.7)$$

The usefulness of KDE has been limited by the difficulty in finding an optimal data-driven bandwidth and a great number of bandwidth selection techniques have been developed. The measurement of the performance of \hat{f} is the mean integrated squared error (MISE) criterion [41],

$$MISE(\mathbf{H}) = E \int_{\mathbb{R}} (f(\mathbf{x}; \mathbf{H}) - f(\mathbf{x}))^2 d\mathbf{x} \quad (2.8)$$

The choice of \mathbf{H} is usually based on minimization of MISE over the space of all symmetric, positive definite $d \times d$ matrices:

$$\mathbf{H}_{MISE} = \underset{\mathbf{H}}{argmin} MISE(\mathbf{H}) \quad (2.9)$$

Plug-in selector is such a data-driven bandwidth selector for multivariate KDE. This method is first introduced by Wand and Jones [41], which minimizes asymptotic MISE under the assumption that f is multivariate normal distribution and K is Gaussian.

2.2 Flow Extraction Methods

This section introduces the work flow used to extract movement trajectory from a geospatial data set using two flow extraction models. Mobility information in a given space is represented by a vector field, which is a common and effective means to depict spatial interactions. A vector field refers to an assignment of a vector to each point in a subset of space. This definition is first introduced in physics to represent force fields and velocity fields. Each vector has a magnitude and a direction, which features flow passing through the corresponding point [42]. Depending on the research subject, magnitude could indicate volume, force, velocity, or any value with physical meaning. In the flow extraction model, the magnitude of the vector reflects the volume of flow.

2.2.1 Gravity Based Flow Extraction Model

Kim et al. presented a gravity-based flow extraction model, which effectively visualizes spatio-temporal patterns in data [23]. This model is essentially based on KDE and simulates Newton’s universal law of gravitation, which measures the attraction of two objects based on their mass and distance apart. Newton’s gravity model has been successfully applied in human geography to estimate the amount of interaction between two cities. Newtons gravity-based flow extraction model is expressed as:

$$Flow(x, y, t) = \sum_{p=-W}^W \sum_{q=-W}^W \sum_{r=-T}^T \frac{(KDE(x, y)|_t)^{a_0} \cdot (KDE(x_p, y_q)|_{t_r})^{a_1}}{d_{ij}^2} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (2.10)$$

where W is the kernel size in the spatial axes and T is the kernel size along the time axis. a_0 and a_1 in the function are control parameters and d_{ij} is the Euclidean distance between (x, y) and (x_p, y_q) . This gravity-based model replaces mass in the original gravity model with the probability density estimated by KDE. This model sums multiple vector fields along the time axis to determine the mobility status at a time stamp instead of during a period of time. In this model, $KDE(x, y)|_t$ is the

probability density of cell (x, y) at time t , whose value does not change with p , q , and r . $KDE(x, y)|_t$ does not affect the direction of the vector, but only its magnitude. Thus, we can move this item before the summation signs without affecting the result. From this perspective, the function of this model is to sum all unit vectors from (x, y, t) to every point in a space-time cube and assign a weight to the unit vectors based on their probability densities. Thus, the resulting vector will point to the direction with the highest probability density sum over the time series.

2.2.2 Flow Extraction from Density Difference

As an alternative to the gravity-based flow extraction model, the density difference-based flow extraction model is built upon the difference of spatial distribution between two time stamps. This model assumes that when objects increase or decrease at a location, they either come from or go to surrounding cells. Spatial distribution in this model is denoted by a matrix of probability density, which can be estimated and smoothed by KDE and visualized as a heat map. For example, Figure 2.1(a) and 2.1(b) are heat maps of population at start time t_1 and end time t_2 , respectively. Figure 2.1(c) shows the KDE difference between the start time and end time. Since the population at location k decreased between the two time stamps, this imposes an incoming flow to location i with a direction from k to i . The magnitude of this flow is the KDE difference at location k . In the same way, an existing flow from i to j is formed due to an increasing population at j , with a length of $KDE(x_j, y_j)|_{t_2} - KDE(x_j, y_j)|_{t_1}$. The final flow vector at i is the sum of flows imposed by all surrounding cells. To restrict the range of surrounding cells, a parameter window of size W is introduced into the model. This parameter decides the furthest cells to be taken into consideration. In other words, neighbors within a specified distance are weighted based on their KDE difference. Therefore, cells outside the

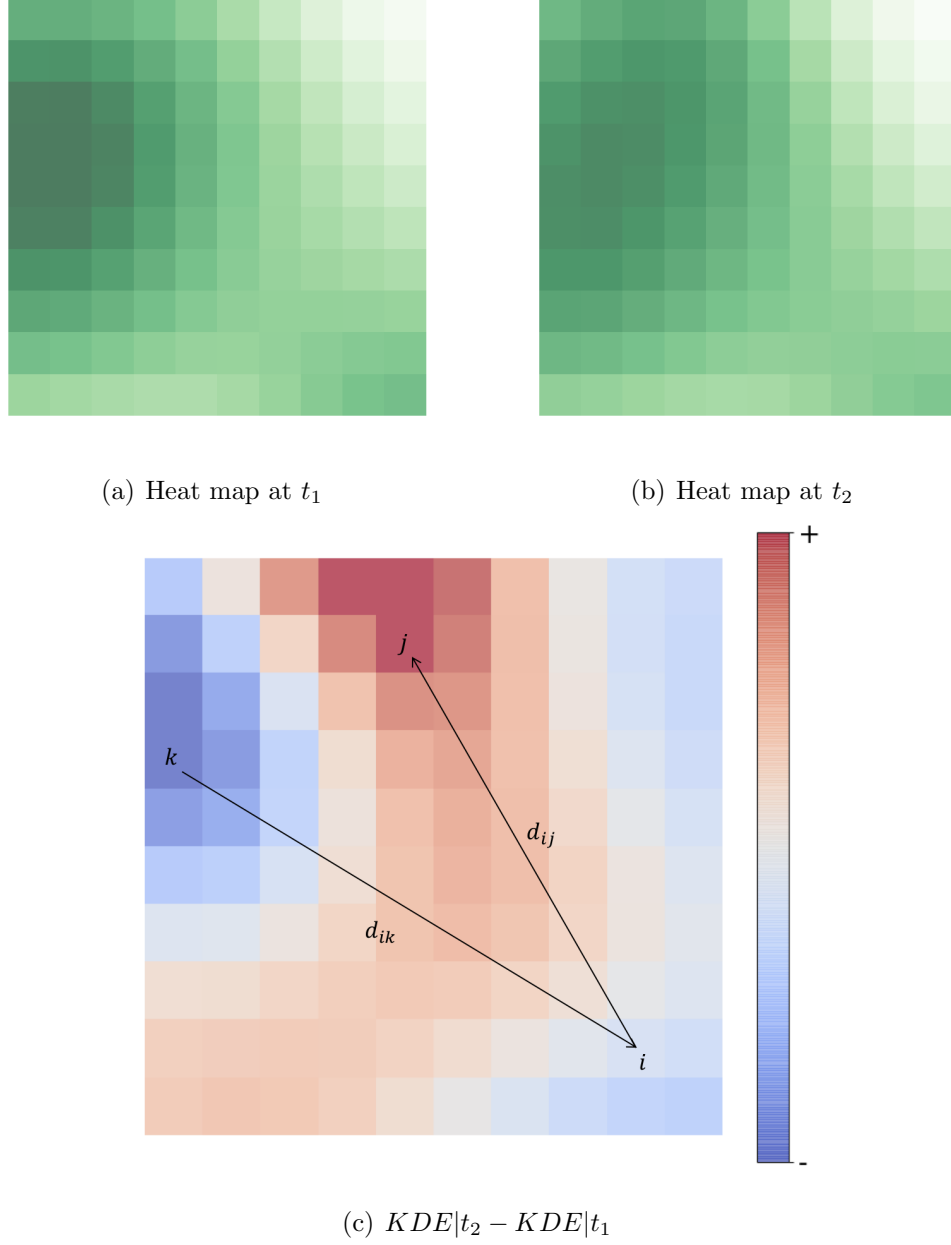


Fig. 2.1.: Flow extraction from density difference

specified distance have no influence on calculations since their weight is zero. The formula for calculating flow vector $[u, v]^T$ at any location i is given by

$$\begin{bmatrix} u \\ v \end{bmatrix} = \sum_{p=-W}^W \sum_{q=-W}^W \frac{KDE(x_j, y_j)|_{t_2} - KDE(x_j, y_j)|_{t_1}}{d_{ij}} \begin{bmatrix} p \\ q \end{bmatrix} \quad (2.11)$$

where (p, q) is the vector from (x_i, y_i) to (x_j, y_j) .

This flow extraction model can be rewritten as a form of convolution to generate a vector field of the entire area. The calculation is divided into two parts along the x-axis and y-axis, as shown in (2.12).

$$\begin{aligned} U &= (KDE|_{t_2} - KDE|_{t_1}) * kernel_p / kernel_d \\ V &= (KDE|_{t_2} - KDE|_{t_1}) * kernel_q / kernel_d \end{aligned} \quad (2.12)$$

U is a matrix of all x coordinates in a vector field and V is the matrix for y. The calculations are implemented with kernels of $(2W + 1) \times (2W + 1)$ matrix. Table 2.2 is a mask of (p, q) , whose origin is at the center. The element of the mask is given by its coordinates relative to the origin. The kernels of p and q can be obtained by splitting this mask according to the x-axis and y-axis. A kernel of distance can also be derived from this kernel through an element-wise calculation $\sqrt{p^2 + q^2}$.

Table 2.1.: Kernels of $W = 2$

Table 2.2.: Kernel (p, q)

$(-2, 2)$	$(-1, 2)$	$(0, 2)$	$(1, 2)$	$(2, 2)$
$(-2, 1)$	$(-1, 1)$	$(0, 1)$	$(1, 1)$	$(2, 1)$
$(-2, 0)$	$(-1, 0)$	$(0, 0)$	$(1, 0)$	$(2, 0)$
$(-2, -1)$	$(-1, -1)$	$(0, -1)$	$(1, -1)$	$(2, -1)$
$(-2, -2)$	$(-1, -2)$	$(0, -2)$	$(1, -2)$	$(2, -2)$

Table 2.3.: Kernel d

$2\sqrt{2}$	$\sqrt{5}$	2	$\sqrt{5}$	$2\sqrt{2}$
$\sqrt{5}$	$\sqrt{2}$	2	$\sqrt{2}$	$\sqrt{5}$
2	1	0	1	2
$\sqrt{5}$	$\sqrt{2}$	2	$\sqrt{2}$	$\sqrt{5}$
$2\sqrt{2}$	$\sqrt{5}$	2	$\sqrt{5}$	$2\sqrt{2}$

Table 2.4.: Kernel p

-2	-1	0	1	2
-2	-1	0	1	2
-2	-1	0	1	2
-2	-1	0	1	2
-2	-1	0	1	2

Table 2.5.: Kernel q

2	2	2	2	2
1	1	1	1	1
0	0	0	0	0
-1	-1	-1	-1	-1
-2	-2	-2	-2	-2

2.3 Spatial Autocorrelation

Spatial autocorrelation is a multi-directional and multi-dimensional concept in geo-statistics, which makes it useful for finding patterns in complicated data sets. Spatial autocorrelation examines the independence of observations made at different locations and measures the correlation of a variable with itself through space. There are two types of spatial autocorrelation: positive and negative. When similar values occur near one another, this variable shows positive spatial autocorrelation. In contrast, negative spatial autocorrelation occurs when dissimilar values occur near one another.

2.3.1 Moran's I

Moran's I is an inferential statistic whose null hypothesis states that the attribute being analyzed is randomly distributed among the features in the study area [5] [43]. Moran's I is a weighted product-moment correlation coefficient, expressed as:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{i,j} z_i z_j}{S_0 \sum_{i=1}^n z_i^2} \quad (2.13)$$

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{i,j}$$

where z_i is the deviation of an attribute for feature i from its mean $x_i - \bar{X}$, $w_{i,j}$ is the spatial weight between feature i and j that indicates geographic proximity, and n is equal to the total number of features. Values of I greater than 0 indicate positive spatial autocorrelation; values less than 0 indicate negative spatial autocorrelation. A z-score and p-value are computed with the Moran's I index to evaluate the significance of that index. Only when the p-value is statistically significant ($p\text{-value} \leq 0.05$) can the null hypothesis be rejected. If the z-score is positive, the spatial distribution of high and low values in the data set is more spatially clustered than would be expected if the underlying spatial processes were random. If the z-score is negative, the data

are clustered in a competitive way. For example, high values may be repelling high values or negative values may be repelling negative values.

2.3.2 Getis-Ord G_i^*

The Getis-Ord G_i^* statistic measures the intensity of clustering of high or low values of a feature relative to its neighboring features:

$$\begin{aligned}
 G_i^* &= \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{[n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2]}{n-1}}} \\
 S &= \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2} \\
 \bar{X} &= \frac{\sum_{j=1}^n x_j}{n}
 \end{aligned} \tag{2.14}$$

where x_j is the attribute value of j and $w_{i,j}$ is calculated based on the conceptualized spatial relationship. In this formula, the sum of a feature and its neighbors is compared proportionally to the sum of all features. G_i^* is essentially a z-score and p-value, which reflect whether the attribute value of a feature is different than expected and whether that difference is too large to be the result of random chance. A value of G_i^* near zero implies that the observed spatial events are randomly distributed. Conversely, positive and negative G_i^* statistics with large absolute values correspond to the clusters of high-valued and low-valued events, respectively. When using G_i^* statistic to identify hot spots and cold spots, features with a statistically significant positive z-score are hot spots and those with statistically significant negative z-score are cold spots.

2.4 Sentiment Analysis

Sentiment is the attitude or emotional reaction of a speaker. Classifying a sentence as expressing a positive, negative, or neutral opinion is known as polarity classification. Polarity is a float value falling in the range of -1.0 to 1.0 . Zero indicates a

neutral attitude, while scores greater than or less than zero represent positive and negative sentiments, respectively. The magnitude of positive or negative polarity represents the intensity of emotion. For example, polarity for “good” is 0.7 and polarity for “great” is 0.8. Although the two words both convey an attitude of praise, the emotion of “great” is stronger than “good”. Natural language processing researchers have proposed and developed many sentiment analysis algorithms, which can be classified into two types: rule-based systems and automatic systems. Rule-based systems perform sentiment analysis based on a set of manually crafted rules such as stemming, tokenization, part of speech tagging, and parsing. Automatic systems rely on machine learning techniques to learn from data. The prerequisite of sentiment analysis is to have a tool which automates the process of handling semantic information and calculating polarity. A python library “TextBlob” serves as a rule-based sentiment analyzer and the implementation is dictated as follows.

This sentiment analyzer first designs a parser to retrieve syntactic and semantic information from text in an efficient way. The parser is required to handle the following tasks:

1. Tokenization. A tokenizer divides text into a sequence of tokens, which roughly correspond to words. This step splits punctuation marks from words and finds sentence periods.
2. Tagging. Based on their use and functions, words are categorized into several types or parts of speech based on English grammar, such as nouns and verbs. This tagging tool applies a universal tagset proposed by Petrov et al., which consists of twelve categories that exists across languages [44].
3. Semantic role labeling. This task assigns labels to words or phrases in a sentence that indicate their semantic role in the sentence. Compared with tagging, role labeling has more specific categories [45]. For example, given the sentence “The boy hit that ball.”, the task would be to recognize the different roles of “boy” and “ball”. In this sentence, “boy” is the subject and “ball” is the object.

4. Lemmatization. Words usually have different forms, based on grammatical usage. Lemmatization refers to the process of finding the base form of a word. For example, the base form of “was” is “be”.

The parser is based on Brill’s algorithm, which automatically acquires a lexicon of known words and a set of rules for tagging unknown words from a training corpus [46]. Lexical rules are used to tag unknown words, based on the word morphology (i.e., prefix, suffix). Contextual rules are used to tag all words according to the role of the word in the sentence. Named entity rules are used to discover proper noun phrases such as a persons name, organizations, and locations. The parser separates and segments a sentence into its subconstituents of semantically related words. The output of this process is a list of multi-token sequences called phrase chunk.

Calculation of polarity is executed at the chunk level. The polarity of a sentence is obtained by averaging each non-zero polarity of a phrase chunk. This sentiment analyzer uses a sentiment lexicon to discern objective facts within a context. Words in this lexicon are tagged per sense. Each sense has scores for polarity and intensity. Polarity is a float value within $(-1.0, 1.0)$, indicating whether a sense is negative or positive. Intensity represents the effect of this word in modifying the next word, which is between 0.5 and 2. For a known word, polarity is used to calculate the average of all its senses in the lexicon. For example, ”good” appears twice in this lexicon with different meanings and the value used to calculate polarity is their average 0.7. Since numerous grammar combinations exist in the English language, this program provides solutions on a case-by-case basis. An adjective may be preceded by a modifier to change its intensity. Typical words used as modifiers are ”very”, ”many”, and ”super”. In this case, polarity is multiplied by the intensity of the modifier. Thus, the polarity of ”very good” is $0.7 \times 1.3 = 0.91$. When an adjective is preceded by a word of negation, such as “not”, sentiment is switched from one side to the opposite side. This program handles this situation by multiplying the original polarity by -0.5 . As a result, polarity of “not good” is supposed to be $-0.5 \times 0.7 = -0.35$. A more complex function is provided when a phrase has both a preceding modifier and a

Table 2.6.: Sentiment lexicon

Word	Sense	Polarity	Intensity
good	having desirable or positive qualities especially those suitable for a thing specified	0.7	1.0
good	tending to promote physical well-being	0.7	1.0
very	used as intensifier	0.2	1.3

negation, which is $-0.5 \times \text{polarity}$ divided by the intensity of the modifier. Other than these three common cases, additional solutions are implemented to solve complicated situations and corner cases.

Let the sentence “There was a very adorable cat sitting on that beautiful mat” be an example. According to Figure 2.2, the parser first assigns each word in the sentence to a part-of-speech tag. Next, the sentence is divided into phrase chunks: “There”, “was”, “a very adorable cat”, “sitting”, “on that”, and “beautiful mat”. The program calculates the polarity of “a very adorable cat” as being $1.3 \times 0.5 = 0.65$ because of “very” (intensity: 1.3) and “adorable” (polarity: 0.5). The polarity of “beautiful mat” is 0.85, which is the same as the polarity of “beautiful” in the lexicon. The sentiment assessment for the rest of the phrase chunks is zero. Thus, the entire sentence is given by $(0.65 + 0.85)/2 = 0.75$. Examples of some tweets are listed in Table 2.8.

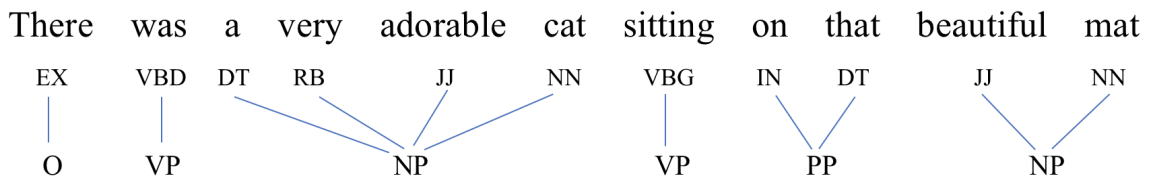


Fig. 2.2.: Sentence parser

Table 2.7.: Sentence parser annotation

Tag	Description
EX	existential there
VBD	verb, past tense
DT	determiner
RB	adverb
IN	conjunction, subordinating or preposition
NN	noun, singular or mass
VBG	verb, gerund or present participle
JJ	adjective
VP	verb phrase
NP	noun phrase
PP	prepositional phrase
O	not part of a chunk

Table 2.8.: Polarity of tweets

Tweet	Polarity
Purdue you failed us all	-0.5
Just look at that smile!!! :D	0.585938
#Boilermaker class of 2017	0

3. TEST AREA AND DATA

Two types of citizen geospatial data were used in the study: GPS records of human trajectory over time and geo-tagged tweets collected via Twitter API.

3.1 GPS Trajectory Data

GPS trajectory data are a set of GPS records provided by SafeGraph [47]. Positions are collected based on the GPS device of Android or iOS mobile applications. Each GPS point is associated with longitude, latitude, epoch, horizontal accuracy, and user id. GPS trajectory data correspond to two areas: 1) West Lafayette, Indiana and 2) Puerto Rico. A summary of both areas is shown in Table 3.1.

Table 3.1.: GPS TRAJECTORY DATA SETS

	Puerto Rico	West Lafayette
Start time	Aug 24, 2017, 20:00:00	Aug 21, 2017, 20:00:00
End time	Sep 29, 2017, 19:59:59	Aug 31, 2017, 19:59:59
# points	65,519,491	2,530,355
# users	168,377	21,473
Average # points per user	389	118
Projected coordinate system	NAD83(NSRS2007) Puerto Rico and Virgin Is. (EPSG:4437)	NAD83 Indiana West (EPSG:26974)

In this data set, horizontal accuracy varies from several meters to one kilometer. The first step of data pre-processing is to remove points with a horizontal accuracy of

100 meters or greater from the data sets. Next, points in both data sets are projected onto a local coordinate system, as shown in Table 3.1. In some cases, the data are not regular, based on when the phone captures and sends a GPS signal, whether the phone is powered on, and whether location services are on. This situation was verified in Figure 4.1, which is a histogram of the number of points per user in Puerto Rico in a period of 36 days. A long tail in the distribution of the number of users as a function of the number of points can be observed, where 35,488 users have less than 10 points and more than 1,000 users have greater than 6,000 points. User variation in point count may bias the study results, especially in group behavior research. Users in the right tail tend to make higher contributions to group movement patterns than users with few records. To avoid such an imbalance, a clean-up mechanism is introduced to remove redundant data points from the data set so that the interval between two consecutive records of any user never exceeds one hour.

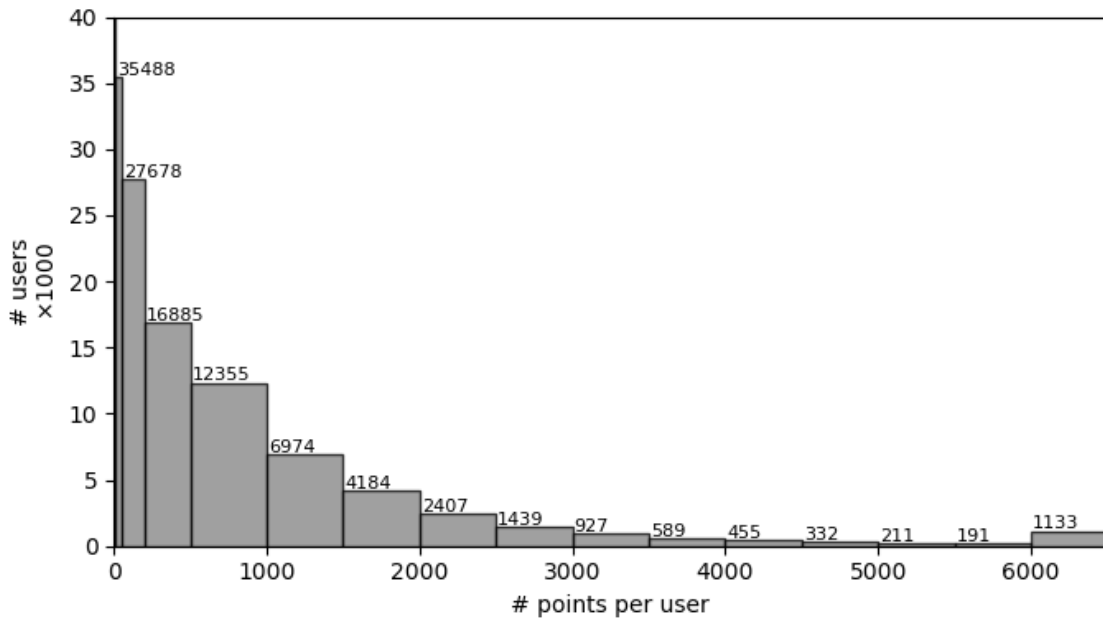


Fig. 3.1.: Number of points against number of users

3.2 Twitter Data

Public conversations have been happening on the Twitter platform since early 2006. Twitter offers a variety of tools and APIs to collect tweets and user information. A Python wrapper of Twitter’s streaming API, Tweepy, is used to download Twitter messages in real time. A geocode parameter is specified by a bounding box to query tweets that fall into that area. This Twitter data set contains tweets collected from 2014 to 2017. The research area corresponds to four US Midwestern college cities: 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). In this data set, each tweet is attached to coordinates, user id, time of posting, text of tweet, and other related information. For convenience, latitude and longitude are projected onto the Universal Transverse Mercator (UTM) coordinate system. Since Twitter’s streaming API has download rate limiting and access levels, the collected data represent a sample of total tweets. Table 3.2 lists the number of tweets collected from 2014 to 2017.

Table 3.2.: Twitter data sets

Area		2014	2015	2016	2017
West Lafayette	# tweets	65594	50735	15193	9804
	# users	3193	4352	2537	1620
	Average # tweets per user	205.43	116.58	59.89	60.52
Ann Arbor	# tweets	411037	214471	103211	53001
	# users	22328	13951	8260	5336
	Average # tweets per user	184.09	153.73	124.95	99.33
Bloomington	# tweets	440323	117270	37609	22472
	# users	11561	7824	4681	2817
	Average # tweets per user	380.87	149.88	80.34	79.77
Columbus	# tweets	3904281	1364573	523468	362904
	# users	71889	51620	29051	19936
	Average # tweets per user	543.10	264.35	180.19	182.03

4. PATTERNS OF HUMAN MOBILITY

This chapter uses GPS trajectory data to study human mobility patterns in West Lafayette and Puerto Rico. In particular, mobility regarding to events and traffics is investigated.

4.1 Density Maps

West Lafayette is home to Purdue University. This city often holds sports games that are attended by people from all around Indiana. According to the West Lafayette event calendar, there was a volleyball game between Purdue University and Alabama University on August 26, 2017. The match started at 3:00 PM and was played in Holloway Gymnasium of Purdue University. This chapter focuses on patterns during two time periods: 14:00 to 15:00 (one hour prior to the start of the match) and 16:00 to 17:00 (one hour after the match started).

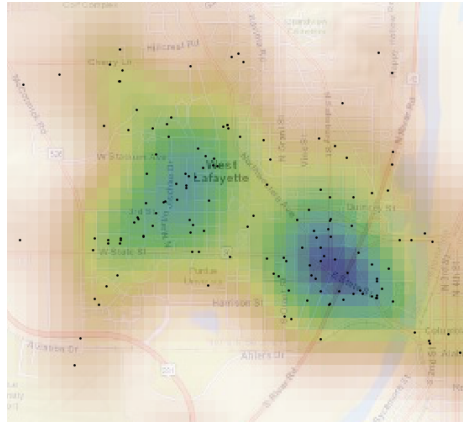


Fig. 4.1.: Map of West Lafayette

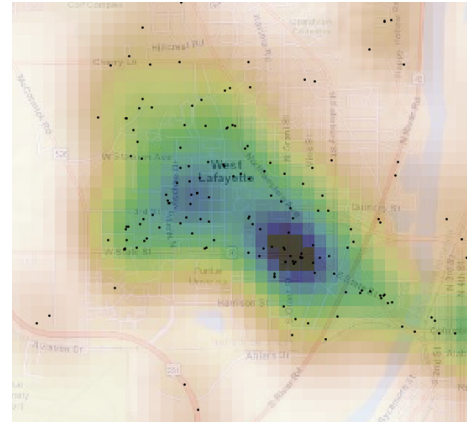
Heat Maps

One approach to simultaneously analyze spatial and temporal patterns of a geospatial data set is to apply animation control on the distribution map followed by visualization of each time step of the data. Mapping heat maps from geospatial data can be accomplished using KDE. The KDE-based implementation of a heat map is used to generate a raster graph, in which each pixel value reflects the density characteristic at the cell's location. An appropriate cell size assures that inside-cell variability is negligible so that the cells property can be represented by a single value. Heat maps of Purdue University are generated with GPS trajectory data of West Lafayette. Heat maps of six time steps in one hour are selected to display. Each pixel corresponds to a $100\text{m} \times 100\text{m}$ square area of earth. A pixel value in a heat map represents the population density evaluated by KDE at the center of each cell. When searching for points stamped exactly at a certain moment in the data set, there may be no point or only a few of such points. Therefore, it is necessary and beneficial to include points around the time of interest. Points in the range of five minutes before and after each time step are used to represent status at the time of interest. For example, a heat map 4.3(a) is actually created with points from 16:00 to 16:10. Spatial change in one hour is represented by heat maps of a time series of every 10 minutes.

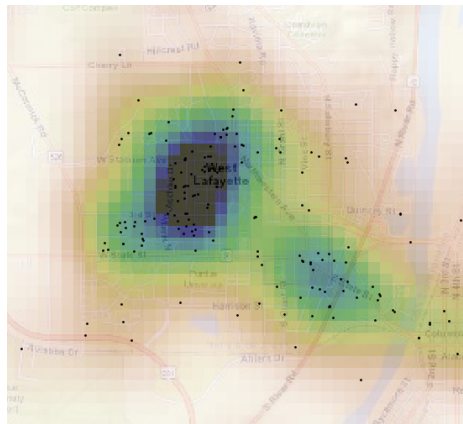
Hot spot is a common concept in spatial statistics, which indicates places with high value or density. A counterpart of this concept, termed cold spot, represents sparse areas on a map. KDE-based heat maps in Figure 4.2 and Figure 4.3 effectively convey information of a hot spot in West Lafayette. According to scatter plots, points are mostly located in the campus area, which contains many educational buildings and school facilities. This area is acknowledged as the most populated area of West Lafayette. Bright yellow circuit outlines the area on the map that contains the largest population. Dark blue marks the locations of population clusters. From 14:00 to 15:00, the spatial pattern of the population experienced a slight change. A hot spot formed near the lower right corner and grew denser. This hot spot moved upward to



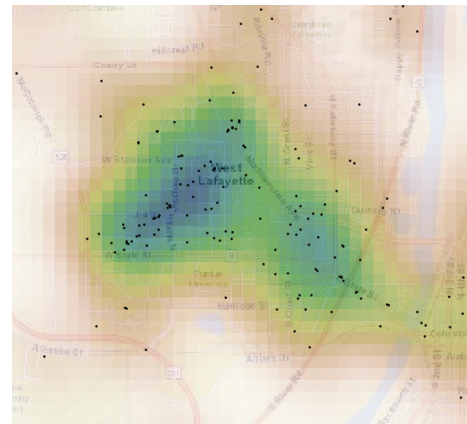
(a) 14:05



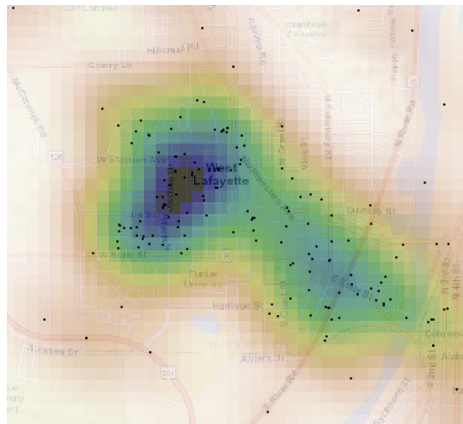
(b) 14:15



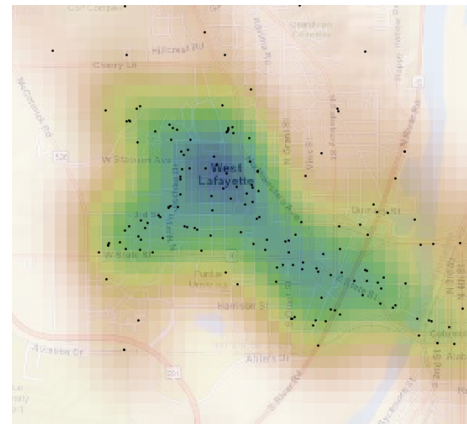
(c) 14:25



(d) 14:35

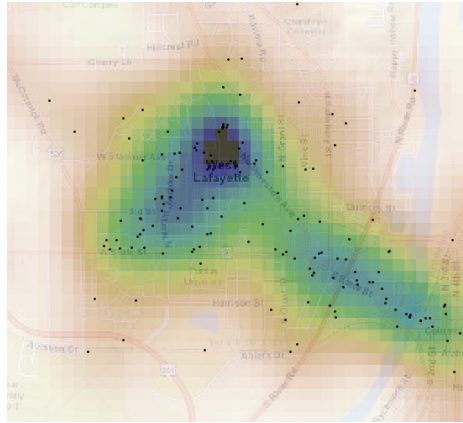


(e) 14:45

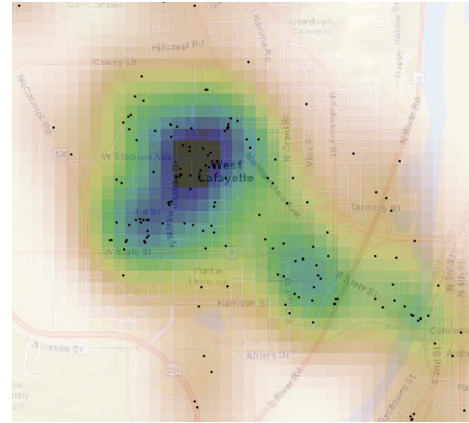


(f) 14:55

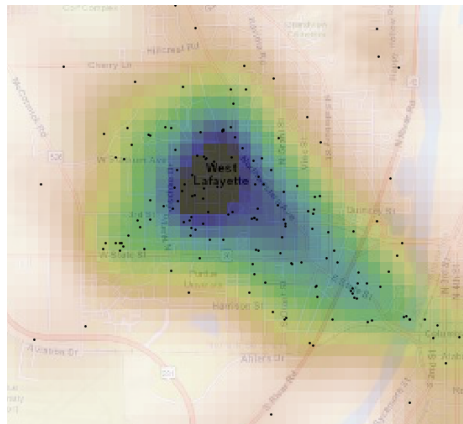
Fig. 4.2.: Heat maps of one hour before the volleyball game



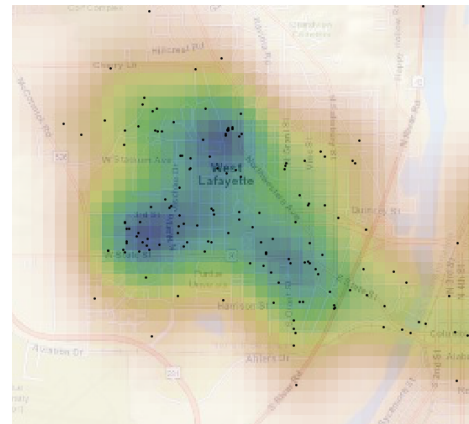
(a) 16:05



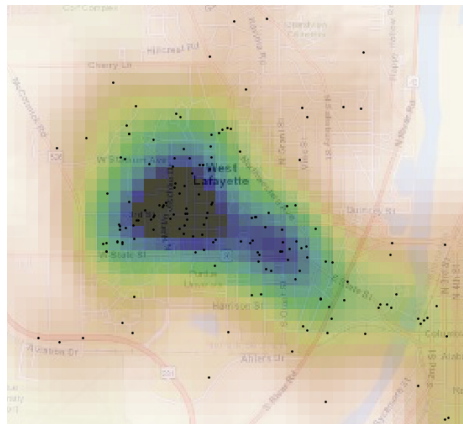
(b) 16:15



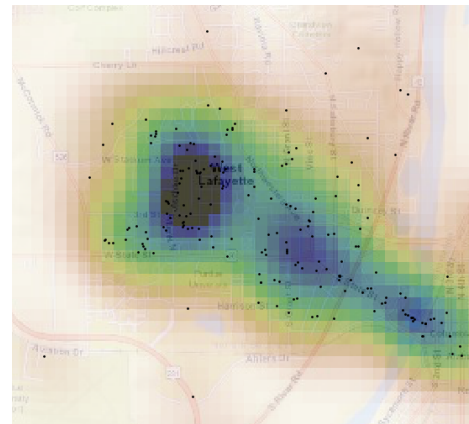
(c) 16:25



(d) 16:35



(e) 16:45



(f) 16:55

Fig. 4.3.: Heat maps of one hour after the volleyball game

the location of Holloway Gymnasium from 14:15 to 14:25, where the density and size of this cluster reached its peak at 14:25. Afterwards, the cluster stayed at the same location and faded slightly. During the one-hour period after the game, the heat map started with a hot spot located at the Holloway Gymnasium. The size of this hot spot expanded in the next twenty minutes and turned into three small clusters at 16:35. Then, the three small clusters merged back into a large cluster and extended along the diagonal. At the end of this one-hour period, a dumbbell-shaped cluster was formed at the right bottom corner. Two ends of the dumbbell were located in the downtowns of West Lafayette and Lafayette. This observation can be explained as the gradual exiting of audiences from the town after the volleyball match. In summary, heat maps visually reproduce movement of hot spots and distortion of clusters.

Emerging Hot Spot Analysis

Viewing heat maps at each time interval is a naive way to identify spatio-temporal patterns. Such an approach requires much human involvement and the explanation of results is subjective. Hence, spatial statistics were developed to study entities based on their topological, geometric, or geographic properties. The emerging hot spot analysis method is used in this section to evaluate spatio-temporal patterns using a combination of statistical measurements: Getis-Ord G_i^* statistic and the Mann-Kendall trend test. This process assigns a subcategory to each location of a hot spot to characterize its temporal trend using GPS trajectory data in West Lafayette.

Emerging hot spot analysis is conducted based on a space-time cube packed by bins containing both temporal and spatial components. Each bin has a fixed position in space and time and its value is calculated by aggregating all points within the same time and distance interval. In this study, the space interval of a bin is 100 meters and the time interval is 5 minutes. Getis-Ord G_i^* statistic measures the clustering intensity of high or low values in a bin relative to its neighboring bins in the time slice. The z-score and p-value reflect whether a bin's sum is different




than what is expected and whether this difference is too large to be the result of random chance. A high z-score and a small p-value ($p\text{-value} \leq 0.05$) indicate spatial clustering of high values. A low negative z-score and small p-value ($p\text{-value} \leq 0.05$) indicate spatial clustering of low values. For statistically significant positive/negative z-scores, the higher the magnitude of the z-score, the greater the clustering intensity of hot/cold spots. Applying the Getis-Ord G_i^* statistic to the cube generates a p-value and z-score for each bin and tags it with one of the three categories: hot spot, not significant, and cold spot.

The Mann-Kendall test is a common method for examining the existence of a trend. The Mann-Kendall test computes the difference between later-measured values and all earlier-measured values, $sign(y_j, y_i)$, where $j > i$ and integer values of 1, 0, or -1 are assigned to positive differences, no differences, and negative differences, respectively [48]. The test statistic, S , is then computed as the sum of the integers:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n sign(y_j, y_i) \quad (4.1)$$

Based on the variance of the values in the bin time series, the number of ties, the number of time periods, and the observed sum is compared to the expected sum (zero) to determine if the difference is statistically significant. The Mann-Kendall trend test is performed independently on every bin time series. Each pair of time steps was compared over the 12 time slots, generating the Mann-Kendall statistic with associated trend z-score and p-value for each bin.

[illegible]

 No Pattern Detected
  New Hot Spot
  Consecutive Hot Spot




 Intensifying Hot Spot
  Persistent Hot Spot
  Sporadic Hot Spot

Fig. 4.4.: Emerging hot spot analysis before and after the volleyball game

Table 4.1.: Emerging hot spot category

Pattern category	Definition
New	A location that is a statistically significant hot spot only for the last five minutes.
Consecutive	A location with a single uninterrupted run of statistically significant hot spot bins in the final time-step intervals. The location has never been a statistically significant hot spot prior to the final hot spot run and less than 11 of the 12 intervals ($< 90\%$) of all bins are statistically significant hot spots.
Intensifying	A location that has been a statistically significant hot spot for more than 11 of the 12 intervals ($> 90\%$), including the last five minutes. In addition, the intensity of clustering of high counts in each time step is increasing.
Persistent	A location that has been a statistically significant hot spot for more than 11 of the 12 intervals ($> 90\%$), with no discernible trend indicating an increase or decrease in the intensity of clustering over time.
Sporadic	A location that is an on-again then off-again hot spot. Less than 11 of the 12 intervals have been statistically significant hot spots.

The emerging hot spot analysis employed ArcGIS as the computational environment. Five hot spot patterns were detected on maps and are explained in Table 4.1. Figure 4.4(a) shows that, during the hour before the game, the campus of Purdue University was not a hot spot at the beginning but became a hot spot after a while. Some on-campus residential areas and commercial areas were identified as hot spots from time to time. Only a few cells were constantly hot spots during the entire one-hour prior to the game. On the second map, the location of the Holloway Gymnasium

became a cluster of persistent hot spots, meaning that this area was a hot spot for more than 90% of the total time. To the lower left of this area is an area of an intensifying hot spot. This area not only is a hot spot for the majority of the time but also displayed an upward trend throughout the hour. Some new hot spots appear near the downtown area of West Lafayette and Lafayette. The rest of the Purdue University campus was covered by sporadic spots.

4.2 Flow Maps

Before Volleyball Game

Figure 4.5 shows heat maps of West Lafayette at 2:00 PM and 3:00 PM, based on GPS trajectory data of West Lafayette. The heat maps were produced with KDE using plug-in bandwidth selector and Gaussian kernel. The pixel size is 100 meters and the value of each cell is the estimated probability density at the center. Since there were only a handful of points stamped at exactly 2:00 PM, we use points from half hour before to half hour after the game to represent the status of 2:00 PM. Similarly, all points between 2:30 PM and 3:30 PM were used to draw the 3:00 PM heat map. These two heat maps do not have any visually observable differences. At both moments, most points cluster in the Purdue campus and the downtown area of West Lafayette, forming an irregular shaped distribution. The rest of the points scatter around the map and reveal some small hot spots near the road network.

Using a window size $W = 20$, the difference-based flow extraction model is applied to the heat maps to calculate vector field. Figure 4.7(a) is the vector field map. Next, a python function "streamplot" is used to extract streamlines from the vector field and draw the flows and arrows shown in Figure 4.6. According to the flow map, the campus of Purdue University does not have the most population mobility even though this area is the most populated area on the heat maps. The most intense population movement happened near the Sagamore Parkway. The right side of 4.6 lists three local patterns on the flow map. (b) indicates a trend of population aggregation

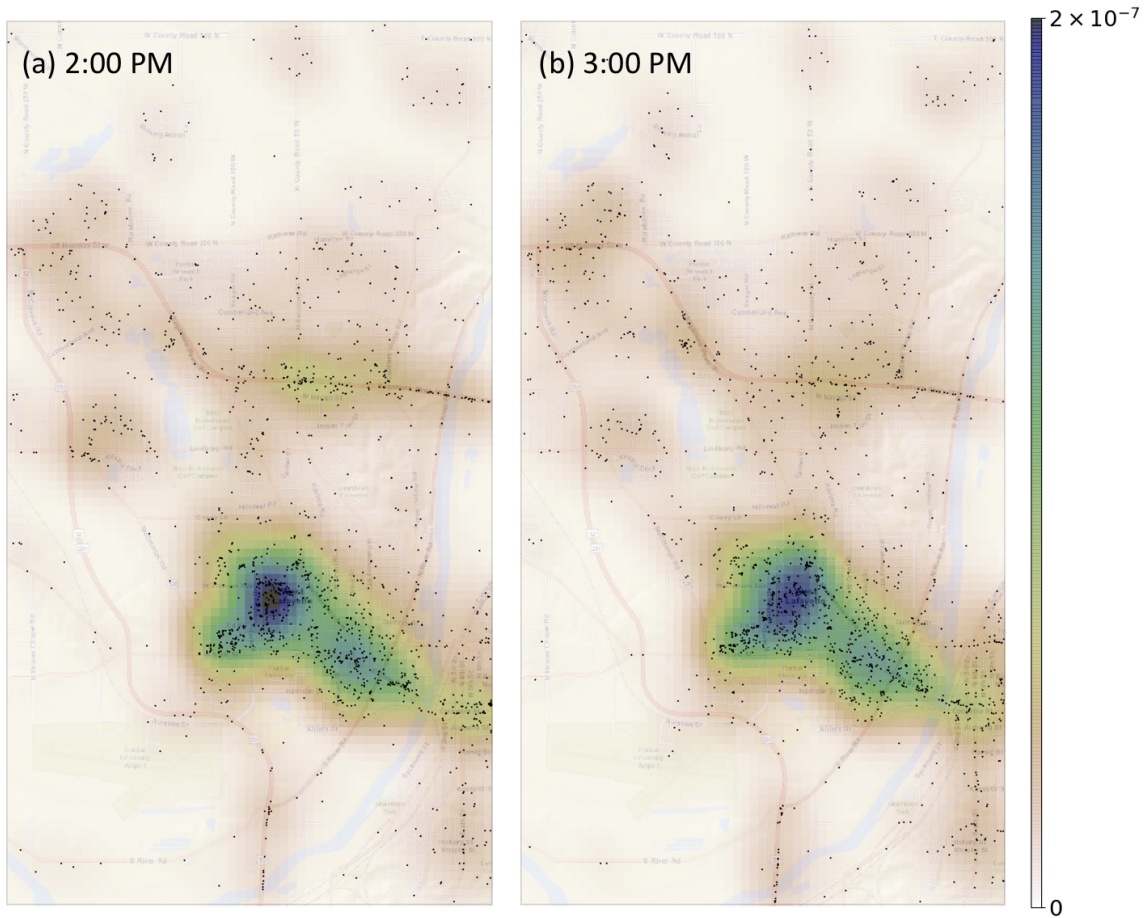


Fig. 4.5.: Probability density before and after volleyball game

around the stadium and flows are mainly from the north and east. (c) shows that, near the northwest corner of the map, there are flows that come from every direction and converge on an apartment community. Flows in (d) form a border along the Sagamore Parkway. Flows above this highway point north and flows below it are directed to the south. This phenomenon indicates that a significant number of people chose to exit this highway near this area during this one-hour period, driving either north or south.

To better identify the area in which the population diverges and converges, a direction map was drawn based on the vector field, as shown in Figure 4.7(b). Vectors

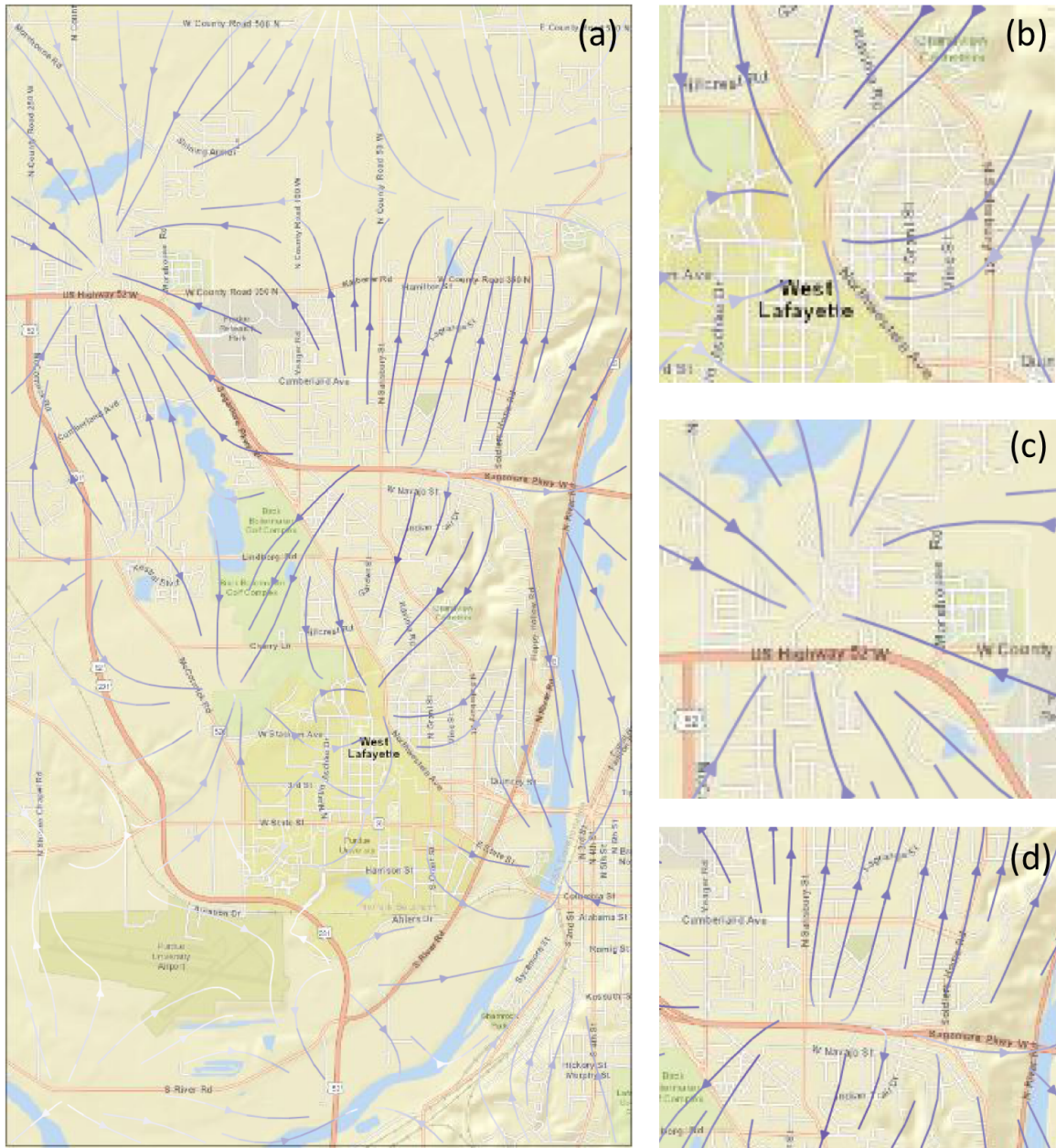
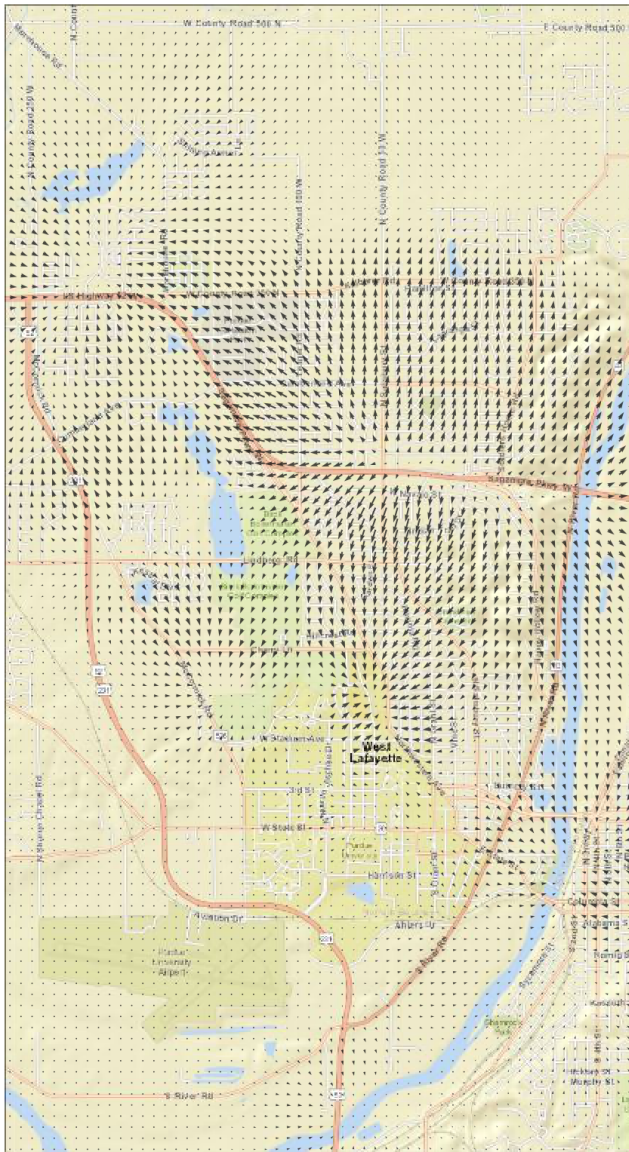


Fig. 4.6.: Density difference-based flow map before volleyball game

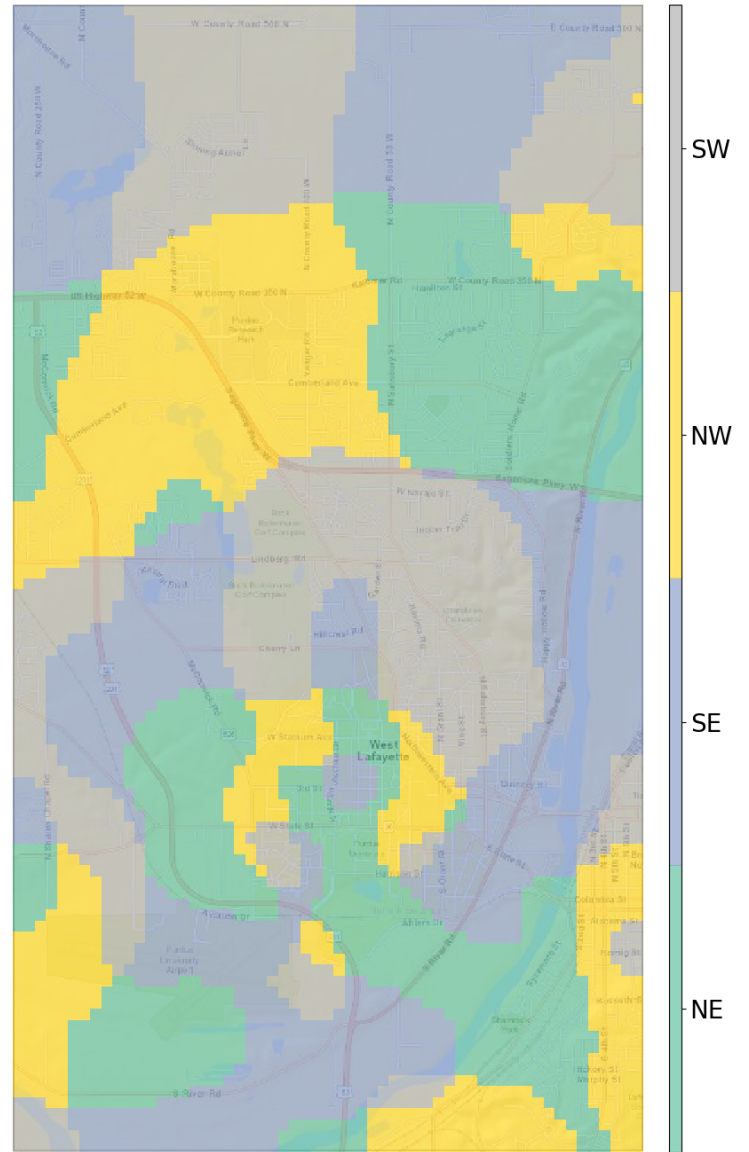
on the map are classified into four categories according to their orientation: northwest, southwest, northeast, and southeast. Four colors on the map represent four directions. There are two patterns that are especially important. The first of these denotes converging flows in Figure 4.8(a). The map is roughly divided into four parts: top

left, bottom left, top right, and bottom right. Each part is occupied by vectors from the corresponding direction. For example, a continuous and integral area painted with grey is located at the top right of the map. This indicates that the direction of population flow in this area is from northeast to southwest. Similarly, all the vectors on the map point to the center of the map such that the intersection of the four colors represents the end point of converging flows. This center is where the population increases the most and is a place that commonly holds events. Figure 4.8(a) is a zoom-in view of the direction map from Figure 4.7(b). The exact location of the stadium is marked by a red circle, which confirms that people gathered at the stadium before the game. An opposite pattern is Figure 4.8(b), which identifies the divergent area. Either pattern can be converted from the other one through a 180-degree rotation. Vectors point to the surrounding area and flows diverge the center of the pattern. As a result, the intersection of this pattern is the point that loses the most population during this period. We define the point identified by Figure 4.8(a) as the convergent point and Figure 4.8(b) as the divergent point.

Window size is a critical parameter in the flow extraction model that may influence the vector field. Figure 4.9 shows flow maps with $W = 10$ and $W = 50$ using the density difference-based method. With a smaller window size, more convergent points and divergent points are detected. Detection of additional convergent and divergent points is sometimes not significant. In contrast, latent and subtle patterns may be generalized when a larger window size is used. For example, latent and subtle patterns above the Sagamore Parkway appear on the map with $W = 10$. However, these patterns are replaced by straight south-north direction flows in Figure 4.9(b). In addition, the length of flow is usually longer with a larger window size. Thus, it is concluded that a smaller window size correlates with a more detailed flow map.



(a) Vector map



(b) Direction map

Fig. 4.7.: Density difference-based vector map and direction map before volleyball game

After Volleyball Game

Figure 4.10 shows the flow map and direction map corresponding to one-hour after the volleyball game using the density difference-based method. On the flow map, not



Fig. 4.8.: Convergent area and divergent area identified by direction map

as many subtle patterns are observed as compared to the flow map corresponding to before the volleyball game. Only one divergent pattern was observed in the campus area. According to the direction map, this divergent area is a line along Northwestern Avenue rather than being point-shaped. Holloway Gymnasium is the most northwest point in the divergent area. Near the southeast corner of the map, there are many flows crossing the river and leaving West Lafayette during this hour. The pattern in Figure 4.6 (d) does not exist on this map any more. Another newly formed divergent area sits near the intersection of Linberg Road and Highway 231. At the intersection of Highway 52 and Sagamore Parkway, there is a point of convergence.

Gravity Based Flow Extraction Model

Figure 4.11 shows flow maps created with the gravity-based flow extraction model. For both moments, the window size in the space dimension is still 20. In the time sequence, the window size is $T = 3$, which indicates that three time steps before and after the time of interest are taken into consideration. The time interval between two consecutive time steps is ten minutes. For example, Figure 4.11(a) is the flow map at

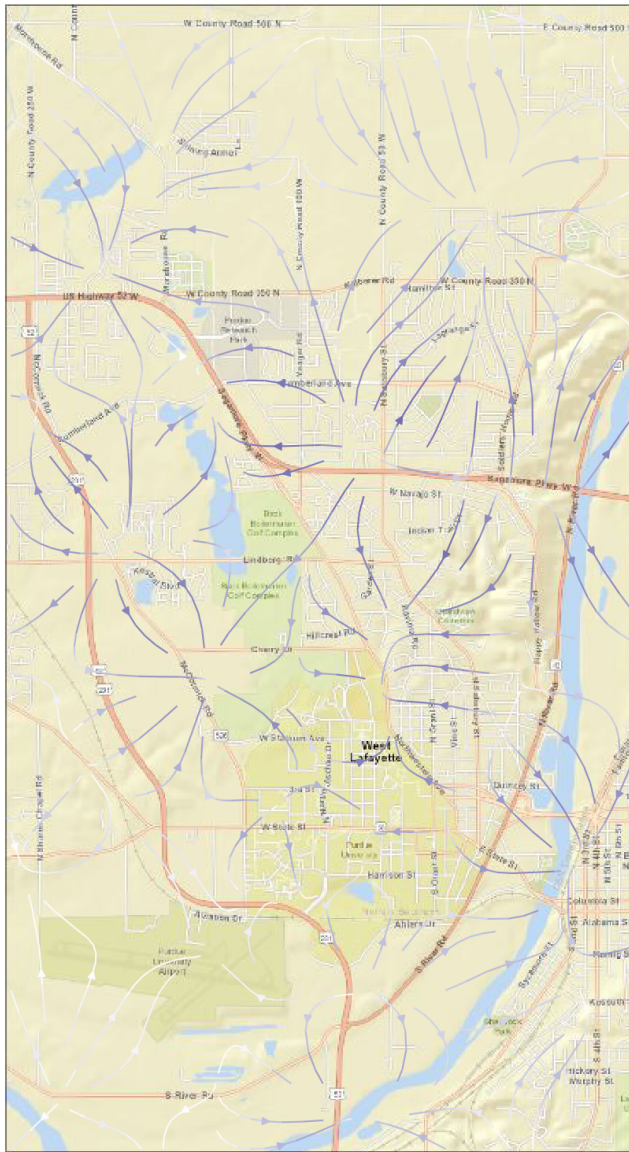
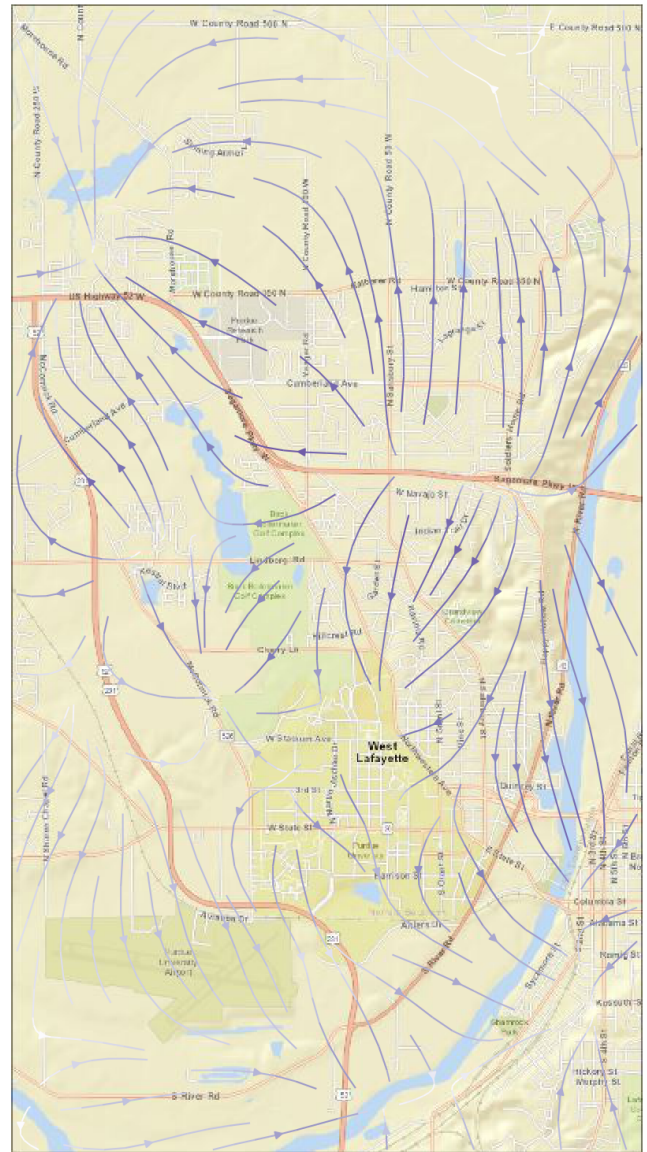
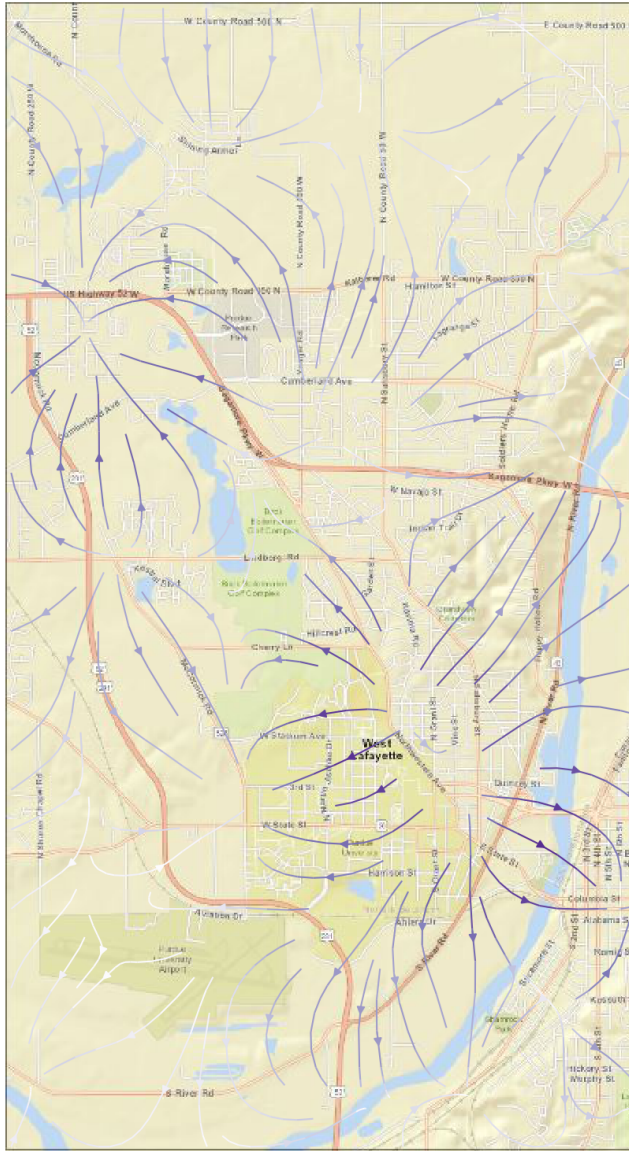
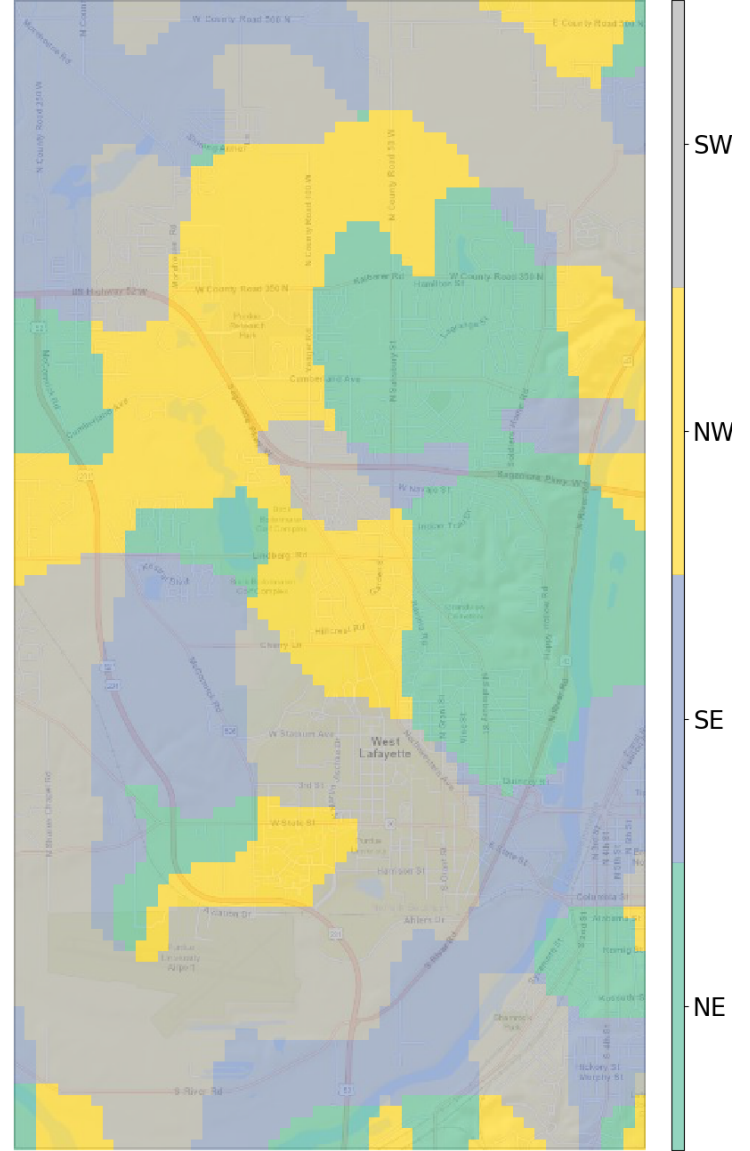
(a) $W = 10$ (b) $W = 50$

Fig. 4.9.: Density difference-based flow maps before volleyball game with different window sizes

15:00, which is made by KDEs at 14:30, 14:40, 14:50, 15:00, 15:10, 15:20, and 15:30. For each moment, KDE is calculated based on points ranging from five minutes before the volleyball game to five minutes after. Surprisingly, the flow map at 3:00 PM and the one at 5:00 PM present similar patterns. Changes in population movement in the



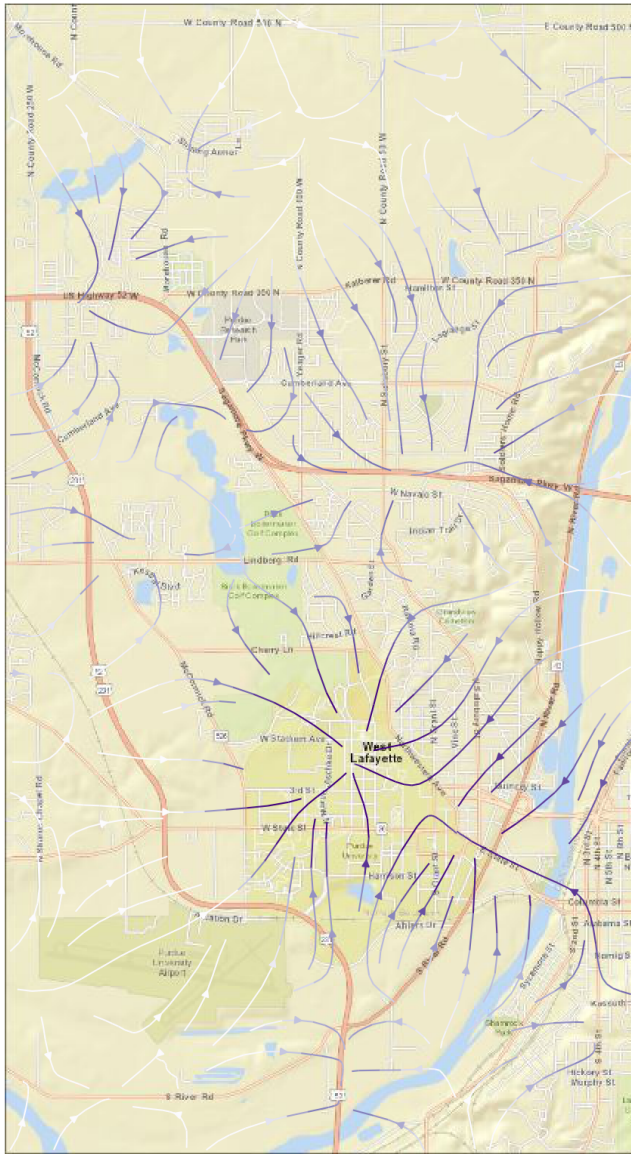
(a) Flow map



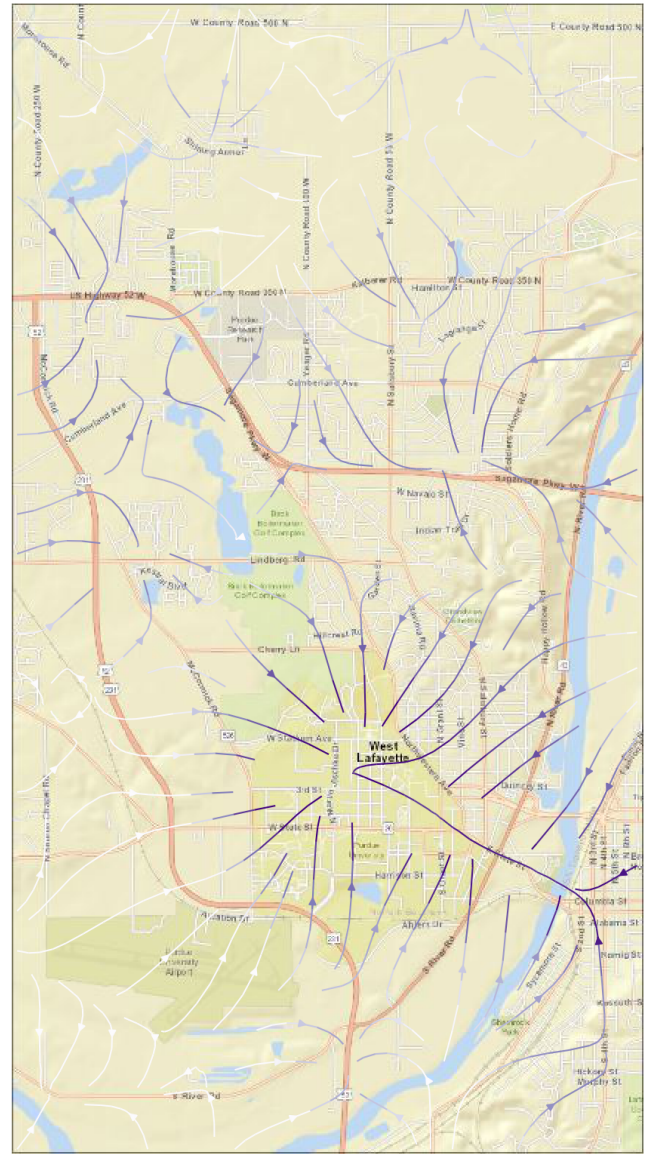
(b) Direction map

Fig. 4.10.: Density difference-based flow map and direction map after volleyball game

time interval between before and after the volleyball game were not observed. Flows in the campus area stand out with similar converging pattern on both maps. In previous work, this method successfully extracted a movement path when clusters shifted from one location to another or when significant dispersion or concentration were observed to occur in clusters [23]. However, when the time series change is too subtle to



(a) Before volleyball game



(b) After volleyball game

Fig. 4.11.: Flow maps with gravity-based flow extraction model

drastically affect spatial distribution, the gravity-based flow extraction model may fail to extract the correct movement direction. This phenomenon results from this model being density-based instead of density difference-based. Flows extracted with this model always point to the direction with the highest probability density sum over time series. This nature of the gravity-based model makes it less sensitive to

spatial change. When there is no dramatic spatial change over time, this model is indicating that people tend to leave sparse areas and travel to populated areas. Thus, it is concluded that the density difference-based model outperforms the gravity-based model when processes geospatial data with slight spatial distribution changes.

Traffic Patterns in Puerto Rico

Puerto Rico is an unincorporated territory of the United States located in the northeast Caribbean Sea. Puerto Rico has a land area of 8,870 square kilometers and a population of 3,337,177 people. The most populous city is the capital, San Juan, with approximately 371,400 people. This case study experiment was conducted using the GPS trajectory data of Puerto Rico. A normal Monday was selected for the study of movement pattern in the context of a day. Figure 4.12 demonstrates the population flows from 7:00 AM to 10:00 AM. In this map, the most noticeable trend is that the population converged in San Juan's metropolitan area in the northeast area of the map. Eight of Puerto Rico's top ten largest cities are in this area, including San Juan, Bayamn, and Carolina. Two additional populous cities are located in the south (Ponce) and west (Mayagez) of the island. They are also convergent area in the morning. In the afternoon, these cities turn into a divergent area. A trend of leaving is discerned for these places. The length of vectors is demonstrated on maps in Figure 4.14 and Figure 4.15. These maps demonstrate locations with the greatest and least change in population and reflect human mobility. Although the flow maps corresponding to the two selected times show opposite patterns, their mobility maps indicate that large cities have the largest population movements.

4.3 Summary

This chapter demonstrates analyzing mobility patterns regarding a volleyball game at Purdue University and traffic patterns in Puerto Rico. The GPS trajectory data set is used for demonstration. The volleyball game at Purdue University was intro-



Fig. 4.12.: Density difference-based flow map of Puerto Rico, 7 AM to 10 AM, Aug 28, 2017



Fig. 4.13.: Density difference-based flow map of Puerto Rico, 4 PM to 7 PM, Aug 28, 2017

duced first. Heat maps of a time series were generated based on probability density calculated by KDE to track the formation and movement of population clusters. The results show that a cluster that originated in the downtown area moved toward the gym before the game and reappeared in the downtown area at the end of the second one-hour period. The emerging hot spot analysis studied the patterns from another

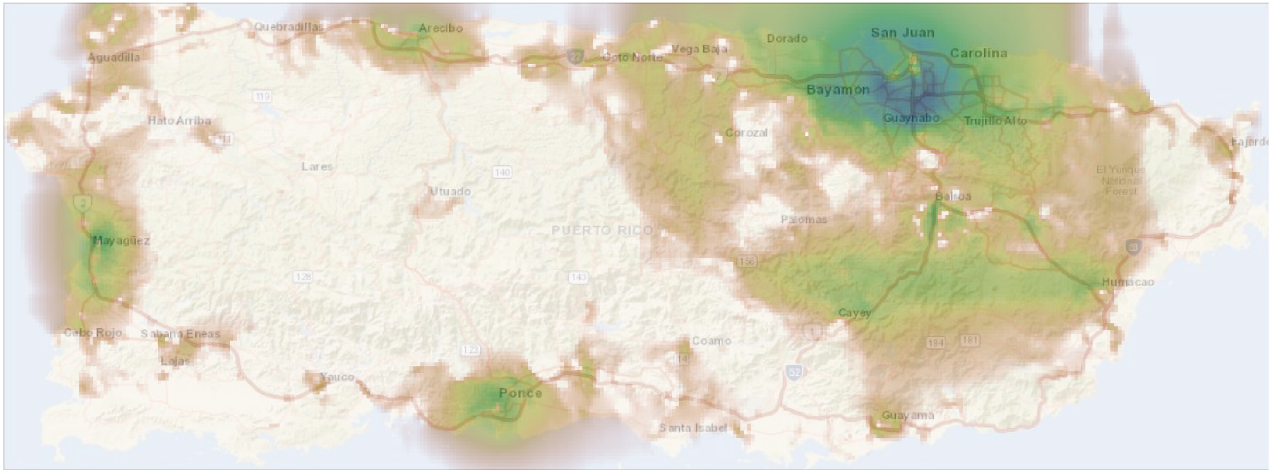


Fig. 4.14.: Density difference-based mobility map of Puerto Rico, 7 AM to 10 AM, Aug 28, 2017

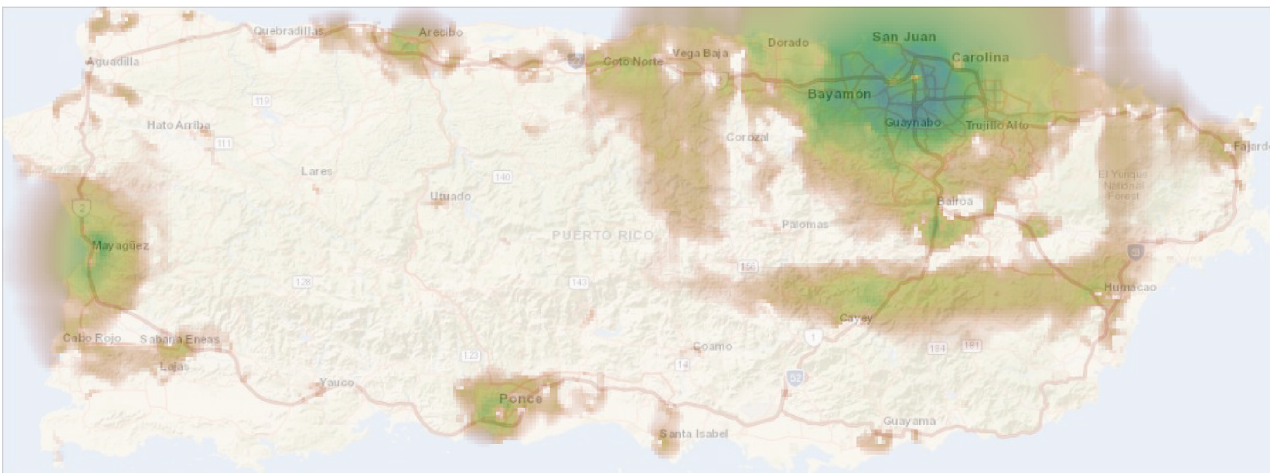


Fig. 4.15.: Density difference-based mobility map of Puerto Rico, 4 PM to 7 PM, Aug 28, 2017

perspective. This method provides a more objective and informative presentation by assigning each cell a hot spot subclass: new, consecutive, intensifying, persistent, or sporadic. The assignment is determined by its spatial measurement of the Getis-Ord G_i^* statistic across the time slice and temporal measurement of the Mann-Kendall test over the time series. The difference-based flow extraction model can extract pop-

ulation flow without using trajectory information. This model adds up vectors from itself to neighboring cells weighted by KDE difference. This section walks through this flow extraction model step-by-step using the case of the volleyball game. The concept of direction map is introduced to identify exact locations of divergent and convergent areas. This case study shows that the stadium attracted people from all directions before the game. The context within which cells are considered for computation is determined by a window size parameter. Flow maps with different window sizes deliver different visual effects and level of detail. A gravity-based flow extraction model is also implemented and compared with the proposed model. This case reveals the weakness of the gravity model when dealing with geospatial data that experience trivial changes along the time line. Another case study aimed to discover daily traffic patterns in Puerto Rico. It is concluded that people travel to big cities in the morning and leave these cities in the afternoon. In the future, we are hopeful that the visualization technique can be further improved. Due to the continuity of KDE, some flows on maps of Puerto Rico appear on the water. Such faults can be eliminated through referring to territory boundary.

5. PATTERNS OF HUMAN SENTIMENTS

Popular social media platforms such as Twitter have permitted users to send messages containing information about location, which creates approximately seven million geo-tagged tweets every day [49]. Scientists are able to access these data via APIs and track commuter patterns, as well as the volume, speed, and occupancy rates of the traffic. The objective of this chapter is to analyze temporal, spatial, and user patterns of tweet sentiment. Every tweet in a Twitter data set is scored with a polarity value according to the rules described in Chapter 2.

5.1 Citizen Sentiment Patterns

This section analyzes the distribution of citizen-based sentiment in different years and cities. Figure 5.1 contains histograms of average sentiment polarity per user in different years in West Lafayette. In these histograms, only those users who posted more than five tweets in a given year are counted. These plots indicate that, in all years, users with positive average sentiment polarity represent the majority of all users. The mean value of average sentiment polarity per user increased from 2014 to 2017.

Figure 5.2 is a histogram showing the percentage of positive and negative tweet sentiment polarity per user. Percentages equal to zero are excluded from the plot. It can be seen that bins of positive polarity percentage distribute more right than bins of negative polarity percentage. This pattern demonstrates that more people tweet with positive emotion with respect to both count and emotion intensity. Figure 5.3 shows histograms of average sentiment polarity per user in 2016 in Bloomington, Columbus, and Ann Arbor. The same conclusion also applies for these three study areas: the average sentiment polarity of most users is positive.

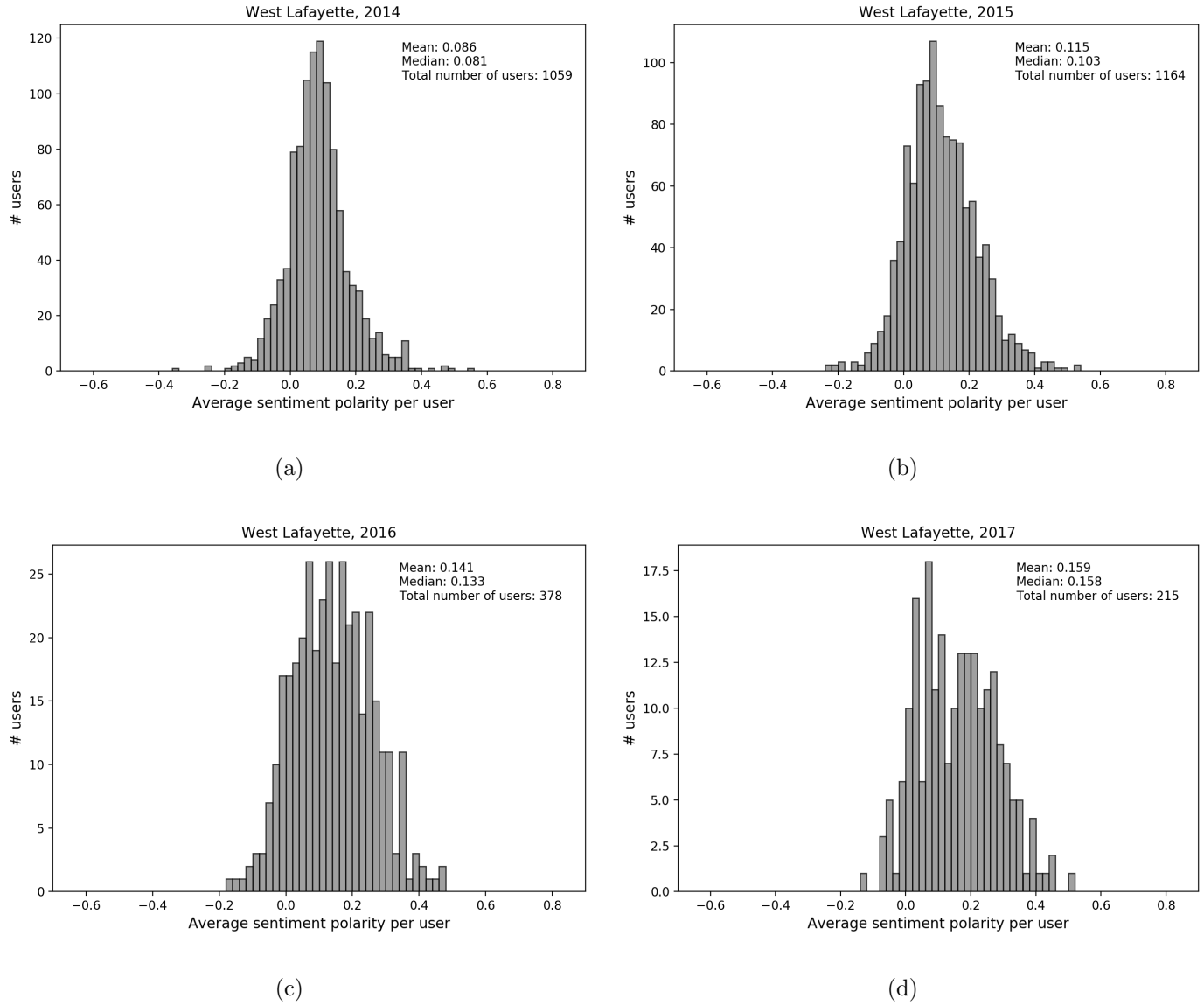


Fig. 5.1.: Histograms of average sentiment polarity per user in West Lafayette

5.2 Temporal Patterns

In Figure 5.4(a), each point is the average of polarity of all tweets posted in the corresponding weekday and year. Similarly, hourly and monthly average polarity are calculated and plotted in Figure 5.4(b) and 5.4(c) to illustrate temporal distribution in West Lafayette. It is interesting to note that a yearly difference in sentiment exists

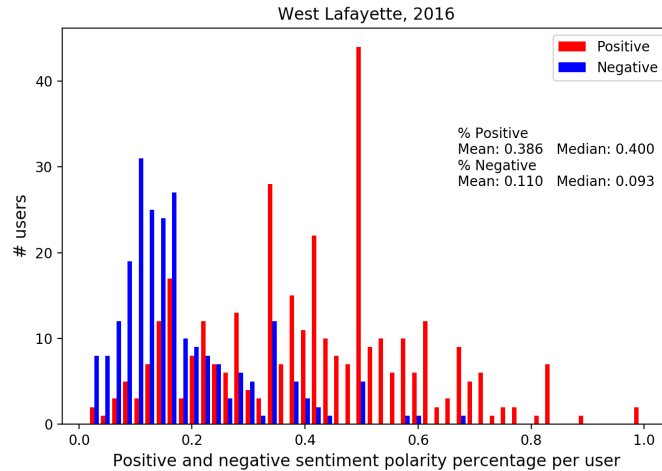


Fig. 5.2.: Histogram of positive and negative polarity percentage per user in West Lafayette, 2016

on no matter what scale. Figure 5.4(a) shows a year-by-year increase in average polarity since there exists a gap between every two consecutive years. Especially, the leap between lines of 2015 and 2016 is not negligible, which separates 2016 and 2017 from 2014 and 2015. Difference between 2016 and 2017 is indefinite since their lines tangle with each other in Figure 5.4(b) and 5.4(c). With this knowledge, it would be useful to examine the correctness of these yearly differences using statistical tests. The task of this statistical test is to determine the years in which the sample differ; in other words, which pair of two years has a significantly different polarity mean.

The pairwise comparison test compares all possible pairs from a set. In this case, four groups generate six pairs. The Tukey test, also called the Tukey's Honest Significant Difference (HSD) test, is a post-hoc test based on the studentized range distribution [50]. A post-hoc test is supposed to be performed after an analysis of variance (ANOVA) test, whose purpose is to determine whether there are any

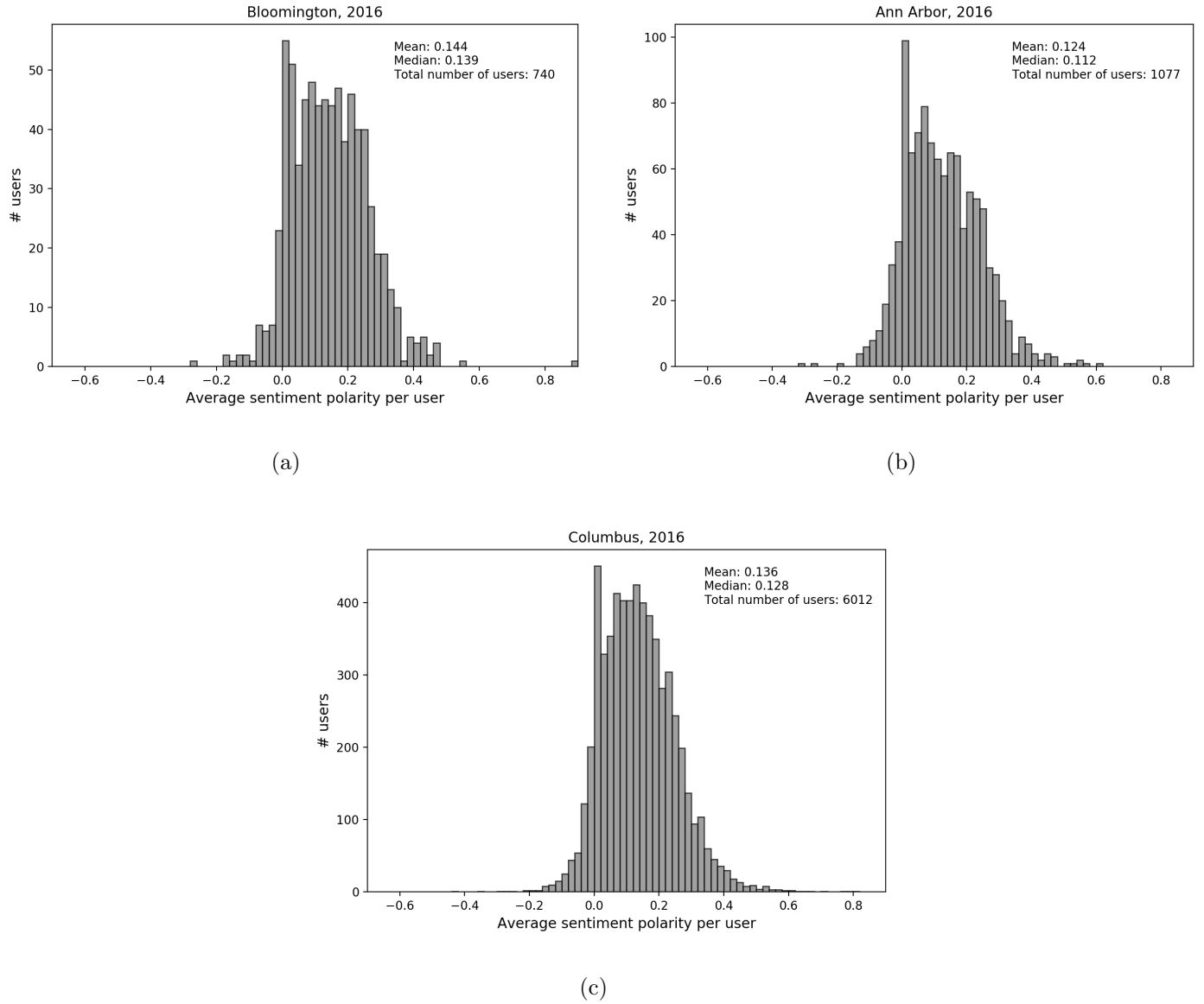


Fig. 5.3.: Histograms of average sentiment polarity per user in 2016 in Bloomington, and Ann Arbor, and Columbus

statistically significant differences between the means of two or more independent groups. Specifically, it tests the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k \quad (5.1)$$



Fig. 5.4.: Temporal distribution of polarity in West Lafayette

where μ is the group mean and k is the number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis so that there are at least two groups whose means are statistically significantly different from each other. Table 5.1 is the one-way ANOVA table of polarity in West

Table 5.1.: ANOVA table

Source	Sum of squares	Degrees of freedom	F	p-value
Between	149.007598	3.0	558.917515	0.000000
Within	12558.564334	141319.0		
Total	12707.571932	141322		

Lafayette, in which the treatment is year. The resultant p value is less than 0.05, which means that the hypothesis is rejected and the average polarity of these four years are not all the same.

Tukey test is set up to test if pairs of means are different. The hypothesis $H_0 : \mu_i = \mu_j$ refers to μ_i and μ_j of any pair. To test all pairwise comparisons, Tukey test calculates Honest Significant Difference (HSD) for each pair of means using the following formula:

$$q = \frac{\mu_i - \mu_j}{\sqrt{\frac{MS_W}{n_h}}} \quad (5.2)$$

where $\mu_i - \mu_j$ is the difference between the pair of means and μ_i should be larger than μ_j . MS_W is the mean square within and n is the number in the group or treatment when sample sizes are equal. The Tukey-Kramer method modifies the Tukey HSD test by replacing $\sqrt{\frac{MS_W}{n_h}}$ with $\sqrt{\frac{2MS_W}{n_i + n_j}}$ in the above formulas. This method tolerates unequal sample size and is more common. If q is larger than the tabulated value, the two means are significantly different. A confidence interval can be estimated by

$$\mu_i - \mu_j \pm a \sqrt{\frac{2MS_W}{n_i + n_j}} \quad (5.3)$$

Accordingly, if the calculated confidence interval contains 0, the difference between the means is not statistically significant. Table 5.3 shows the results from applying Tukey's test on the West Lafayette data set. The only two years that do not have significantly different means are 2016 and 2017.

Table 5.2.: Tukey's test for polarity mean in West Lafayette

Group 1	Group 2	Difference	Lower bound	Upper bound	Reject
2014	2015	0.0192	0.0146	0.0237	True
2014	2016	0.0868	0.0799	0.0937	True
2014	2017	0.0956	0.0873	0.1039	True
2015	2016	0.0677	0.0606	0.0747	True
2015	2017	0.0765	0.068	0.0849	True
2016	2017	0.0088	-0.0011	0.0187	False

The relation between average polarity and percentage of positive polarity can be positive or negative. Specifically, a higher average polarity can result from more positive polarity in the data. It is also possible that the absolute value of positive polarities is larger than negative ones in general. Thus, the percentage of positive polarity is another key factor for comparing sentiment. Figure 5.5 indicates that the weekday distribution of positive polarity percentage is visually similar to the figure of mean in West Lafayette. A new data set is computed through mapping all positive polarity values to 1, otherwise 0. In this case, the mean value of the new data becomes the percentage of positive polarity of the original data. Performing Tukey's test on the new data set outputs Table 5.2 with the same result. This result indicates that, among all years, only 2016 and 2017 did not have a significantly different percentage of positive polarity. Thus, the conclusion drawn from previous plots is verified by a statistical test: tweet sentiment of 2016 in West Lafayette is close to 2017 and higher than in 2015 and 2015 is higher than 2014.

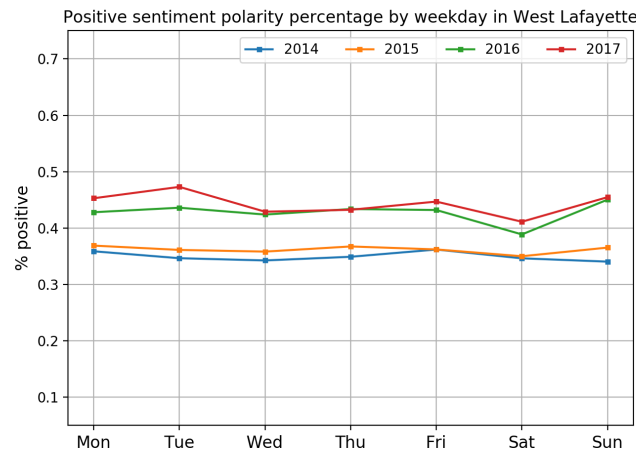


Fig. 5.5.: Weekday pattern of positive polarity percentage in West Lafayette

In addition to West Lafayette, the procedure was also applied to the other three study areas. Since the results of mean and percentage of positive polarity are not different in West Lafayette, only the mean will be considered in the following work.

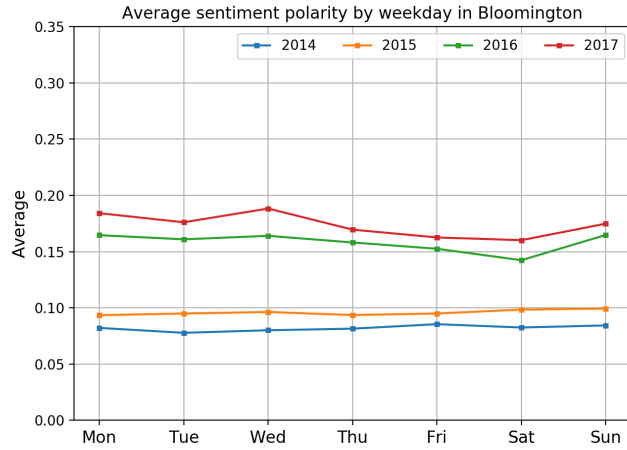
Table 5.3.: Tukey's test for percentage of positive polarity in West Lafayette

Group 1	Group 2	Difference	Lower bound	Upper bound	Reject
2014	2015	0.0121	0.0048	0.0195	True
2014	2016	0.0762	0.0651	0.0874	True
2014	2017	0.0922	0.0788	0.1056	True
2015	2016	0.0641	0.0527	0.0755	True
2015	2017	0.0801	0.0664	0.0937	True
2016	2017	0.016	-0.0001	0.032	False

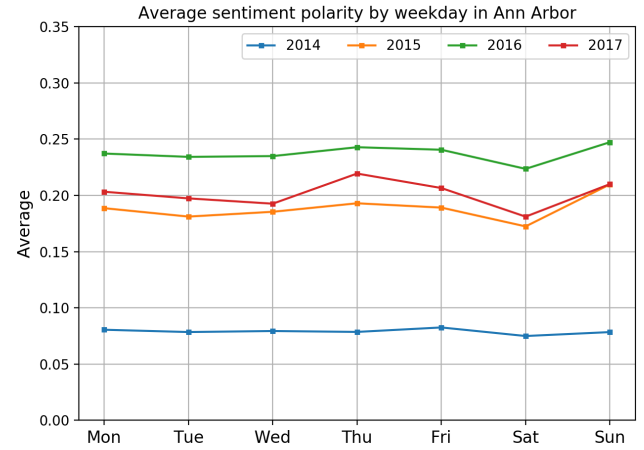
Table 5.4.: Tukey’s test result for polarity mean of Bloomington, Ann Arbor, and Columbus

ANOVA	$p < 0.05$		
Tukey’s test	2015	2016	2017
2014	True	True	True
2015		True	True
2016			True

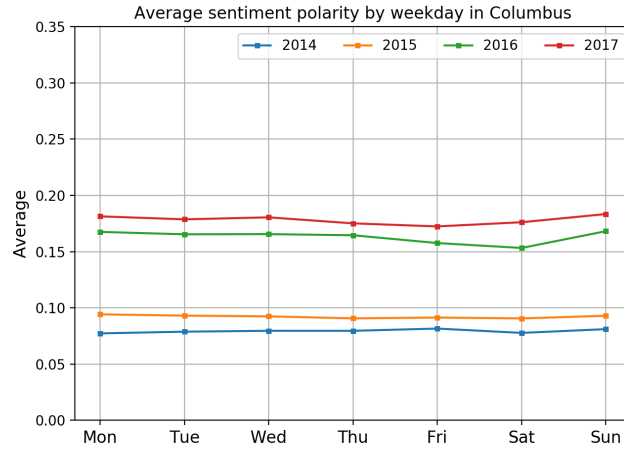
ANOVA and Tukey’s test result for all these areas are the same for the three areas and are shown in Table 5.4. There was a statistically significant difference between groups as determined by one-way ANOVA. According to Tukey’s test, there are no two years whose means are not statistically significantly different. Their weekday patterns show a yearly difference and contain specific information about pairwise relation. Plots of Columbus and Bloomington are very similar to West Lafayette, showing that the polarity mean increases from 2014 to 2017 and there is also a big gap between 2015 and 2016. However, Ann Arbor does not have the same yearly trend. Although 2014 is the lowest, the big gap appears below 2015 and above 2014, which groups 2015, 2016, and 2017 together. Year 2016 instead of 2017 has the highest mean polarity and year 2016 is between 2015 and 2017 and closer to 2015.



(a)



(b)



(c)

Fig. 5.6.: Weekday pattern of polarity in Bloomington, Ann Arbor, and Columbus

Besides yearly difference, tweet sentiment along the time line may reveal additional information. The pattern along the time line is built based on dates in the range from January 16, 2014 to December 16, 2017. For each day, average polarity is represented by the mean of all tweets posted from 15 days before to 15 days after the corresponding date. In other words, values on the y-axis indicate average polarity of

a month centered at the corresponding date. This convolution method can smooth out the curve and provides a more accurate representation of sentiment change over time. Figure 5.7 is the plot of time series distribution of four study areas. Patterns of West Lafayette, Bloomington, and Columbus are close, smooth, and steady most of the time. The plot indicates that there was fluctuation of emotion in 2014 in West Lafayette. From April to July 2014, the sentiment average of West Lafayette is comparably high and decreases steeply after August of the same year. The data set shows that the number of tweets from August to October was significantly low due to data collection (August: 336, September: 112, October: 27). It is inappropriate to draw a decisive comment about this phenomenon in case the results were biased by small-size sampling. Patterns of Bloomington and Columbus have a trend that increases over the entire time period. This increasing trend is three-step, where May 2015 and July 2016 are two turning points. In this plot, the line of Ann Arbor is prominent with many ups and downs. The middle part of the line is obviously higher than the other three starting around May 2015 till March 2017. The line goes up sharply and then drops down to near 0.2 in August 2015 and increases again before reaching its peak near October 2015. After that, the values stay at a high level and reduce gradually.

Table 5.5.: Global Moran's I result

Moran's Index	z-score	p-values
0.061633	138.395739	0.000000

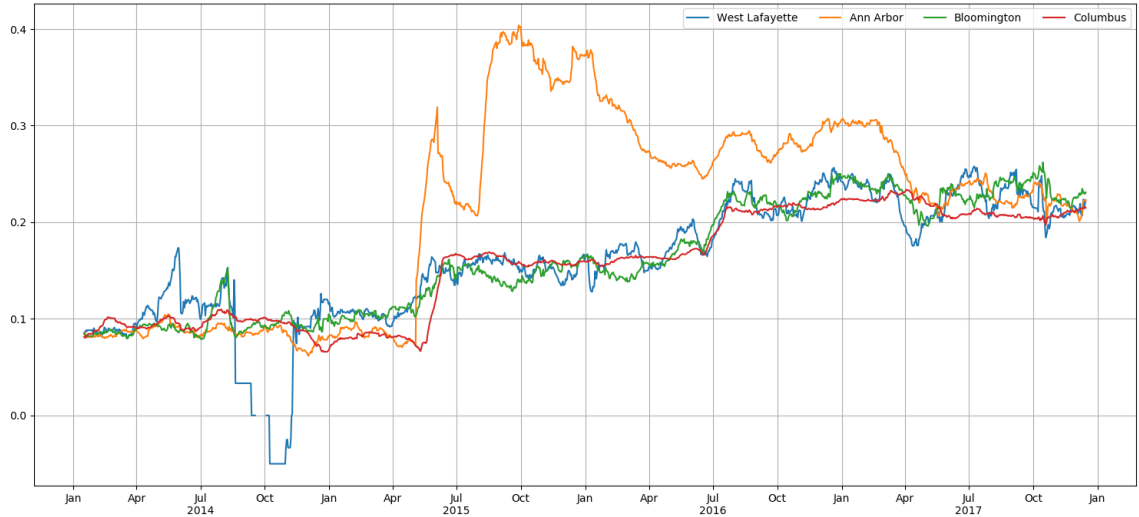


Fig. 5.7.: Sentiment polarity distribution over time

5.3 Spatial Patterns

This section analyzes spatial distribution of tweet sentiment. Specifically, this section examines if spatial autocorrelation exists in tweet polarity. It is assumed that people would be happier at some location and be in a bad mood at other locations due to characteristics and functions of different places.

Global Moran's I simultaneously measures overall spatial autocorrelation based on tweet locations and polarity. This statistic evaluates whether the spatial pattern of tweet sentiment is clustered, dispersed, or random. Spatial cluster means that high values cluster near other high values and low values cluster near other low values. If tweet sentiment clusters spatially, the Moran's Index will be positive. Given Table

5.5, there is a less than a 5% likelihood that this clustered pattern could be the result of random chance. Thus, overall, tweet sentiment is spatially autocorrelated.

Getis-Ord G_i^* statistic was introduced in the second chapter, which identifies statistically significant spatial clusters of high values and low values. In Figure 5.8, tweets are assigned to seven classes based on the resultant z-score and p-value. A G_i^* value close to zero is identified as not being significant on the map, implying that the occurrence of this tweet polarity is a random event. If a z-score is statistically significantly positive, the corresponding tweet is a hot spot and the larger the z-score is, the more intense the clustering of high values. Conversely, if a z-score is statistically significantly negative, the point will be classified as a cold spot. On the map, many points are not significant, meaning that they are neither a hot spot cluster nor a cold spot. Some red and blue points cluster around the Purdue campus and the downtown of West Lafayette, representing hot and cold spot clusters. Figure 5.9 zooms into area of Purdue University campus, giving a clearer view of cluster distribution. Blue circles and red circles mark clusters of cold spots and hot spots, respectively. Cold spot clusters are found around educational buildings, including buildings that house engineering departments and the undergraduate library. One hot spot cluster appears around an on-campus residential area at the center of the map. Another hot spot cluster is located near the Wabash Landing plaza, which is a commercial area for recreational activities. Figure 5.10 also shows that local autocorrelation exists in Columbus. This result verifies that tweet sentiment is locally autocorrelated. Thus, there are places where people tend to tweet more positive or negative statements.

5.4 Summary

This chapter places emphasis on semantic information of tweets, where sentiment of tweets is the object of study. Polarity, a float value from -1 to 1 , is a numeric representation of tweet sentiment. Sign and magnitude of the value work together to indicate sentiment polarity and intensity. This value is calculated using natural

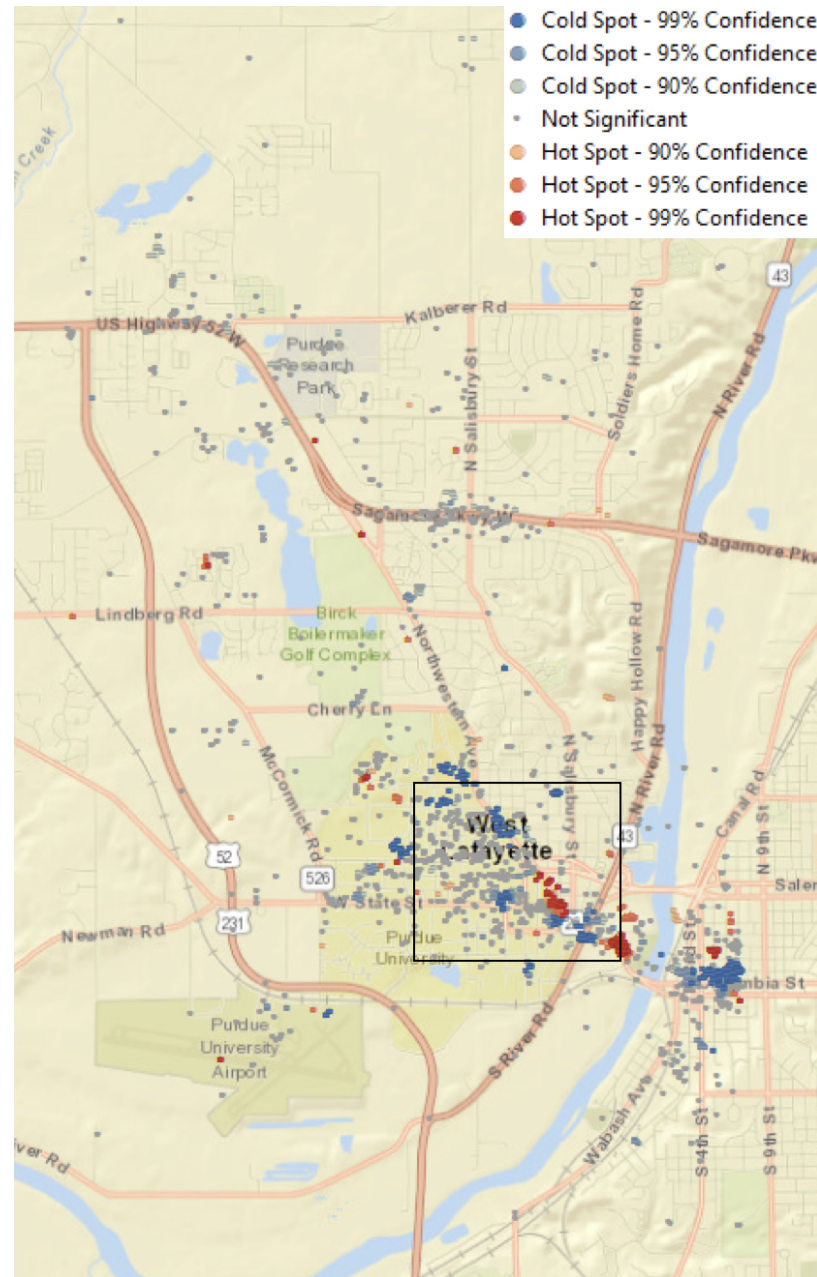


Fig. 5.8.: Hot spot analysis of West Lafayette in 2016

language processing techniques based on linguistic rules. Based on the calculated polarity of tweets, distribution of tweet sentiment is viewed from perspectives of user, time, and space. The first section analyzed sentiment distribution among users. Histograms of polarity average against number of users were generated, which supported

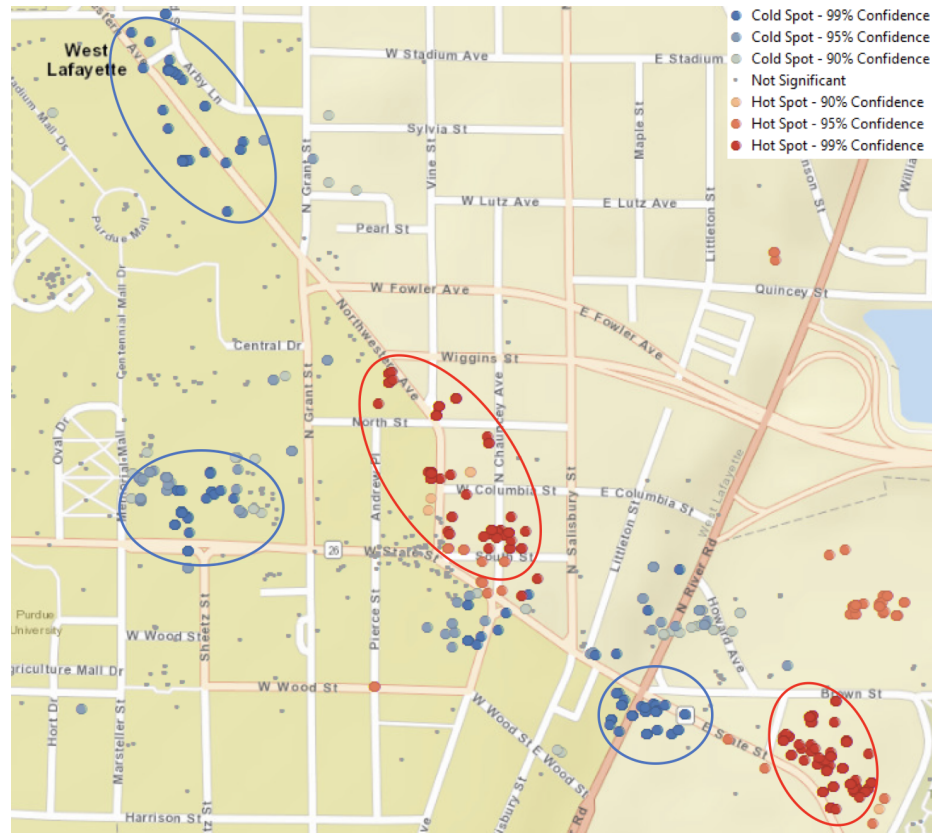
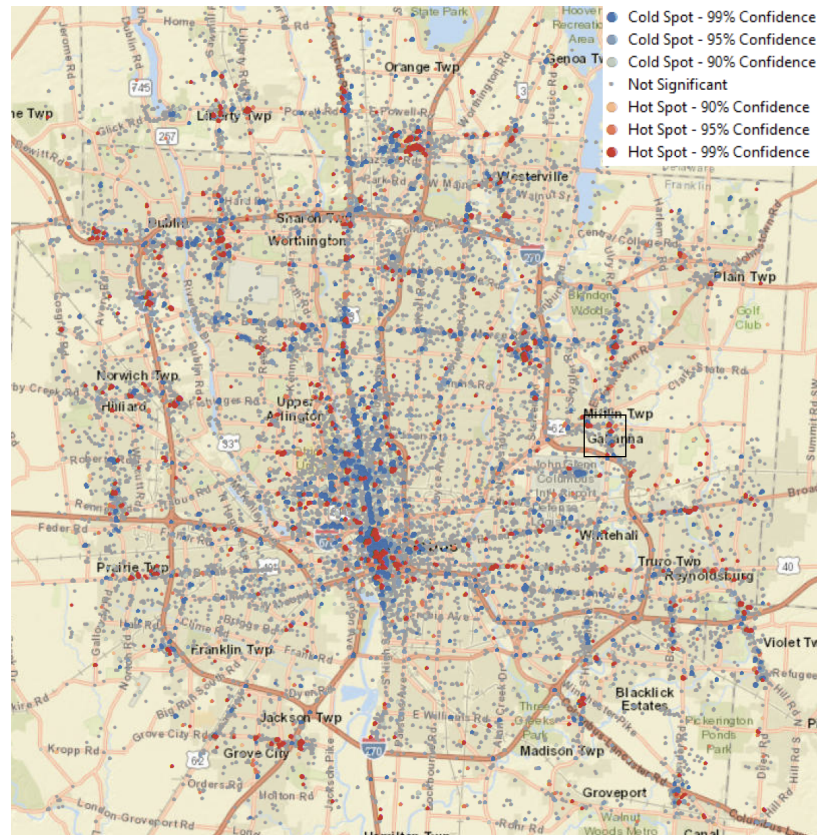
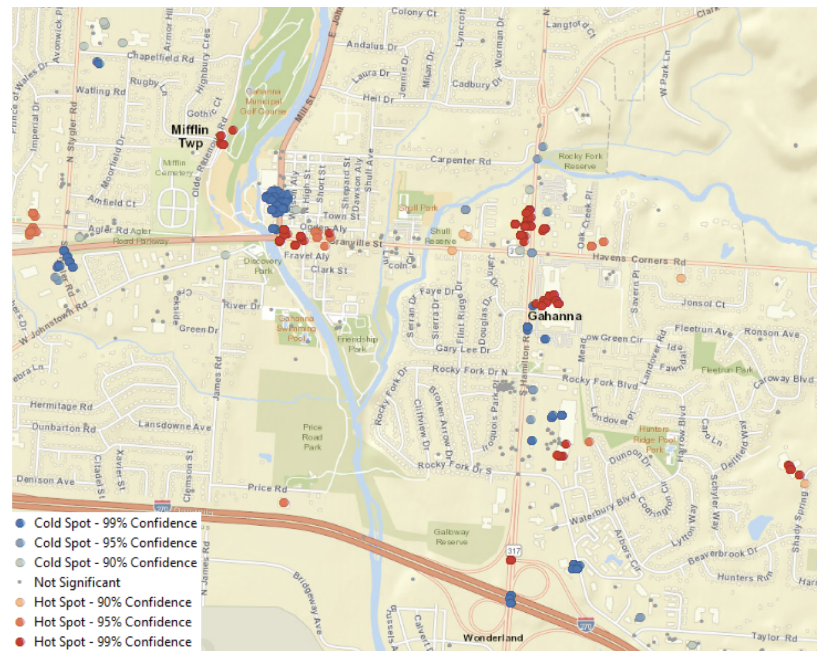


Fig. 5.9.: Hot and cold clusters of West Lafayette in 2016

the conclusion that most Twitter users post more positive statements. Temporal plots of sentiment reveal possible yearly differences. Tukey's test and ANOVA were used to test the means and percentages of positive polarity. The result showed that yearly difference does exist in all study areas with variation in the trend. Then, a time series plot was generated with a convolution method. This plot demonstrated a more detailed overview of temporal change, where Ann Arbor showed a unique and more irregular pattern than the other three. The last section of this chapter focused on spatial distribution of Twitter sentiment. Moran's I and Getis-Ord G^* statistics were employed to verify the existence of overall and local spatial autocorrelation.



(a)



(b)

Fig. 5.10.: Hot and cold clusters of Columbus in 2016

6. CONCLUSION

Describing and understanding patterns from geospatial data raises important and complex questions. This thesis explored human behavioral patterns based on citizen geospatial data. Population and sentiment were two study objects in this work. Data used for research included GPS trajectory data and social media posts collected from Twitter.

Mobility patterns associated with a volleyball game at Purdue University were explored. Spatio-temporal patterns before and after the event were analyzed using KDE-based heat maps of a time series and emerging hot spot analysis. These two approaches consistently indicated that a population cluster formed around the stadium before the game. Flow extraction models based on density difference and the gravity model were applied to GPS trajectory data to simulate a vector field representation of population movement. A density difference-based model effectively demonstrated a reverse in the direction of the flow as the game progressed. Direction maps generated based on the orientation of vectors precisely identified the location of the event. However, with the gravity-based flow extraction model, no trend was found that indicated that people left the stadium after the volleyball game. Application of the density difference-based flow extraction model on Puerto Rico indicated that human mobility is greater near big cities. This trend showed that citizens converged in San Juans metropolitan area in the morning and left this area in the afternoon.

This thesis also studied sentiment patterns of geo-tagged tweets. Each tweet was converted to a numeric value according to linguistic rules to indicate the sentiment conveyed by the tweet. Patterns of sentiment polarity were expanded according to user, time, and space. User patterns were analyzed by plotting histograms of average polarity per user. It was concluded that people tend to express more positive statements on Twitter rather than negative ones. The temporal pattern of the data

showed statistically different average polarity over four years. A time series plot detected unique trends in Ann Arbor as compared to the other three college cities. A spatial pattern study focused on spatial association of tweet sentiment. With Moran's I and Getis-Ord G_i^* statistics, tweet sentiment demonstrated both overall and local spatial autocorrelation.

There are several areas in which this research could be improved. First, one limitation of this research concerns the visualization of flows. On the flow maps of Puerto Rico, many flows were drawn on the sea. It would be useful to include the territory and land use information to adjust the result. Although transforming a vector field to flow map is beyond the scope of this thesis, a more effective visualization method should be tried in the future. Second, another limitation of the flow extraction model is that the model does not consider trajectory information. Points in the GPS data set are associated with a user id. Tracking the trajectory of a frequent user is closer to the truth of individual movement. Extraction and generalization of this information would be a great benefit to the flow extraction model. Another improvement concerns the sentiment of tweets. Since the aim of this research was to study happiness patterns of citizens, only tweets posted by individuals were supposed to be considered. A more complicated preprocessing should be implemented to filter out tweets posted by robots. A thorough explanation as to the representativeness of tweets and reasons for the detected patterns remain to be identified.

REFERENCES

REFERENCES

- [1] E. Order, “Coordinating geographic data acquisition and access, the national spatial data infrastructure, executive order 12906,” *Federal Register*, vol. 59, p. 1767117674, 1994.
- [2] N. R. Council *et al.*, *IT roadmap to a geospatial future*. National Academies Press, 2003.
- [3] W. R. Tobler, “Cellular geography,” in *Philosophy in geography*. Springer, 1979, pp. 379–386.
- [4] H. J. Miller, “Potential contributions of spatial analysis to geographic information systems for transportation (gis-t),” *Geographical Analysis*, vol. 31, no. 4, pp. 373–399, 1999.
- [5] P. A. Moran, “Notes on continuous stochastic phenomena,” *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.
- [6] R. C. Geary, “The contiguity ratio and statistical mapping,” *The incorporated statistician*, vol. 5, no. 3, pp. 115–146, 1954.
- [7] A. S. Fotheringham, M. E. Charlton, and C. Brunsdon, “Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis,” *Environment and planning A*, vol. 30, no. 11, pp. 1905–1927, 1998.
- [8] H. B. Mann, “Nonparametric tests against trend,” *Econometrica*, vol. 13, no. 3, p. 245, 1945.
- [9] M. G. Kendall, “Rank correlation methods.” *Biometrika*, vol. 44, no. 1/2, p. 298, 1957.
- [10] K. Hamed, “Exact distribution of the mannkendall trend test statistic for persistent data,” *Journal of Hydrology*, vol. 365, no. 1-2, p. 8694, 2009.
- [11] M. Fuentes-Vallejo, “Space and space-time distributions of dengue in a hyper-endemic urban space: the case of girardot, colombia,” *BMC Infectious Diseases*, vol. 17, no. 1, 2017.
- [12] L. Li, M. F. Goodchild, and B. Xu, “Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr,” *cartography and geographic information science*, vol. 40, no. 2, pp. 61–77, 2013.
- [13] A. Mitra, A. Apte, R. Govindarajan, V. Vasan, and S. Vadlamani, “A discrete view of the indian monsoon to identify spatial patterns of rainfall,” *arXiv preprint arXiv:1805.00414*, 2018.

- [14] J. Yang, W. Wang, and P. S. Yu, "Mining asynchronous periodic patterns in time series data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 613–628, 2003.
- [15] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang, "Semantic annotation of mobility data using social media," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1253–1263.
- [16] D. Guo, J. Chen, A. Maceachren, and K. Liao, "A visualization system for space-time and multivariate patterns (vis-stamp)," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, p. 14611474, 2006.
- [17] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. Cleveland, S. Grannis, and D. Ebert, "A visual analytics approach to understanding spatiotemporal hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, p. 205220, 2010.
- [18] T. Hgerstrand, "What about people in regional science?" *Papers of the Regional Science Association*, vol. 24, no. 1, p. 621, 1970.
- [19] P. Gatala, N. Andrienko, and G. Andrienko, "Interactive analysis of event data using space-time cube," *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004*.
- [20] E. Bogucka and M. Jahnke, "Feasibility of the spacetime cube in temporal cultural landscape visualization," *ISPRS International Journal of Geo-Information*, vol. 7, no. 6, p. 209, 2018.
- [21] M.-J. Kraak and I. Kveladze, "Narrative of the annotated spacetime cube revisiting a historical event," *Journal of Maps*, vol. 13, no. 1, p. 5661, 2017.
- [22] O. Huisman, I. F. Santiago, M.-J. Kraak, and B. Retsios, "Developing a geo-visual analytics environment for investigating archaeological events: Extending the spacetime cube," *Cartography and Geographic Information Science*, vol. 36, no. 3, p. 225236, 2009.
- [23] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. S. Ebert, "Data flow analysis and visualization for spatiotemporal statistical data without trajectory information," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 3, pp. 1287–1300, 2018.
- [24] X. Liu, W. Y. Yan, and J. Y. Chow, "Time-geographic relationships between vector fields of activity patterns and transport systems," *Journal of Transport Geography*, vol. 42, pp. 22–33, 2015.
- [25] B. Shen, X. Liang, Y. Ouyang, M. Liu, W. Zheng, and K. M. Carley, "Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 724–733.
- [26] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 247–256.

- [27] Z. Li, J. Han, B. Ding, and R. Kays, "Mining periodic behaviors of object movements for animal and biological sustainability studies," *Data Mining and Knowledge Discovery*, vol. 24, no. 2, pp. 355–386, 2012.
- [28] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.
- [29] P. Beineke, T. Hastie, and S. Vaithyanathan, "The sentimental factor: Improving review classification via human-provided information," in *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2004, p. 263.
- [30] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [31] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.
- [32] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsn*, vol. 11, no. 538-541, p. 164, 2011.
- [33] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 70–77.
- [34] N. Öztürk and S. Ayvaz, "Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136–147, 2018.
- [35] K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in new york city: A high resolution spatial and temporal view," *arXiv preprint arXiv:1308.5010*, 2013.
- [36] H. E. Froehlich, R. R. Gentry, M. B. Rust, D. Grimm, and B. S. Halpern, "Public perceptions of aquaculture: evaluating spatiotemporal patterns of sentiment around the world," *PloS one*, vol. 12, no. 1, p. e0169281, 2017.
- [37] J. S. Simonoff, *Smoothing methods in statistics*. Springer, 2012.
- [38] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," Sep 2017. [Online]. Available: <https://arxiv.org/abs/1704.03924>
- [39] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," p. 153158, 1969.
- [40] J. S. Racine, "Nonparametric econometrics: A primer," *Foundations and Trends in Econometrics*, vol. 3, no. 1, p. 188, 2007.
- [41] M. P. Wand and M. C. Jones, "Multivariate plug-in bandwidth selection," pp. 97–116, 1994.

- [42] A. Galbis and M. Maestre, “Vector analysis versus vector calculus,” *Universitext*, 2012.
- [43] H. Li, C. A. Calder, and N. Cressie, “Beyond moran’s i: testing for spatial dependence based on the spatial autoregressive model,” *Geographical Analysis*, vol. 39, no. 4, pp. 357–375, 2007.
- [44] S. Petrov, D. Das, and R. McDonald, “A universal part-of-speech tagset,” *arXiv preprint arXiv:1104.2086*, 2011.
- [45] D. Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.
- [46] E. Brill, “A simple rule-based part of speech tagger,” in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 152–155.
- [47] “Safegraph,” <https://www.safegraph.com/>.
- [48] T. Tech, “Statistical analysis for monotonic trends,” 2011.
- [49] L. Sloan and J. Morgan, “Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter,” *PloS one*, vol. 10, no. 11, p. e0142209, 2015.
- [50] J. W. Tukey, “Comparing individual means in the analysis of variance,” *Biometrics*, pp. 99–114, 1949.