DATA-INTENSIVE SPATIAL PATTERN DISCOVERY BASED ON GENERALIZED SPATIAL POINT REPRESENTATIONS

BY

YIZHAO GAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Geography
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Shaowen Wang, Chair
Professor Mei-Po Kwan
Associate Professor Bo Li
Professor Sara McLafferty

ABSTRACT

Geospatial big data consisting of records at the individual level or with fine spatial resolutions, such as geo-referenced social media posts and movement records collected using GPS, provide tremendous opportunities to understand complex geographic phenomena and their space-time dynamics. Such data have been widely used in many real-world applications, such as event detection and population migration analyses. These applications require not only efficient data handling and processing capabilities, but also innovative data models and analytical approaches that satisfy application-specific requirements. The aim of this dissertation research is to establish a suite of innovative methods for analyzing geospatial big data that can be modeled as generalized spatial points while addressing the following key research questions: how to estimate the spatial and spatiotemporal patterns of geographic phenomena from geospatial big data based on spatial point models? How to compare these patterns to gain insights into complex geographic phenomena? How to estimate the computational intensity of the methods? How can cyberGIS be advanced to resolve the computational intensity? Specifically, novel methods are designed in this dissertation research to exploit spatial data characteristics, innovate spatial point pattern analytics, and resolve computational intensity through high-performance spatial algorithms. Such methods are evaluated in the context of several real-world applications, including event detection from social media data and spatial movement pattern detection. Experiment results demonstrated that fine-scale spatial patterns can be revealed from geospatial big data using the proposed approaches. Novel cyberGIS software capabilities are also created as a result of this dissertation research.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and help from the following nice people.

I would first like to thank my advisor, Dr. Shaowen Wang, and my Ph.D. committee members, Dr. Mei-Po Kwan, Dr. Bo Li, and Dr. Sara McLafferty. Shaowen has been guiding me through my Ph.D. journey in the past several years. He has provided tremendous support to improve my research ability, technical and communication skills, and projection management knowledge. I am also grateful to Mei-Po, Bo, and Sara for their support and guidance during this process. Their suggestions and insights are invaluable for this dissertation.

I appreciate all their help and support from current and former members in the Department of Geography and Geographic Information Science, the CyberInfrastructure and Geospatial Information (CIGI) Laboratory, and the CyberGIS Center for Advanced Digital and Spatial Studies. I am especially thankful to Dr. Yan Liu, who has shared me with tremendous technical knowledge, and Dr. Anand Padmanabhan, Dr. Junjun Yin, Dr. Myeonghun Jeong, and Kiumars Soltani, who have collaborated with me in research papers that contributed significantly to my dissertation research. I am also fortunate to work with Dr. Shakil Bin Kashem as his teaching assistant, and from this experience I learned a lot about teaching. I would like to thank Susan Etter, Matthew Cohn, and Denise Jayne for their support to make my Ph.D. study smooth.

This dissertation research is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 0846655, 1047916, 1354329, 1429699, and 1443080. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Finally, I thank my family for their endless love and support during my Ph.D. journey. I would like to give special thanks to my wife and colleague Ting. Her unending love and support

are my constant source of courage and inspiration. She does not only manage things for our family, but also provides countless insights into my dissertation and servers as the most accessible person I could turn to for any thoughts. I also own many thanks to my parents and Ting's parents. They offered their unconditional love and support for us to study at the University of Illinois at Urbana-Champaign far away from home. They are the most fantastic parents in the world.

To my family, for their love and support

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND BACKGROUND 1

1.1 MOTIVATION 1
1.2 RESEARCH QUESTIONS 3
1.3 BACKGROUND 4
1.4 GENERALIZED SPATIAL POINT MODEL 10
1.5 THESIS ORGANIZATION 15

CHAPTER 2: MAPPING SPATIOTEMPORAL PATTERNS OF EVENTS USING SOCIAL MEDIA: A CASE STUDY OF INFLUENZA TRENDS 18

2.1 INTRODUCTION 19
2.2 RELATED WORK 21
2.3 METHODS 24
2.4 CASE STUDY: DETECTING INFLUENZA PATTERNS USING *TWITTER* DATA 30
2.5 RESULTS 36
2.6 CONCLUSIONS 47

CHAPTER 3: A MULTIDIMENSIONAL SPATIAL SCAN STATISTICS APPROACH TO MOVEMENT PATTERN COMPARISON 51

3.1 INTRODUCTION 52
3.2 RELATED WORK 55
3.3 DATA MODEL 57
3.4 METHOD 59
3.5 CASE STUDIES 66
3.6 CONCLUSIONS AND FUTURE WORK 76

CHAPTER 4: SCANNING WINDOW DESIGNS FOR MULTIDIMENSIONAL SCAN STATISTICS APPROACH TO MOVEMENT PATTERN ANALYSIS 81

4.1 INTRODUCTION 82
4.2 MULTIDIMENSIONAL SPATIAL SCAN STATISTICS AND OD DATA MODEL 84
4.3 SCANNING WINDOW DESIGNS 85
4.4 ALGORITHM AND COMPUTATION 99

4.5 EXPERIMENTS AND RESULTS 107
4.6 CONCLUSION 121
CHAPTER 5: CONCLUSION AND FUTURE DIRECTIONS 124

5.1 SUMMARY OF CONTRIBUTIONS 124
5.2 FUTURE WORK 131
REFERENCES 134

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 MOTIVATION

The increasingly availability of geospatial big data consisting of records at individual level or fine spatial units, such as geo-located social media posts and movement records collected using GPS, provide unique opportunities to understand complex geographic phenomena and their space-time dynamics (Dodge et al. 2016, Gonzalez et al. 2008, Guo and Zhu 2014). Extensive research using such data have been conducted for many applications, such as event detection, movement flow summarization and mapping, and community detections. These applications require not only efficient data handling and processing capabilities, but innovative data models and analytical approaches that satisfy application-specific requirements. However, conventional data models, analytical approaches, and computing resources are often unable to handle such big data and to answer new research questions that emerge as a result of new datasets and computing resources. The focus of this dissertation research is therefore to develop spatial pattern analytical approaches by bridging cyberGIS, new methods for spatial pattern analytics, and novel applications, in the context of geospatial big data.

The specific aim of the dissertation is to establish scalable approaches to analyzing geospatial big data that can be modeled as generalized spatial points, and to detect patterns based on spatial point representations. Being the simplest but widely used form of spatial data models, a spatial point model fits well with many geospatial big data sources containing individual-level records (Illian et al. 2008, Diggle 2013). Many of these records represent phenomena each of which is associated with one location, and thus can be modeled as points in conventional geographic (mostly 2D) spaces, such as individual geo-located social media posts; yet this dissertation demonstrates that many other more complicated phenomena that are associated with

multiple locations can also be modeled as generalized points in a higher conceptual spaces, such as origin-destination (OD) movements or bivariate spatial interactions, each individual record of which can be modeled as a point within 4 dimensions (Gao et al. 2018). The point models enable spatial pattern discovery through point pattern analysis approaches, in order to answer research questions such as pattern comparison and change detection. These point models make the detection of fine-scale spatial patterns possible. They can also ensure that no aggregation to areal units is necessary and thus the results are less prone to the Modifiable Areal Unit Problem (MAUP, Openshaw 1984). The drawbacks of using point models on geospatial big data, however, are that the quantity of data used through analytical processes is usually large, and thus point analysis approaches such as kernel density estimation (KDE) and scan statistics are computationally intensive.

Performing data-intensive spatial and spatiotemporal point analysis requires computation- and data-intensive capabilities and thus cyberGIS (that is, geographic information science and systems based on advanced computing and cyberinfrastructure, Wang 2010; Wright and Wang 2011). CyberGIS has been advanced to resolve the challenges of analyzing geospatial big data by seamlessly integrating advanced cyberinfrastructure and high-performance computing resources to achieve computation- and data-intensive spatiotemporal analytics. CyberGIS does not only speed-up conventional spatial analysis tasks (e.g., to performance kernel density estimations from millions of input points), but provides unprecedented opportunities for answering research questions that were impossible before (e.g., detecting spatial clusters in multidimensional conceptual spaces using scan statistics approaches). Thus, this dissertation research is well positioned to advance cyberGIS by transforming data-intensive spatial analytics while posing new questions and enabling new approaches to such questions (Wang and Goodchild 2018).

Algorithms and software tools are increasingly essential components of scientific research (Lazer et al. 2014). New data-intensive pattern analysis tasks are impossible without the support of efficient and reliable software. The technical implementations, assumptions, and specifications of software tools are also demonstrated to have a large influence on research results (Kwan 2016). Developing and sharing open-source software codes and tools are a core contribution of this dissertation research since other researchers can easily reproduce our methods and understand the technical details associated with the dissertation. These software resources will also benefit the scientific community in that other researchers can apply our methods to new applications or to continue modifying and improving our methods. Hence, in this dissertation research, significant scientific and engineering efforts are also put into the development of new software for achieving the aforementioned purpose.

1.2 RESEARCH QUESTIONS

The overarching goal of this research is to establish a set of novel data-intensive methods for knowledge discovery from geospatial big data based on a generalized spatial point model. Using real-world applications such as event detection from social media data, taxi traffic analysis, and migration pattern analysis, this dissertation demonstrates show how hidden spatial patterns can be detected through methodological innovation. Specifically, novel methods are designed to synergistically advance cyberGIS and geospatial data science in the big data era through innovative integration of spatial analysis methods, algorithms, computation, and application-specific characteristics by addressing the following research questions:

- How to reveal spatial and spatiotemporal patterns from geospatial big data based on spatial point models?

- How to compare these patterns for gaining insights into the complexity of such patterns?
- How to estimate the computational intensity of the methods?
- How can cyberGIS be advanced to resolve the computational intensity?

In order to answer these research questions, the scope of the dissertation research encompasses the following objectives:

- Develop a set of models, algorithms, and methods for data-intensive spatial point pattern analysis;
- Establish novel cyberGIS approaches to resolving related computational challenges;
- Develop software codes and tools for the developed approaches; and
- Evaluate the approaches in the context of multiple real-world applications.

1.3 BACKGROUND

1.3.1 Geospatial Big Data Analytics

Recently, complex and massive geospatial data, such as GPS tracking records, large-scale surveys, and social media, have become increasingly available due to technological advances and significant application demands (Kwan 2016). Geospatial big data collected from various sources, have the general properties of big data, that are the 4Vs (Shu 2016, Lee and Kang 2015, Barwick 2012, Hilbert 2015). Volume refers to data amount or quantity. Variety means that data can be of multiple types or from multiple sources. Velocity refers to the high speed at which data are generated. Veracity means that data conform to facts with desirable accuracy. The availability of big data and new data analytics provide tremendous opportunities for

understanding and modeling complex geographic phenomena. They challenge established epistemologies and engender the paradigm of data-driven science (Kitchin 2014, Miller and Goodchild 2015).

Many types of geospatial big data are in the form of individual-level records (or records with fine spatial resolutions). For instance, a geo-located Twitter dataset contains tweets as individual records, and a taxi-trip dataset contains the pick-up and drop-off locations of each individual taxi trip. These individual records collectively may provide valuable insights into complex geographic phenomena and processes. When analyzing spatial patterns using these individual-level records, previous research often aggregates them into predefined administrative boundary or grids. For example, individual social media posts are aggregated to the collective behavior at grid cells (Liu et al 2015); individual movement records are aggregated to origin-destination matrices between regions (Yin et al. 2016). While these aggregation strategies dramatically decrease the complexity of spatial analysis problems and reduces big data to small data (Lazer et al. 2014), the MAUP is not adequately addressed (Openshaw 1984). In addition, the detailed spatial patterns within the aggregate units might be lost.

The increasing availability and popularity of such geospatial big data have significantly increased the importance of computation and algorithms in geographic research (Kwan 2016). The use of computational processes and associated algorithms are essential to generate results and in many situations, the results are highly mediated by algorithms and computational configurations (Lazer et al. 2014). Thus, optimal algorithms and computational approaches with high efficiency and effectiveness are critical for spatial pattern discoveries from geospatial big data. CyberGIS, a new generation of GIS based on advanced computing and cyberinfrastructure

approaches, has emerged to enable computation- and data-intensive spatial pattern discoveries (Wang and Liu 2009; Wang 2010; Wang et al. 2013).

1.3.2 Spatial Point Pattern Analysis

Spatial point pattern analysis is an important component of spatial analysis and GIScience (Illian et al. 2008). It analyzes spatial arrangements, patterns and distributions of a set of points across space and time. Most applications assume planar (2D) space while there are also one-dimensional and three-dimensional cases (Diggle 2013). In addition, spatiotemporal point pattern analysis with one temporal dimension in addition to usual 2D or 3D space is catching increasing attention (Diggle 2013, Tango 2010).

Density estimation is one of the fundamental tasks in spatial point pattern analysis. It estimates the underlying probability density function from observed point data (events). Density estimation provides a straightforward way to estimate the spatial patterns of point events, which describes how the intensity of them varies across space (Cromley and McLafferty 2011). Among a variety of density estimation approaches, kernel density estimation (KDE) is one of the most widely used methods (Silverman 1986, Shi 2010). KDE is a non-parametric density estimation method that can be considered as a sum of 'bumps' placed at observations (Silverman 1986). KDE can also be used to estimate the relative concentration of certain point events (cases) to a background population, by calculating the density ratio between cases and population (Kelsall and Diggle 1995a, Kelsall and Diggle 1995b, Shi 2010).

Clustering analysis of spatial point data plays a key role in many applications of spatial point pattern analysis (Diggle 2013). It can be further divided into the detection of clustering and the detection of clusters (Cromley and McLafferty 2011). The former type detects whether there is a tendency for point events to occur closely together (Diggle 2013, Cromley and McLafferty

2011). Classic methods to test spatial clustering include the Ripley's K and L functions (Ripley 1976) and spherical contact distribution function (Stoyan 1995, Baddeley 2004). The later type detects areas with local excesses of point events (Kulldorff 1999). Spatial scan statistics (Kulldorff 1997, Kulldorff 1999) are the most popular method for detecting spatial clusters. Density-based clustering methods from data mining literature, such as DBSCAN (Ester et al. 1996), OPTICS (Ankerst et al. 1999) and DENCLUE (Hinneburg and Keim 1998) are widely applied. Clustering and clusters may also be a result of point density fluctuation. There is a fundamental ambiguity between heterogeneity and clustering, the first corresponding to the spatial variation of the intensity function, the second to the stochastic dependence amongst the points of the process (Illian et al. 2008).

1.3.3 Spatial Point Process Models

A number of spatial point process models have been proposed to describe the spatial distributions of point events, primarily in 2D space (Illian et al. 2008, Diggle 2013). These models can be expanded to adapt to multidimensional space for the purpose of our generalized spatial point processes.

1.3.3.1 Complete spatial randomness

Complete spatial randomness (CSR) is one of the most important concepts in spatial point pattern analysis. It describes that point events within a given study area are distributed completely randomly. Extended from Diggle (2013), CSR in a multidimensional space has the following two characteristics:

- The number of events in any region (hyper-region) A with area |A| follows a Poisson distribution with mean $\lambda|A|$ (Equation (1.1)), where λ is the average of event density.

- Given *n* events in a region A, they are an independent random sample from the uniform distribution on A.

$$P(n_A = k) = \frac{(\lambda|A|)^k e^{-\lambda|A|}}{k!} \quad (1.1)$$

Based on these two characteristics, the intensity of events is constant across space and events do not have any interaction with each other.

CSR is also called homogeneous Poisson point process, and usually serves as a benchmark process to test whether points are clustered or dispersed. On one hand, if the inter-event distances are smaller than that of CSR, the point distribution is clustered. On the other hand, if the inter-event distances are larger than that of CSR, the point distribution is dispersed (uniform).

1.3.3.2 Poisson point process model

A Poisson point process model indicates that the number of events in any region *A* follows a Poisson distribution with mean according to a local underlying intensity λ_A (Equation (1.2)). CSR is a special case of Poisson point process (homogeneous Poisson process), where the event density is constant and the local intensity is proportional to the area size. In a more general case (i.e. a heterogeneous Poisson process), such density is varying across space and can be used to represent a changing population density from which events are generated.

$$P(n_A = k) = \frac{(\lambda_A)^k e^{-\lambda_A}}{k!} \quad (1.2)$$

A Poisson process model is used to evaluate whether the spatial distribution of generalized spatial points is different from a known underlying intensity that generates these

points. Spatial clusters are then detected against the null-hypothesis that the expected value of the Poison distribution is proportional to the known underlying intensity.

1.3.3.3 Bernoulli point process model

A Bernoulli point process model is used to compare distributions of two types of generalized spatial points. For the convenience of reference, the two types are called type α and type β in the remaining parts of this dissertation. A Bernoulli model states that each randomly chosen event, regardless of its location, has a constant possibility p (following a Bernoulli distribution) to be of type α. It infers that the two type events have exactly the same spatial distributions. Given a region A of n total events, a Bernoulli model states that the number of type α events follows a Binomial distribution of Equation (1.3):

$$P(n_\alpha = k) = \frac{n!p^k(1-p)^{(n-k)}}{k!(n-k)!} \tag{1.3}$$

A Bernoulli model is widely used in case-control studies. It can be used to detect clusters in bivariate marked spatial point processes, by testing against a constant possibility (random-labeling) null-hypothesis. Clusters representing local concentrations of one event type relative to the other type can be detected to represent spatial pattern differences between the two types of events.

1.3.4 Spatial Scan Statistics

Scan statistics was originally proposed to detect clusters in a point process in one-dimensional (Naus 1965b) or two-dimensional space (Naus 1965a). It was further extended into a spatial scan statistic by Kulldorff (1997), which allows the area of scanning windows to vary and can detect clusters in spatiotemporal point processes. Spatial scan statistics have then been

applied to a wide range of research domains, including but not limited to epidemiology, public health, ecology, crime analysis, and astronomy, to find clusters of events (Kulldorff 2015).

The original and most popular baseline process models in spatial scan statistics are (homogeneous or inhomogeneous) Poisson and Bernoulli (Kulldorff 1997). A Poisson model deals with the number of events occurring within a time interval and a spatial region. A Bernoulli model handles events that are in either one of two states (i.e., belonging to either of two categories), which is often used to compare the spatial distributions of two types of events, such as in a case-control study. Other models include space-time permutation (Kulldorff et al. 2005), ordinal (Jung et al. 2007), exponential (Huang et al. 2007), normal (Kulldorff et al. 2009) and multinomial (Jung et al. 2010).

The most widely used software program of spatial scan statistics is SaTScan™ developed by Kulldorff (2015) (the user guide listed hundreds of research papers in a number of domains using the software). However, this software only supports analysis in geographic space (with two spatial dimensions and optionally one temporal dimension). Furthermore, SaTScan™ is designed based on desktop computing, and thus is not scalable to large datasets. Finally, despite being free to use, SaTScan™ is not open source (Baker and Valleron 2014). With unclear implementation details, it is difficult to adapt or modify its method for serving diverse needs.

1.4 GENERALIZED SPATIAL POINT MODEL

This research deals with geographical phenomena involving one or multiple locations that can be modeled as a set of multidimensional spatial points $P = \{P_1, P_2, \ldots, P_N\}$. Each of these points represents one observation or event of the phenomena, such as a disease case or a flow trip. It contains two or more spatial dimensions, and may also include one or more temporal dimensions as well as some additional (non-spatiotemporal) attribute values, $P_i =<$

$x_{1i}, x_{2i}, \ldots t_{1i}, \ldots I_{1i}, \ldots >$ where x_{ji} is the *j*th spatial dimension, t_{ji} is the *j*th temporal dimension and $I_{ji}$is the *j*th attribute value. While it seems counterintuitive to have more than one temporal dimensions in a spatial data model, it is necessary if multiple components involved in the phenomena do not happen simultaneously. For example, suppose an analysis intends to detect spatiotemporal patterns of traffic and to model each origin-destination record as a spatial point, such a record will have two temporal dimensions, one for the trip start time and the other for the trip end time.

As an example of social media analytics, each geo-located social media post can be modeled as a point $P_i =< x_{1i}, x_{2i}, t_{1i}, I_{1i} >$. x_{1i} and x_{2i} are the 2D spatial coordinates of a social media post, $t_{1i}$is the timestamp at which the post is generated, and I_{1i} may be used to indicate either the hashtag of the post, or whether the post is talking about an event (e.g., a disease outbreak and natural disaster). Considering OD movement pattern analysis, each OD movement flow can be modeled as a point with four spatial dimensions $P_i =< x_{oi}, y_{oi}, x_{di}, y_{di} >$. $< x_{oi}, y_{oi}, x_{di}, y_{di} >$ is a point in a 4D hyperspace $X_o \times Y_o \times X_d \times Y_d$, which results from the Cartesian product of the origin's 2D geographic space $X_o \times Y_o$ and the destination's 2D geographic space $X_d \times Y_d$. When comparing the spatial distributions of two OD movement datasets, a two-valued indicator variable I_{1i}, which indicates the dataset that each movement record belongs to, is added such that $P_i =< x_{oi}, y_{oi}, x_{di}, y_{di}, I_{1i} >$. This 4D movement data model matches with the common sense of OD movement analysis – two movement flows are near to each other only if they have both near origins and near destinations, and thus all their four spatial dimensions should have similar values, which means their 4D distance is small. If two movement flows have the same origin but their destinations are far away from each other, they

are not considered as near to each other, and will be modeled as two far apart 4D points whose 4D distance is their destination's distance.

A clarification needs to be made for the usage of "events" and "points". In point pattern analysis literature, it is a convention to use events or point events to represent input data or observations. Such an event has a different meaning from "event detection" or "event-related social media points". However, when multidimensional point models or related computation are addressed, we usually call them points, since a point is straightforward to understand with geometric meaning.

1.4.1 The Necessity of the 4D Point Models in Movement Analysis

Using movement analysis as an example, this subsection explains why a multidimensional point model is necessary for spatial pattern analysis of geographical phenomena involving multiple locations. On the one hand, with a 4D point model, a pair of origin and destination in one flow is naturally considered as a single analytical unit, and the spatial interactions between the origin and the destination are modeled with the 4D point locations. Such OD integration is essential in movement analysis since the pairwise connections between origin and destination are what define movement flows. The spatial patterns of two OD movement datasets can be different even if they have exactly the same origin and destination distributions. The information about the spatial distributions of origins and destinations, and pairwise connections between them are all captured in the 4D point models, and thus researchers no longer need to analyze them separately. On the other hand, with a 4D point model, movement patterns can be represented in 4D point patterns, and analyzed through 4D point pattern analysis. Each flow represents a spatial interaction between two locations. Movement pattern analysis (e.g., flow clustering) often requires modeling the relationships a pair of flows, and thus requires

describing the spatial relationships among four 2D locations (two for each flow). It is challenging for the lack of methods to analyze such four-order relationships. The 4D point model converts movement pattern analysis from the four-order relationship between 2D locations to the spatial relationship between 4D points. Hence methods designed for analyzing 2D point patterns can be extended to analyze movement flows. For instance, Figure 1.1 shows two movement datasets. While these two movement patterns appear to be different, describing and quantifying such differences are difficult. Through our 4D model, the task to compare these two movement patterns can be converted to the comparison of the spatial patterns of two 4D point datasets. It is then possible to use a 4D bivariate point process model, and methods such as cross K-function or scan statistics for pattern comparison.

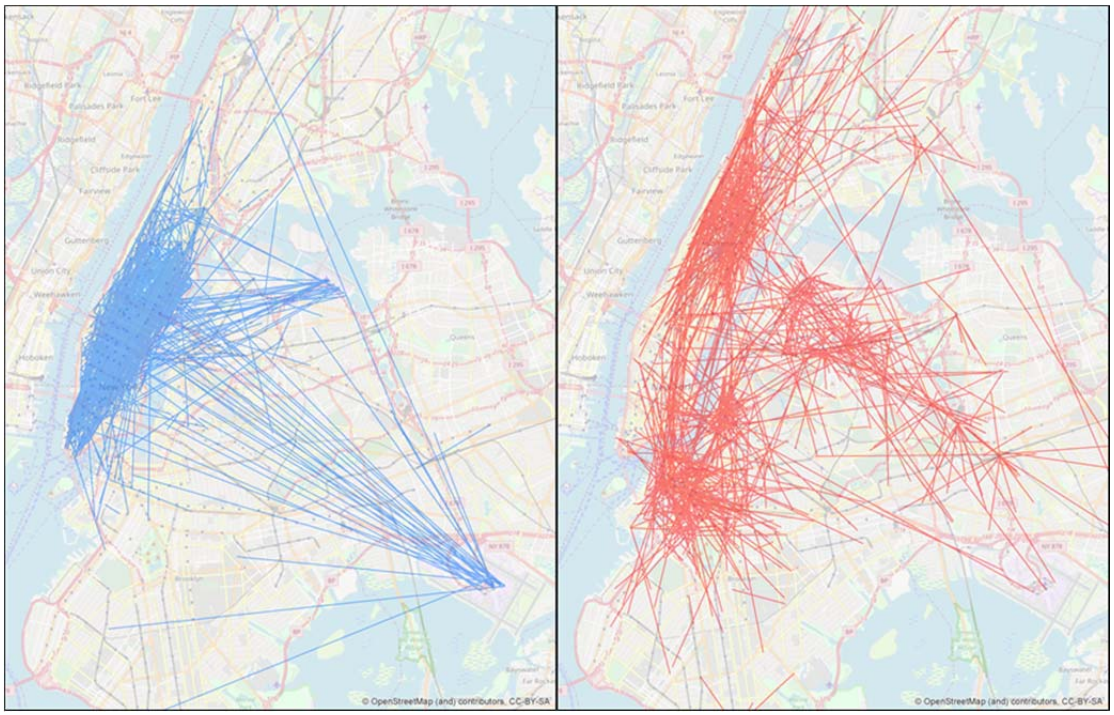

Figure 1.1: An example of two different movement patterns.

1.4.2 Comparison with Flow Clustering Literature

In movement analysis literature, it is a common practice to combine origin coordinates and destination coordinates together to calculate the distance or the similarity between movement flows for pattern discovery and clustering analysis (Berglund and Karlstrom 1999, Lu and Thrill 2008, Liu et al. 2015, Tao and Thrill 2016). Conventional point analysis approaches such as Getis-Ord's *G* (Berglund and Karlstrom 1999), Moran's *I* (Liu et al. 2015) and Ripley's *K* (Tao and Thrill 2016) are applied to movement datasets for movement cluster detection. However, without a 4D point data model, there are two major issues related to these studies. First, it is difficult to justify and extend the usage of point pattern analysis approaches for non-point datasets. This is because these methods are designed and tested based on point process models (such as CSR) and there is no guarantee that these assumptions are still valid when analyzing OD pairs. Hence, a point pattern analysis method applied to OD point pairs is a fundamentally different method, and needs its own justification. It is also generally believed that flow data is different from single point data, and that methods designed for points cannot be directly applied to flow data (Tao and Thrill 2016). Yet, in reality, point analysis methods are always applied to flow data without sufficient justification. For instance, it is not clear why the point distance in *K*-function can be replaced by a movement similarity measure (Tao and Thrill 2016), and what are the properties of this new but actually very different *K*-function. Second, without the help of a 4D space and its geometry, it is difficult to evaluate the complicated spatial relationships between movement flows in a straightforward fashion. Sophisticated categorizations are usually necessary in order to describe the spatial associations between individual movement flows (e.g., Lu and Thrill 2008). It is even more difficult to evaluate the spatial relationships between collections of flows, such as whether two flow clusters are overlapping with each other.

The 4D point data model hence can benefit movement pattern analysis research in two major ways. First, the 4D point data model builds a straightforward connection between flows and simple point data, and makes the extension of 2D point pattern analysis to movement analysis much easier both conceptually and practically. Through this data model, spatial movements can thus be understood simply as spatial points with more spatial dimensions. Hence, the concepts, point process models, pattern indicators and analytical procedures from decades of point pattern analysis research can be naturally extended to study movement patterns, just as many of the current 2D spatial pattern analysis approaches (such as KDE, scan statistics) are extended from 1D. Second, with the help of 4D geometry, it is much easier to understand and analyze the spatial patterns and relationships of flows. In 4D space, the similarity between two movement flows can be directly assessed by their 4D point distance. A 4D region can be used to represent a collection of OD movements from some origin area to some destination area. When detecting movement clusters, each cluster corresponds to one 4D region, which can be used to describe the location, size, and shape of the cluster. With the support of 4D geometry, whether a specific flow is in a cluster can be evaluated through a 4D point in region check, and the similarity between two movement clusters can be measured by computing the intersected areas of their 4D regions.

1.5 THESIS ORGANIZATION

This dissertation consists of three papers that are centered around the development of new conceptual models, methods, and algorithms for spatial pattern detection from geospatial big data, and the evaluation of them in real-world applications.

Chapter 2 describes a systematic approach to detecting spatiotemporal patterns of events from massive geo-located social media data. The approach resolves the challenges of estimating

potential time spans and influenced areas of an event from social media through several interrelated strategies: modeling social media points as space-time points; using kernel density estimation for smoothed social media intensity surfaces; utilizing event-unrelated social media posts to support mapping of relative event prevalence; and normalizing event indicators based on historical fluctuation. It is applied to detect influenza activity patterns in the conterminous US using Twitter data. Experiment results demonstrate that fine-scale influenza activity patterns consistent with available ground truth data can be detected using this approach.

Chapter 3 presents a multidimensional spatial scan statistics approach to comparing spatial movement patterns based on origin-destination (OD) representation. This approach evaluates differences and similarities between the spatial distributions of a pair of OD movement datasets and detects areas where the two spatial distributions differ the most. It is based on a multidimensional spatial point model for OD movement, where each OD record is modeled as a single point in the multidimensional OD space with four spatial dimensions – two dimensions for origins and two for destinations. A multidimensional scan statistics approach is hence developed to compare movement patterns by analyzing the multidimensional point patterns. The effectiveness of this approach is demonstrated in case studies with both individual-level and aggregated movement datasets.

Chapter 4 proposes and evaluates multiple scanning window designs for multidimensional scan statistics and provides efficient algorithms and parallel computing approaches for them. From shape, location, and size perspectives, this chapter reviews commonly used approaches in 2D spatial scan statistics and 3D space-time scan statistics, analyzes their advantages and disadvantages, and proposes 4D extensions that are valid for movement pattern analysis. Six scanning window designs are then proposed in this chapter based

on the analysis in these three perspectives. These scanning windows designs are evaluated and compared both analytically and using large real-world OD datasets. The results provide insights into the choice of scanning window designs and the trade-off between computing cost and the ability to detect high-quality clusters for movement pattern analysis.

Finally, chapter 5 concludes the dissertation, summarizes the major findings, and discusses future research directions.

CHAPTER 2: MAPPING SPATIOTEMPORAL PATTERNS OF EVENTS USING SOCIAL MEDIA: A CASE STUDY OF INFLUENZA TRENDS[1]

Abstract. *Tracking spatial and temporal trends of events (e.g. disease outbreaks and natural disasters) is important for situation awareness and timely response. Social media, with increasing popularity, provide an effective way to collect event-related data from massive populations and thus a significant opportunity to dynamically monitor events as they emerge and evolve. While existing research has demonstrated the value of social media as sensors in event detection, estimating potential time spans and influenced areas of an event from social media remains challenging. Challenges include the unstable volumes of available data, the spatial heterogeneity of event activities and social media data, and the data sparsity. This paper describes a systematic approach to detecting potential spatiotemporal patterns of events by resolving these challenges through several interrelated strategies: using kernel density estimation for smoothed social media intensity surfaces; utilizing event-unrelated social media posts to help map relative event prevalence; and normalizing event indicators based on historical fluctuation. This approach generates event indicator maps and significance maps explaining spatiotemporal variations of event prevalence to identify space-time regions with potentially abnormal event activities. The approach has been applied to detect influenza activity patterns in the conterminous US using Twitter data. A set of experiments demonstrated that our approach produces high-resolution influenza activity maps that could be explained by available ground truth data.*

Keywords: CyberGIS, Movement Analysis, Spatial Analysis and Modeling, Spatial Scan Statistics

[1] Reprint, with permission, from Gao *et al*., 2018. Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *International Journal of Geographical Information Science*, 32 (3), 425-449.

2.1 INTRODUCTION

Social media, such as *Facebook*, *Twitter* and *Flickr*, are widely popular (e.g. *Twitter* has 320 million monthly active users as of September 30, 2015[2]) and have dramatically simplified the way that information is generated, disseminated and exchanged (Kaplan and Haenlein 2010). By nature, these social media tend to provide timely individual-level information, and with the proliferation of location-aware mobile devices, they are increasingly capturing detailed spatial contexts. With spatiotemporal information and rich user-generated content, social media offer a distinct source of ambient geospatial information (Stefanidis *et al.* 2013), and a proxy for detecting spatial and temporal trends of events like disease outbreaks and natural disasters (McIntosh and Yuan 2005). The value of social media as sensors in event detection has been demonstrated by extensive research (e.g. Sakaki *et al.* 2010, Culotta 2010a, Lee and Sumiya 2010, Cheng and Wicks 2014). However, mapping potentially unknown spatiotemporal patterns of events based on social media remains challenging for the following reasons. First, both the total volumes and spatial distributions of available social media posts can change from time to time. Second, raw event magnitude estimated from social media have spatial heterogeneity, since both the underlying event activities and the popularity of event-related topics vary across space. Third, social media posts about an event can be sparse and thus event mapping may suffer from small number problems when posts are aggregated at fine spatial resolutions. Therefore, it is difficult to find a consistent event indicator based on social media for revealing informative spatiotemporal patterns of events that are worth following investigations at a fine resolution.

This paper describes a systematic approach to mapping potential spatiotemporal patterns of events by tackling the aforementioned challenges. Specifically, a set of interrelated strategies is employed in this approach. First, individual social media posts are modeled as space-time

[2] https://about.twitter.com/company visited on 01/10/2016

points, and a kernel density estimation (KDE) method is used to generate smoothed social media intensity surfaces. Second, both posts that reveal observations of a target event (event-related posts) and ones that do not (event-unrelated posts) are utilized in this approach (Albuquerque *et al.* 2015). Different from Albuquerque *et al.* (2015), which uses the frequency of event-related tweets in aggregated spatial blocks, this paper uses them to map the relative spatial distribution (ratio map) of event-related posts through KDE. Third, these ratio maps are tested against the null hypothesis that the value at each location is not different from the local historical values where there are no severe events. A normalization process based on local historical fluctuation is used to generate the final event indicator from the ratio maps. This normalization process tackles the spatial heterogeneity challenge by presenting how abnormal the event activity is, rather than providing an absolute event measure.

Hence, our approach includes the following major phases. First, both event-related posts and event-unrelated posts are extracted from geo-located social media datasets by analyzing textual contents of these posts. Then, for each period, spatial density maps of both event-related posts and all posts are estimated using KDE respectively; maps representing the proportion of posts that are event-related (referred to as **social media event rate**, or **SMER**) are generated as the ratio between the two density maps. Finally, a local baseline of SMER is formed for each location according to its historical trends, and SMER maps are tested against and normalized based on the local baseline to get both the final event activity maps and the *p*-value maps indicating the significance of event activities. These resulting maps reveal potential spatiotemporal patterns of the underlying event.

Our approach was evaluated in a real-world problem of detecting influenza activity patterns in the conterminous US using *Twitter* data. Influenza-related tweets were first extracted

from collected *Twitter* datasets. These posts were then used to estimate the magnitude, spread and trend of influenza activities from 2012 to 2014. Experiments showed that our approach can estimate potential spatiotemporal patterns of influenza activities with unprecedented spatiotemporal resolutions, many of which can be explained by ground truth data.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 explains the proposed approach. Section 4 describes a case study of monitoring influenza activities in the US using *Twitter* data, applying the approach from section 3. Section 5 analyses the results from the case study. Finally, section 6 concludes the paper.

2.2 RELATED WORK

Social media provide alternatives to traditional data collection approaches like questionnaires or interviews for understanding people's opinions and observations (Lampos 2012), and are increasingly used in many research domains. Analyzing correlations between social media and real-world events has become an active field of research. Extensive research has been conducted using social media data to estimate or predict various events, including influenza activities (Corley *et al.* 2009, Lampos and Cristianini 2010, Lampos and Cristianini 2012, Culotta 2010a, Culotta 2013, Signorini *et al.* 2011, Achrekar *et al.* 2011, Nagel *et al.* 2013, Aslam *et al.* 2014, Paul et al. 2014), box-office revenues (Asur and Huberman 2010), stock markets (Bollen *et al.* 2011) and election results (Tumasjan *et al.* 2010, and Tsou *et al.* 2013). Most of these studies estimate statistical relationships between event measurements and attributes retrieved from social media in predefined study areas, and use these relationships for prediction. However, the previous research has the following limitations. First, these methods can only be used to estimate event activities at aggregated regions where there are real event measurements to train models. Second, due to spatial heterogeneity, an estimation model may

not be generalizable to regions other than those where training is conducted. Consequently, it is challenging to find a consistent event indicator for revealing spatiotemporal patterns of events at a fine resolution using these methods.

Mining event patterns from social media requires finding the subset of posts that are relevant to target events. Methods have been developed to extract event-related social media posts. Culotta (2010a, 2010b) estimated a logistic regression model from a small set of labeled messages to tell whether a tweet was reporting a flu symptom. Sakaki *et al.* (2010) used a support vector machine (SVM) method to find tweets that report actual earthquake occurrences. Aramaki *et al.* (2011) compared different classification methods for finding influenza-related tweets and found that an SVM classifier outperformed other methods. Their classification method is adapted in this paper to find social media posts related to particular events. A similar method was also used by Sadilek *et al.* (2012). Doan *et al.* (2012) and Lamb *et al.* (2013) went further in analyzing semantic features in tweets to find flu-related tweet contents. However, their method requires in-depth domain knowledge, and thus cannot be easily generalized to other events.

Spatial event (anomaly) detection with social media data represents another related research area. In this area, unusual bursts of social media usage are detected to indicate possible real-world events. It can be either an open domain event detection where abnormal behaviors of social media users (e.g. clusters of posts, active movements of social media users) are considered as signals of events (Lee and Sumiya 2010, Cheng and Wicks 2014), or targeting a particular kind of events (Earle et al. 2010, Sakaki et al. 2010). While open-domain methods require limited prior knowledge about events, it is consequently difficult to make detection results informative. Further analyses are often needed to find what each anomaly or cluster means, and

there is no guarantee that events of interest can be identified. Finally, event (anomaly) detection research usually only answers a yes-or-no question – whether an event occurs at certain locations. In contrast, our approach focuses on detecting how the magnitude of particular events changes spatially and temporally.

A number of studies used KDE for exploratory analysis of geo-located social media (Pozdnoukhov and Kaiser 2011, Chae *et al.* 2012, Hwang *et al.* 2013, Li *et al.* 2013). These studies addressed a variety of application problems such as user location recommendation or prediction (Kurashima et al. 2010, Zhang and Chow 2013, Lichman and Smyth 2014, Thom *et al.* 2014), human activity pattern analysis (Hasan *et al.* 2013) and understanding spatial relationships between places (Li and Goodchild 2012). Different from previous research, our approach exploits the analytical capabilities of KDE through the calculation of density ratios followed by a normalization procedure based on historical trends, in response to the challenges of event mapping based on social media.

Other related research includes data searching, visualization and analysis tools for event situation awareness based on social media (MacEachren *et al.* 2011, Tsou *et al.* 2015), demonstrations of the value of twitters as sensors (Crooks *et al.* 2013), statistical models for analyzing spatiotemporal trends of social media data (Helwig *et al.* 2015) and cyberGIS for processing and analyzing massive social media data (Wang 2010, Wang *et al.* 2013, Hwang *et al.* 2013, Padmanabhan *et al.* 2014, Cao *et al.* 2015). Different from these studies, this research aims to establish a holistic suite of methods for mapping spatiotemporal patterns of events by combining social media data with spatiotemporal analysis.

2.3 METHODS

2.3.1 Overview

Our approach is applicable to events that have the following three key characteristics: (1) they spread across space and time, (2) they can be observed by social media users and (3) these observations can be embedded in social media posts. Disease epidemics, natural disasters and festival celebrations are examples of these events. For example, influenza epidemics meet these requirements since (1) they spread spatially and span for a period, (2) flu symptoms such as coughing and fever are easily observable and (3) these flu observations exist in social media posts.

The overarching workflow (Figure 2.1) includes the following three phases. The first phase extracts event-related posts using a classifier that combines keyword filtering and supervised classification methods, which is described in Section 3.2. The second phase estimates the spatial densities of both event-related posts and all posts, and then generates SMER maps, which is described in Section 3.3. Finally, the third phase is described in Section 3.4. In this phase, local event baseline is first estimated by using a sequence of historical SMER maps. Based on the baseline, both normalized event activity maps and *p*-value maps are generated to quantify the severity and significance of event activities.

2.3.2 Extraction of Event-Related Social Media Posts

This section describes how to extract event-related posts from social media datasets. Different from previous research (e.g. Aramaki et al., Sakaki et al. 2010, Signorini et al. 2011) that only considers event-related posts, our study requires event-unrelated posts to serve as controlling points to map the relative concentration of event-related posts, similar to Albuquerque et al. (2015). While social media datasets are usually huge in size, only a small

proportion of posts are related to any particular event. These related posts, besides event observations, also include general discussions, comments and even stories about the event. However, these posts do not necessarily reflect the current event situation and thus should not be considered as event-related posts. Hence, in this paper, an event-related post needs to be about a recent actual event observation. This constraint also implies that an event-related post is usually close to the observed phenomenon in space and time.

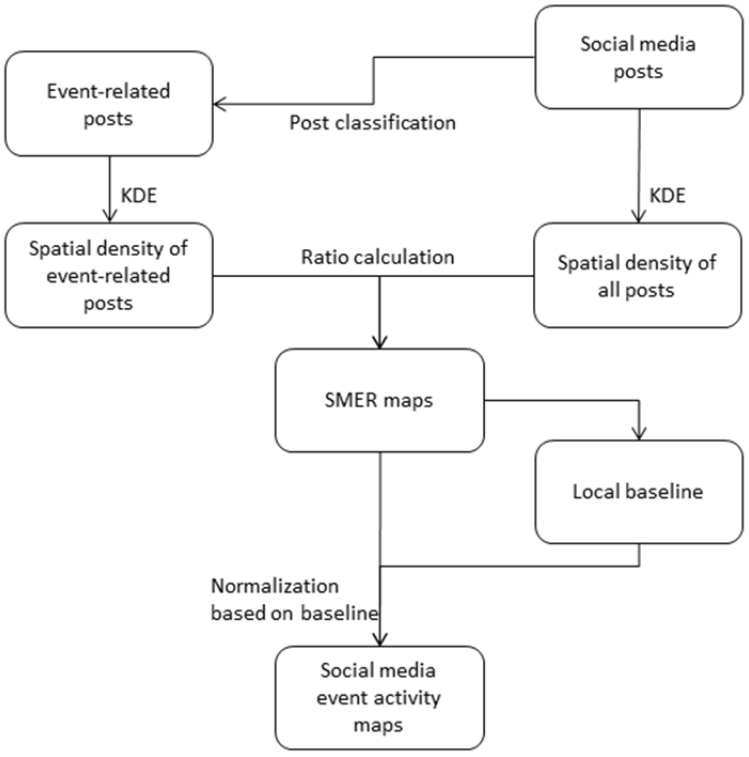

Figure 2.1: Overarching workflow.

In this paper, extracting event-related posts is formulated into a two-class classification problem to identify each post as either event-related or event-unrelated based on the textual content of the posts. The classification problem can be tackled by representing each social media post using a bag-of-words model and applying machine learning methods such as SVM classifier (Hsu *et al.* 2003) to these feature vectors. However, this classification problem is often highly

imbalanced, because event-unrelated posts can usually be hundreds or thousands of times more than event-related ones. The imbalance does not only limit the performance of classification methods (Tang *et al.* 2009), but also increases the difficulty in preparing training datasets for model training. A randomly chosen training dataset with thousands of posts is expected to contain only a limited number of event-related posts, which can be highly unrepresentative.

To resolve such imbalance, a keyword filtering is applied to social media posts before supervised classification is used. Keyword filtering is based on the observation that most social media posts describing an event contain at least one word from a small set of event-related keywords. To do keyword filtering, a keyword list needs to be prepared for a target event. Such a keyword list needs to include the most commonly used words to describe the event. It also needs to have as fewer keywords as possible to reduce potential noises. Posts without any of these keywords are marked as event-unrelated directly. The remaining ones are further classified using the supervised classification methods. The supervised classification methods are trained from a training dataset of posts that are already labeled as event-related or not. This labeling process controls that only posts about recent actual event observations are marked as event-related. Mentioning terms like ‘news’, ‘last year’ or ‘everywhere’ usually makes a post event-unrelated.

2.3.3 Estimation of SMER Maps

SMER maps are computed as the spatial density of event-related posts (cases) divided by the spatial density of all posts (population), and KDE is used to estimate both spatial densities (Bithell 1990 and 1991). KDE is a non-parametric density estimation method that can be considered as a sum of ‘bumps’ placed at observations (Silverman 1986). It is widely used in many research domains to explore spatial patterns of point events and can handle inhomogeneous background populations (Shi 2010). This paper utilizes the same kernel

bandwidth for both cases and population, which is justified by Kelsall and Diggle (1995a and 1995b) and Shi (2010). When the same bandwidth is used, Equation (2.1) (Silverman 1986, Bithell 1990 and 1991) is used, where $r(x, y)$ is the estimated SMER value at location (x, y), $K()$ is the kernel function, h is the bandwidth, (X_i, Y_i) is the location of ith social media post and z_i indicates whether it is event-related ($z_i = 1$) or not ($z_i = 0$). This paper uses an Epanechnikov kernel function (Silverman 1986).

$$r(x, y) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}, \frac{y - Y_i}{h}\right) z_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}, \frac{y - Y_i}{h}\right)} \quad (2.1)$$

The use of SMER instead of count or density of event-related posts is due to the instability of daily social media post volumes. On the one hand, the number of social media posts depends on the popularity of specific social media services and may change from time to time. On the other hand, both people's posting behaviors (e.g. differences between weekdays and weekends) and uncontrollable technical issues (e.g. data volume controlled by data providers) influence the number of available posts per day. Consequently, SMER provides a more robust event activity measure that can be compared across time.

Multiple reasons contribute to the choice of KDE as the density estimation method in this paper. One reason is that non-parametric density estimation methods are preferred to model complex spatial patterns such as event activities. The distribution of social media posts (either raw or event-related) across an entire study area cannot be assumed to follow density functions of particular forms like Gaussian or to have linear or quadratic spatial trends. Another reason is that the sparsity of event-related posts makes methods based on binning ineffective (Helwig *et al.* 2015). For any small area like a raster cell on the output map, the number of social media posts can be very small. This problem makes the estimation of each bin highly unreliable.

Consequently, KDE, which incorporates neighboring points through a distance decay kernel function, is argued to be the most proper method for this research.

2.3.4 Mapping Event Activities

Mapping event activities from social media data requires the quantification of how the local prevalence of event-related posts (i.e. SMER) in each location deviates from its normal situation. Thus, the SMER at each location is tested against the null hypothesis that it is not different from its normal values where there are no severe events. Such a null hypothesis is different from existing research using KDE (e.g. Kelsall and Diggle 1995a, Lloyd and Cheshire 2017), where the spatial variations of risks (i.e. the ratio of one density to another) are tested against a constant risk null hypothesis. In this research, the SMER at each location is compared with and tested against the empirical distribution of the SMERs during a long historical baseline period with no severe events.

There are several reasons why SMER maps are not directly used as event activity indicators and why a constant risk is not a proper null hypothesis. First, event magnitude has different baselines across space. A common event (e.g. 3 in of snow) in one place can be extraordinarily and highly influencing somewhere else. Second, the popularity of event-related topics varies across space. How likely a social media user posts about event-related observations can be different in different regions. Third, due to the small number problem, the ratio in less populous areas can fluctuate greatly and may easily stand out as extreme event activities due to random chances (Cromley and McLafferty 2011, p.153-157).

In this paper, two measures are used to quantify how event activities deviate from their normal (baseline) situations at each individual location. The first measure is a normalized event activity level, which is defined as the *z*-score of SMER - the number of standard deviations the

SMER is above the mean. A higher event activity level indicates that the SMER is higher above its normal value range when there are no severe event activities. The formula of the normalized event activity level $l(x, y)$ is expressed by Equation (2.2), where $r(x, y)$ is the SMER at (x, y), $mean(x, y)$ is the baseline mean and $sd(x, y)$ is the baseline standard deviation. The necessity of such normalization is also demonstrated in the case study in Section 4.4.

$$l(x, y) = \frac{r(x,y) - mean(x,y)}{sd(x,y)} \tag{2.2}$$

The second measure is the *p*-value indicating the probability to encounter a higher or equal SMER value when a SMER is hypothesized to be from its normal situation without severe events. Once a baseline period is chosen and the SMER values (e.g. weekly, daily or even hourly) during the baseline period are calculated, the *p*-value at (x, y), $p(x, y)$, can be calculated by Equation (2.3), where $N_{Baseline}(x, y)$ is the number of baseline SMER values, and $N_{Higher}(x, y)$ is the number of baseline SMER values that are higher or equal to the target SMER value. The *p*-value ranges between $1/(N_{Baseline}(x, y) + 1)$ and 1, with the lowest value indicating that the target SMER value is the highest among all baseline values.

$$p(x, y) = \frac{N_{Higher}(x,y)+1}{N_{Baseline}(x,y)+1} \tag{2.3}$$

The two measures are complementary to each other. While the *p*-value provides the statistical significance to judge whether a SMER is different from its normal situation, it does not inform how severe the event activity is. Two SMERs can have the same *p*-values if one is slightly above the highest baseline value and the other is five times higher. On the other hand, the normalized event activity level provides a straightforward indicator of how abnormally high the SMER is, but lacks the power to represent pattern significance as the *p*-value does.

2.4 CASE STUDY: DETECTING INFLUENZA PATTERNS USING *TWITTER* DATA

The methods described in the previous section were illustrated and evaluated in a real-world problem of detecting spatiotemporal patterns of influenza activities in the conterminous US using data collected from Twitter. Seasonal influenza epidemics cause respiratory illnesses and deaths worldwide each year (CDC 2014). Traditional influenza surveillance systems, such as the ones used by the US Centers for Disease Control and Prevention (CDC) or the European Influenza Surveillance Scheme are based on virological and clinical data, which typically take 1-2 weeks to generate reports only at large aggregated regions (Ginsberg et al. 2009). Data from social media services like Twitter might be used to supplement traditional surveillance systems by providing near real-time estimation of influenza activities with fine-resolution spatiotemporal details.

2.4.1 Data

Geo-located tweets were collected continuously since 11/01/2012 using the *Twitter* streaming API, with a bounding box covering the conterminous US. The data spanned two influenza seasons: 2012-2013 and 2013-2014. After retweets, tweets outside the bounding box and non-English language tweets were eliminated; on average over 2 million tweets were collected every day (Figure 2.2). The tweet volume is not constant with an overall increasing trend and several noticeable drops.

Weekly influenza surveillance reports in the US provided by CDC[3] were considered as ground truth in this paper. These reports provide the proportion of patients visiting health care providers for influenza-like illness (ILI) from the US Outpatient Influenza-like Illness Surveillance Network (ILINet) at the US national level and 10 Health and Human Services

[3] Data source: http://www.cdc.gov/flu/weekly/fluviewinteractive.htm

(HHS) regions each containing several states. The proportion is referred as CDC ILI value in the remaining parts of this paper. The CDC's reports also include the influenza activity levels in each state. The influenza activity level is defined as the number of standard deviations that a CDC ILI value is below or above the mean value in the last non-influenza season in the same state (CDC 2015).

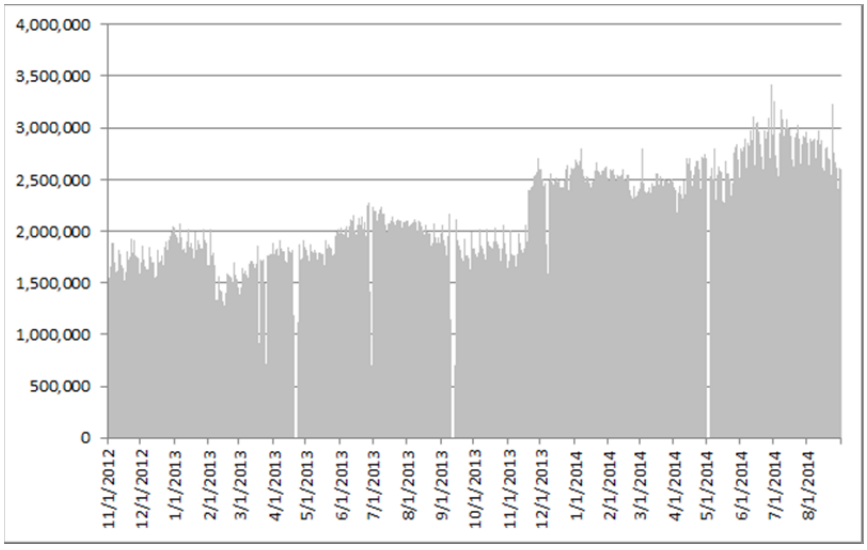

Figure 2.2: Number of tweets per day after filtering.

2.4.2 Tweet Classification

This paper used the following keywords: 'FLU', 'INFLUENZA', 'FEVER', 'COUGH', 'SORE THROAT', 'RUNNY NOSE', 'STUFFY NOSE' and 'COLD' to identify potential flu-related contents based on the influenza symptoms described by CDC (2015) and previous research (Culotta 2010b, Signorini *et al.* 2011, Lampos and Cristianini 2012). The inclusion of 'COLD' was because flu and cold share many symptoms and that it is very difficult for common people to tell their differences (CDC 2011). Tweets without any of these keywords were marked as influenza-unrelated and filtered out directly.

An SVM classifier was applied to the remaining tweets to decide whether they were influenza-related. An influenza-related tweet reports or implies an observation of flu cases or symptoms. The SVM classifier was trained using a manually labeled training dataset of 6502 tweets. These tweets were chosen randomly from the tweets that passed the keyword filtering in the 2012-2013 influenza season. A total of 3542 tweets were labeled as event-related and 2960 as event-unrelated. Examples of event-related tweets include 'I have the flu :(' and 'Now I have a stuffy nose' while examples of event-unrelated tweets include 'Going to get a flu shot this year... is it worth it?' and 'This person has a serious case of the Bieber fever'. Tweets were represented using a bag-of-words model; the feature vectors for the SVM model were binary indicators of whether words exist. A polynomial kernel with degree 3 was used in the SVM model. A grid search was performed during the training process to find the best model parameters based on a fivefold cross-validation (Hsu *et al.* 2003).

Finding the best tweet classification method or improving the accuracy of specific methods is beyond the scope of this paper. However, having a tweet classifier of adequate quality is important for spatiotemporal pattern analyses. To evaluate the performance of the SVM classifier, a testing dataset of 1000 tweets independent of the training dataset was used and manually labeled similarly as the training dataset. The trained SVM classifier was applied to classify these 1000 tweets. The F-score is 0.814, showing that our application of the method by Aramaki et al. (2011) is of adequate quality as their best F-sore is 0.76 on Japanese tweets.

2.4.3 Evaluating Regression Analysis Methods to Detect Spatial Patterns of Influenza Activities

Before applying the proposed method, we first evaluated the performance of regression analysis methods for detecting spatial patterns of influenza activities. Regression analysis

requires the existence of some ground truth influenza activity measures (like CDC ILI values) to train and evaluate models. However, these influenza activity measures are not available except in a handful of regions where there are official reports. Thus, if regression analysis methods are to be used to estimate spatial patterns of influenza activities, some generic relationships between social media responses and influenza activity measures need to exist, regardless of spatial scales or areas. This requirement is necessary since mapping the spatial extents of influenza outbreaks requires an estimation of influenza activities at any location, not merely in the handful of regions with reports.

In order to evaluate whether generic relationships between social media and influenza activity measures exist, we performed cross-validations between regions. For each of the 11 regions (the entire US and 10 HHS regions), regression models were trained to estimate the CDC ILI values based on social media. Then each of the models was tested in all 11 regions, and their prediction mean squared errors (MSE) were recorded and compared.

Three models were trained for each region. First, a linear regression model used the SMER of flu (the proportion of tweets that are influenza-related) in each region to predict the CDC ILI values. Second, an autoregressive with exogenous input (ARX) model (Gilbert 1995) was trained to model how the CDC ILI values could be estimated from its past values with the input of SMER. ARX models have been demonstrated to predict influenza measures more accurately than using only historical data or social media (Achrekar *et al.* 2011, Paul *et al.* 2014). Using this model, the most recent CDC ILI value was estimated using both lagged (previous) flu reports and social media. It helps to model the complex relationships between the time series of CDC ILI values and social media measures, since flu prevalence and social media responses in previous periods may influence how likely people post about flu. Third, for comparison, an

autoregressive (AR) model was trained to estimate CDC ILI values purely based only on its past values. Social media were not used in this model. For both the AR and ARX models, the best number of lags (the number of previous values used to predict the current value) was determined by Akaike Information Criterion. It is worth noting that neither AR nor ARX model can be directly used to estimate the influenza activities in regions where no official report is available, since they rely on past reported values for prediction.

2.4.4 Detecting Spatiotemporal Patterns of Influenza Activities

The methods described in Section 3.3 and 3.4 were used to estimate influenza activity maps. SMER maps were first estimated for each week using the KDE method. The KDE used the Epanechnikov kernel function and a bandwidth of 200km. Then, the non-influenza season in 2013 (from 03/13/2013 to 11/26/2013, for a total of 37 weeks) was treated as the baseline period, and the baseline mean and baseline standard deviation of SMERs in each map cell were calculated. Finally, the SMER values were used to generate both influenza activity maps and *p*-value maps based on local baselines. These maps were then used to interpret spatiotemporal patterns of influenza activities.

Although a fixed bandwidth KDE is usually believed to under smooth in areas with few observations and over smooth in areas with more observations compared with an adaptive bandwidth KDE (Davies and Hazelton 2010), it is used in this research for the following reasons. First, the event activity indicator in this research is based on the comparison with local historical values, and thus using the same bandwidth per location ensures the SMERs in different time periods are estimated from observations in the same spatial extent. Furthermore, both the daily available social media volumes and their spatial distributions are not constant given a long time period, and thus calculating one bandwidth for each cell and using it extensively over time is

problematic. Second, based on our experiments, adaptive KDE has a tendency to predict the risk in less populous areas exactly the same as their nearest large cities. This is because the bandwidths in these areas need to expand to populous cities to have sufficiently smoothed results. Third, the under smoothing in areas with few observations can be mitigated by the normalization procedure, since it takes into account the large variance in these areas by estimating their SMERs' distribution during a long historical baseline period.

The bandwidth of 200km was chosen based on the following justifications. First, according to the widely used bandwidth selection method by Bailey and Gatrell (1995), the optimal bandwidths for all social media posts within a week are around 50km, and the optimal ones for event-related posts are around 350km. The 200km bandwidth lies in between the two values. Second, leaving-one-out cross-validation was conducted for bandwidths ranging from 5km to 600km. The bandwidth that minimizes the MSE was 200km for more than half of the weekly datasets being tested. Finally, the 200km bandwidth is an optimal choice because a larger bandwidth (e.g. 400km) would generate highly smoothed results and conceal spatial patterns within a state while a smaller bandwidth (e.g. 100km or 50km) would lead to patterns with noticeable sharp changes.

The necessity of using the normalization strategy to handle spatial heterogeneity was demonstrated using the data in all the 10 HHS regions. By combining the weekly SMERs and CDC ILI values in 10 regions altogether, we had 880 (88 weeks × 10 regions) pairs of values. The correlation between the SMERs and the CDC ILI values was low (0.5894). This was mainly because different regions have different value ranges for SMER and CDC ILI. After normalizing both SMERs and CDC ILI values using Equation (2.2), the correlation between the normalized

pairs was much higher (0.7463). It demonstrated that the normalization strategy produces a better measure of influenza activity that reduces the influence of spatial heterogeneity.

2.5 RESULTS

2.5.1 Tweet Classification

The numbers of extracted influenza-related tweets per day are shown in Figure 2.3. A temporal trend of influenza activity is clearly shown in this figure. The tweet number ranged from below 500 in non-influenza seasons to above 2000 in influenza seasons. The number of influenza-related tweets is low compared with the total number of tweets collected (Figure 2.2) - overall less than 0.05% (1 in 2000) tweets are influenza-related.

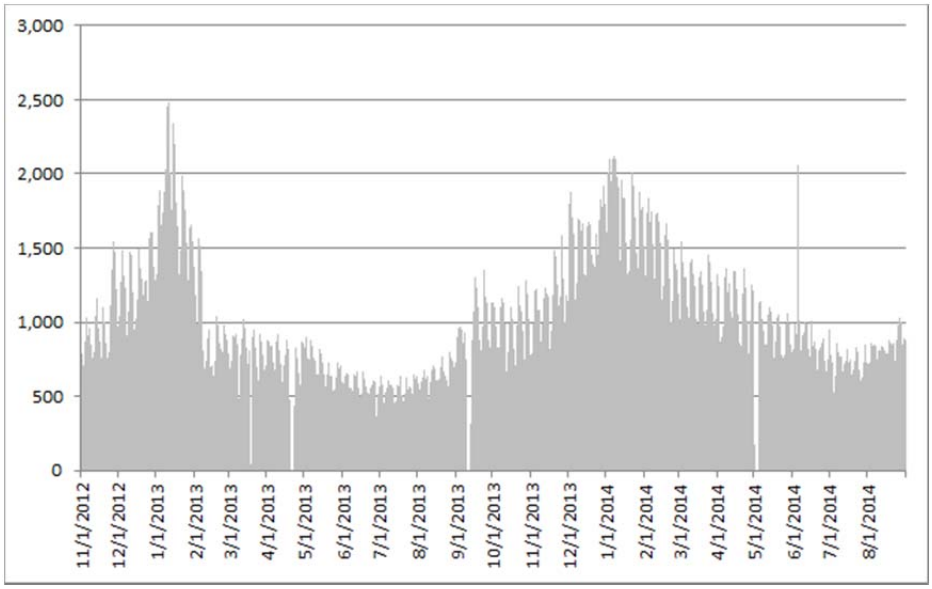

Figure 2.3: Number of influenza-related tweets per day.

2.5.2: Regression Analysis Evaluation

The cross-validation results for linear regression models, ARX models and AR models are shown in Table 2.1, 2.2 and 2.3 respectively. In these tables, Ntn represents the US national dataset, and Rgn represents the data in each HHS region; each row represents a model that is trained using the data in the region of the row header, and each column represents the MSE of models when applied to the region of the column header; the diagonal elements indicate the results when the model is trained and applied in the same region. When only the diagonal elements were considered, both ARX and AR models had better performance than linear regression models, and ARX models outperformed AR models in 9 of the 11 regions (except for Region 1 and Region 3). This result demonstrated that models incorporating data from both Twitter and historical CDC flu reports reduced prediction error, which is consistent with Paul et al. (2014).

Table 2.1: Cross-validation MSEs of linear regression models

	Ntn	Rgn1	Rgn2	Rgn3	Rgn4	Rgn5	Rgn6	Rgn7	Rgn8	Rgn9	Rgn10
Ntn	3.73E-05	2.66E-04	7.20E-05	5.55E-05	4.23E-05	7.80E-05	3.85E-04	1.41E-04	1.31E-04	5.79E-05	2.26E-04
Rgn1	1.78E-04	1.97E-05	2.47E-04	1.66E-04	1.46E-04	8.87E-05	8.93E-04	1.63E-04	5.06E-05	2.74E-04	4.68E-05
Rgn2	5.51E-05	2.84E-04	4.83E-05	7.69E-05	7.90E-05	1.18E-04	3.28E-04	2.05E-04	1.81E-04	4.51E-05	2.90E-04
Rgn3	4.01E-05	2.01E-04	7.80E-05	5.34E-05	4.45E-05	6.04E-05	4.37E-04	1.24E-04	9.79E-05	6.88E-05	1.80E-04
Rgn4	3.97E-05	3.10E-04	9.01E-05	6.00E-05	3.95E-05	8.20E-05	3.90E-04	1.37E-04	1.40E-04	6.73E-05	2.38E-04
Rgn5	7.18E-05	9.69E-05	1.35E-04	7.38E-05	6.52E-05	4.09E-05	6.07E-04	9.55E-05	4.31E-05	1.32E-04	9.12E-05
Rgn6	3.66E-04	1.56E-03	3.04E-04	4.51E-04	3.14E-04	7.12E-04	8.22E-05	8.52E-04	9.92E-04	2.85E-04	1.23E-03
Rgn7	1.05E-04	1.63E-04	2.16E-04	1.07E-04	9.38E-05	5.45E-05	6.85E-04	8.01E-05	5.38E-05	1.88E-04	1.00E-04
Rgn8	1.22E-04	5.14E-05	2.04E-04	1.15E-04	1.04E-04	5.38E-05	7.61E-04	1.03E-04	3.07E-05	2.07E-04	5.27E-05
Rgn9	5.19E-05	3.47E-04	5.12E-05	7.71E-05	6.93E-05	1.30E-04	2.95E-04	2.11E-04	2.06E-04	4.19E-05	3.23E-04
Rgn10	1.95E-04	3.05E-05	2.92E-04	1.80E-04	1.62E-04	9.22E-05	9.38E-04	1.40E-04	4.50E-05	3.05E-04	3.95E-05

The real CDC ILI values are usually below 10% (0.1). If a model predicts the real value with an error of 1% (0.01) in every week, which is of poor performance, the MSE is 1.00E-04.

Linear regression methods (Table 2.1), which only used social media to predict CDC ILI values, had the worst cross-validation results. A model trained in one region usually performed poorly in other regions. When a model trained from another region was applied to a target region, the prediction MSE could often be 5 or 10 times larger than that of the model trained

from the target region. Furthermore, some models consistently overestimated or underestimated the real CDC ILI values. For example, a model trained from region 6 on average predicted the CDC ILI values in region 1 one fourth of the real values. ARX models (Table 2.2), which combined social media and past CDC ILI values in the prediction model, had much better cross-validation results. The MSE differences between models trained from the same and different regions were much smaller. Finally, AR models (Table 2.3), which used no social media data, had the best cross-validation results. In our experiment, applying a model trained from a different region could in most cases have a very similar MSE compared with applying the model trained from the same region.

Table 2.2: Cross-validation MSEs of ARX models

	Ntn	Rgn1	Rgn2	Rgn3	Rgn4	Rgn5	Rgn6	Rgn7	Rgn8	Rgn9	Rgn10
Ntn	1.24E-05	1.38E-05	1.57E-05	2.47E-05	1.53E-05	1.31E-05	4.18E-05	2.35E-05	1.20E-05	1.56E-05	2.22E-05
Rgn1	2.01E-05	8.21E-06	2.73E-05	3.18E-05	2.06E-05	1.45E-05	7.19E-05	2.70E-05	1.09E-05	2.85E-05	1.56E-05
Rgn2	1.39E-05	2.05E-05	1.39E-05	2.58E-05	1.75E-05	1.73E-05	4.86E-05	3.03E-05	1.75E-05	1.38E-05	2.86E-05
Rgn3	2.30E-05	2.27E-05	2.87E-05	3.68E-05	2.50E-05	2.13E-05	7.17E-05	4.27E-05	1.97E-05	3.33E-05	3.36E-05
Rgn4	1.53E-05	2.74E-05	2.23E-05	2.81E-05	1.47E-05	1.78E-05	5.23E-05	2.73E-05	1.46E-05	1.86E-05	2.63E-05
Rgn5	1.30E-05	7.61E-06	1.81E-05	2.66E-05	1.57E-05	1.12E-05	4.86E-05	2.19E-05	9.21E-06	1.80E-05	1.66E-05
Rgn6	2.56E-05	6.31E-05	2.60E-05	4.18E-05	2.84E-05	3.75E-05	3.22E-05	4.98E-05	4.46E-05	2.48E-05	6.40E-05
Rgn7	1.23E-05	7.94E-06	1.76E-05	2.94E-05	1.59E-05	1.16E-05	3.77E-05	2.17E-05	8.93E-06	1.59E-05	1.61E-05
Rgn8	1.41E-05	6.84E-06	1.77E-05	2.51E-05	1.61E-05	1.26E-05	4.55E-05	2.32E-05	8.69E-06	1.85E-05	1.59E-05
Rgn9	1.38E-05	2.57E-05	1.49E-05	2.52E-05	1.71E-05	1.85E-05	4.41E-05	3.01E-05	1.78E-05	1.38E-05	3.17E-05
Rgn10	1.65E-05	6.27E-06	2.20E-05	2.86E-05	1.80E-05	1.30E-05	5.73E-05	2.34E-05	9.59E-06	2.25E-05	1.50E-05

These results showed that spatial heterogeneity exists in the statistical relationships between social media measures and CDC ILI values. The more a model relied on social media, the worse it fit into other areas. It indicated that regression models to predict influenza activities based on social media are highly localized. They only reflect the situation in a specific aggregated region and may not be generalized to any sub regions or any other regions. Hence, since official influenza reports are not available in all the areas of interest to train or evaluate models, regression analysis methods may not be effective to estimate spatiotemporal patterns of influenza activities.

Table 2.3: Cross-validation MSEs of AR models

	Ntn	Rgn1	Rgn2	Rgn3	Rgn4	Rgn5	Rgn6	Rgn7	Rgn8	Rgn9	Rgn10
Ntn	1.33E-05	7.17E-06	1.75E-05	2.85E-05	1.71E-05	1.23E-05	3.80E-05	2.27E-05	9.70E-06	1.72E-05	1.58E-05
Rgn1	1.39E-05	6.92E-06	1.79E-05	2.74E-05	1.72E-05	1.29E-05	3.98E-05	2.46E-05	9.55E-06	1.82E-05	1.59E-05
Rgn2	1.40E-05	7.50E-06	1.73E-05	2.73E-05	1.77E-05	1.34E-05	4.35E-05	2.49E-05	1.02E-05	1.74E-05	1.62E-05
Rgn3	1.80E-05	8.24E-06	2.06E-05	2.46E-05	1.92E-05	1.72E-05	5.94E-05	3.41E-05	1.18E-05	2.50E-05	1.99E-05
Rgn4	1.35E-05	7.08E-06	1.75E-05	2.63E-05	1.64E-05	1.29E-05	4.14E-05	2.38E-05	9.43E-06	1.81E-05	1.61E-05
Rgn5	1.33E-05	7.18E-06	1.78E-05	2.89E-05	1.71E-05	1.22E-05	3.74E-05	2.25E-05	9.72E-06	1.74E-05	1.58E-05
Rgn6	1.35E-05	7.32E-06	1.85E-05	3.07E-05	1.77E-05	1.22E-05	3.66E-05	2.25E-05	9.91E-06	1.80E-05	1.64E-05
Rgn7	1.36E-05	7.25E-06	1.83E-05	2.94E-05	1.72E-05	1.22E-05	3.75E-05	2.24E-05	9.82E-06	1.79E-05	1.58E-05
Rgn8	1.36E-05	6.97E-06	1.74E-05	2.60E-05	1.65E-05	1.31E-05	4.18E-05	2.47E-05	9.36E-06	1.82E-05	1.62E-05
Rgn9	1.32E-05	7.82E-06	1.67E-05	3.07E-05	1.79E-05	1.27E-05	4.06E-05	2.31E-05	9.80E-06	1.62E-05	1.57E-05
Rgn10	1.40E-05	7.37E-06	1.83E-05	2.80E-05	1.72E-05	1.27E-05	4.06E-05	2.31E-05	9.98E-06	1.81E-05	1.55E-05

2.5.3 Spatiotemporal Patterns of Influenza Activities

The empirical distribution of SMER values during the baseline period of 37 weeks varies greatly for different locations. The baseline mean and standard deviation of SMER are shown in Figure 2.4 and Figure 2.5 respectively. These figures indicated that spatial heterogeneity exists in SMER values. Generally, the proportion of influenza-related tweets in the North is larger than that in the South during the non-influenza season. For the standard deviation, it is the smallest in South-eastern parts of the US, and is usually smaller in large cities compared to nearby less populous areas. These results also demonstrated that a constant risk is not an informative null hypothesis, because some areas always had a higher SMER.

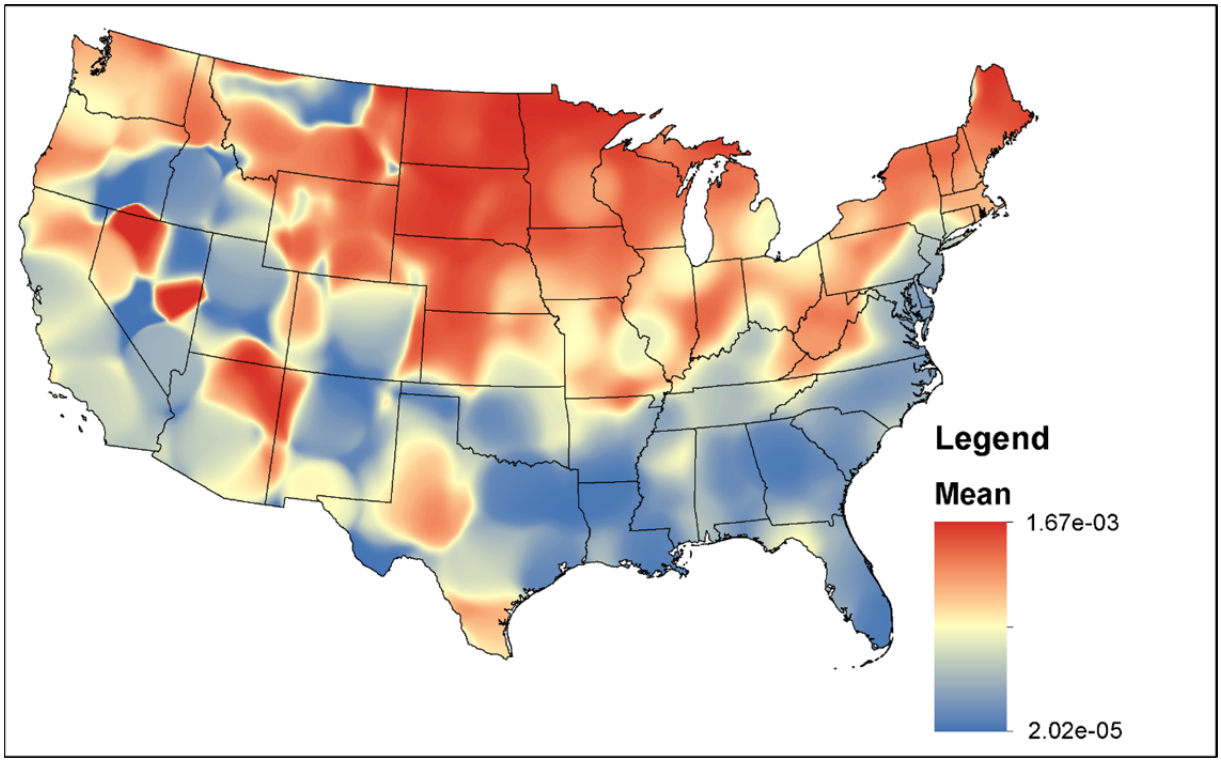

Figure 2.4: Baseline mean.

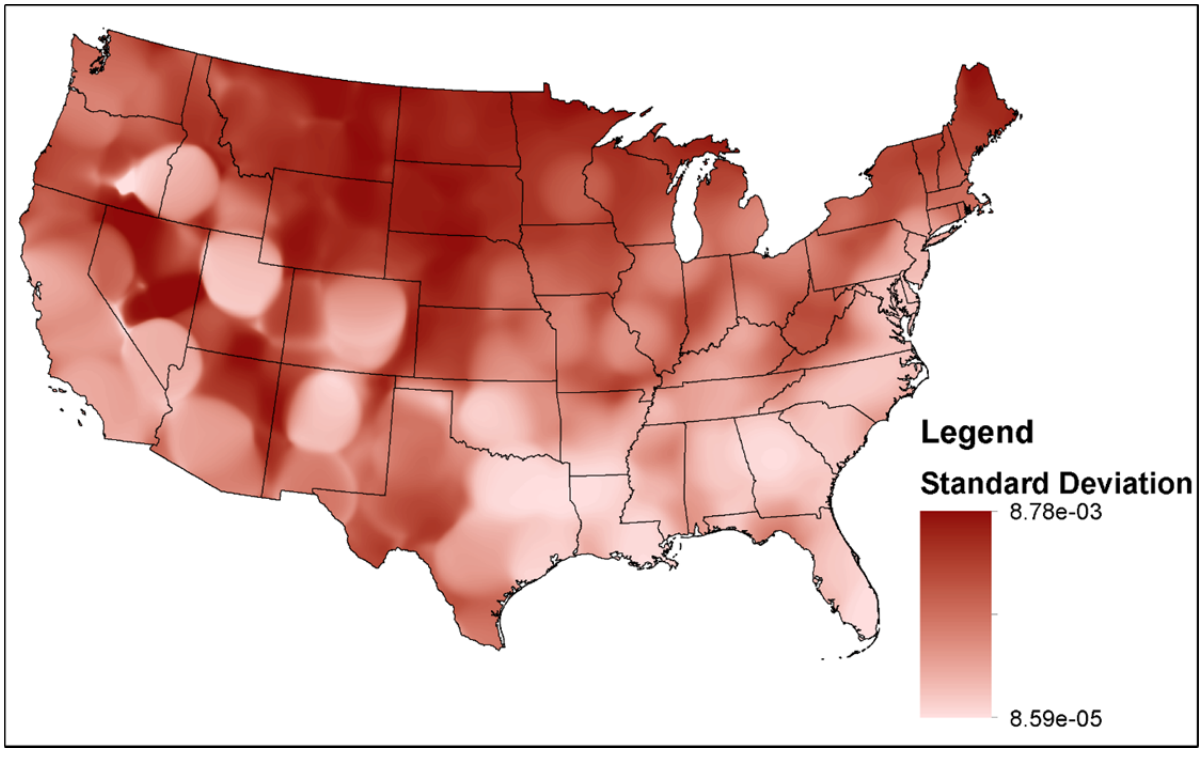

Figure 2.5: Baseline standard deviation

Figure 2.6 shows the estimated influenza activity maps from 09/07/2013 to 10/04/2013. These weeks still belong to our baseline period, and thus are shown here for comparison. Based on the estimation, in the week ending 09/13/2013 (Week 37), most of the US had a low influenza activity level. Beginning from the week ending 09/20/2013 (Week 38), the influenza activity level increased in several parts of the country, including the New England region, New York State, New Jersey, Michigan, Minnesota, Texas and Washington. In the next week (Week 39), this increasing trend continued especially in the New England region, Texas, Washington and Oregon. Finally, in the week ending 10/04/2013 (Week 40), the influenza activity was calming down in most regions except the west coast and the northeast part of the country. The *p*-value maps during these weeks are shown in Figure 2.7. Since these weeks are still within the baseline period, none of these maps have areas with a *p*-value less than 0.05. However, areas with *p*-value between 0.05 and 0.1 expand in the aforementioned areas.

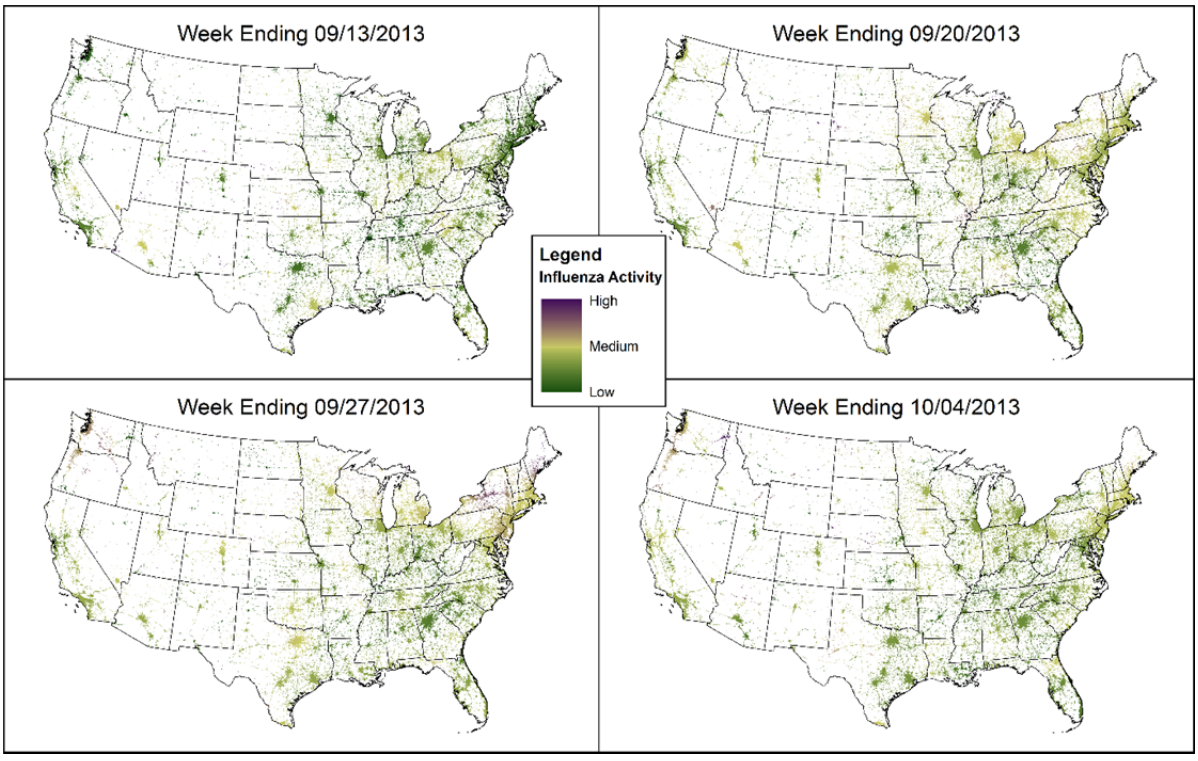

Figure 2.6: Influenza activity maps from 09/07/2013 to 10/04/2013.

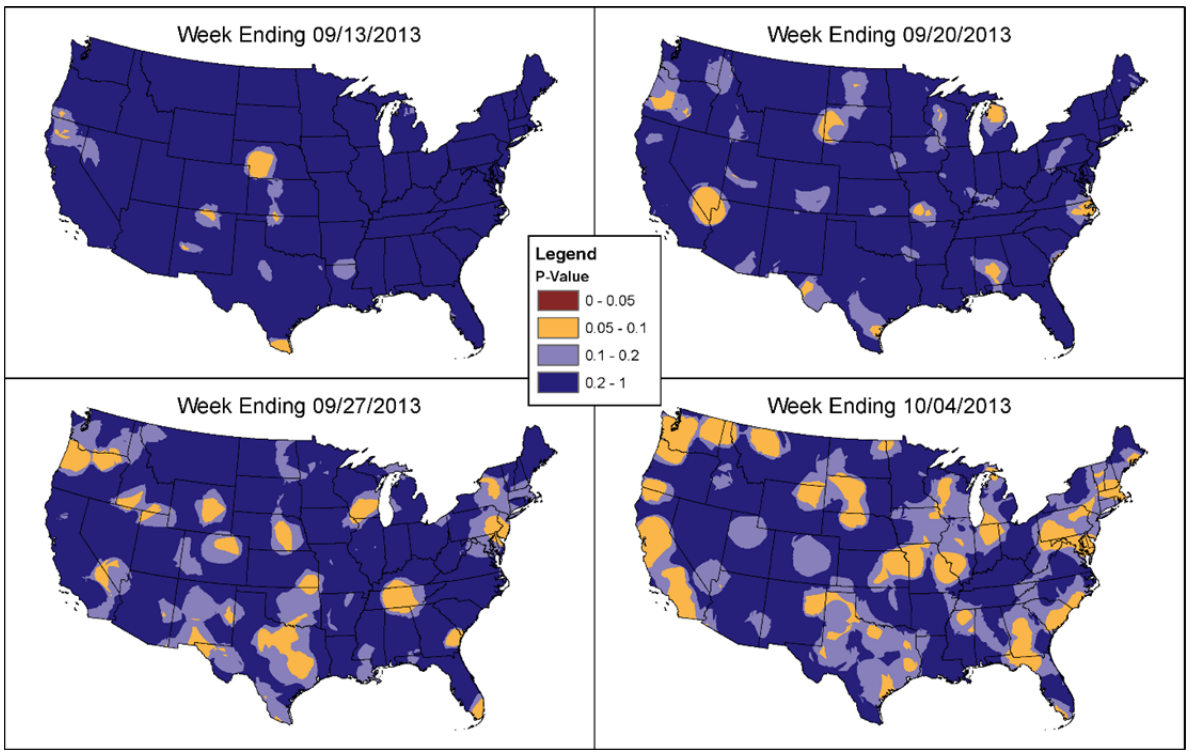

Figure 2.7: P-value maps from 09/07/2013 to 10/04/2013.

These detected patterns were similar compared to what reported by the CDC and state health departments. As reported by the CDC, the CDC ILI value in the New England region increased from 0.3% in Week 37 to 0.5% in Week 38. It kept on increasing to 1.0% in Week 39 and decreased to 0.6% in Week 40. The CDC ILI values in HHS region 2 (which mainly includes New York and New Jersey) were 0.7, 0.8, 0.9 and 1.6 in these four weeks. In Washington State, based on the reports from the Washington State Department of Health, the CDC ILI value also increased for almost three folds from Week 37 to Week 40 (Washington State Department of Health 2014). In Texas, the CDC ILI values were 4.15%, 4.49%, 4.63% and 3.69% in these four weeks, respectively (Texas Department of State Health Services 2013, 2014). These patterns were all captured in the estimated influenza activity maps.

Figure 2.8 and Figure 2.9 show the estimated influenza activity maps and *p*-value maps from 12/07/2013 to 01/03/2014 respectively. In this period, as reported by the CDC, areas with high influenza activity expanded from only several Southern states to most parts of the US, including the entire Southern parts of the US, the Midwest states, and the west coast. These patterns were depicted on the estimated influenza activity maps. On the map of the week ending 12/13/2013 (Week 50), mostly only the Southern parts of the country from Texas to South Carolina were having a higher influenza activity and a low *p*-value. The high influenza prevalence areas spread in the next three weeks similar to what was reported by the CDC. As of the week ending 01/03/2014 (Week 01), most parts of the conterminous US were having relatively high influenza prevalence and the SMERs were significantly higher than their baseline values, with noticeable exception in the northeast part of the country. Such results were consistent with the CDC's reports.

Figure 2.10 and Figure 2.11 show the estimated influenza activity maps and *p*-value maps from 01/25/2014 to 02/21/2014. The influenza activity level was decreasing in almost all parts of the US. In the week ending 01/31/2014 (Week 5), Texas, Oklahoma and Arkansas still had high influenza activity levels. States in the south and in the northeast were having moderate influenza activities. These areas also had high significance. Influenza activity in these parts decreased in the next few weeks. In the week ending 02/21/2014 (Week 8), most parts of the US were only having low influenza activity, except for certain parts of Minnesota, where the influenza activity increased slightly during this period. The *p*-value maps also indicated that SMERs in most of the US were no longer different from the SMERs in the last non-influenza season. These results were also consistent with the CDC's reports.

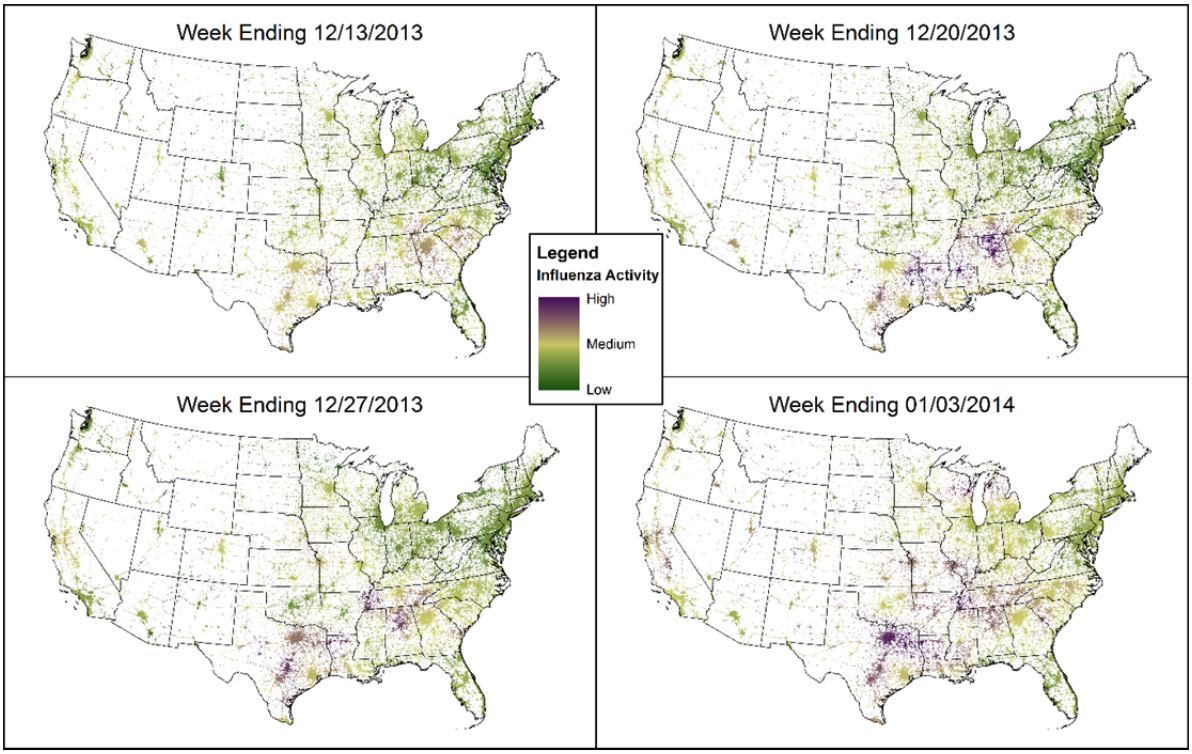

Figure 2.8: Influenza activity maps from 12/07/2013 to 01/03/2014.

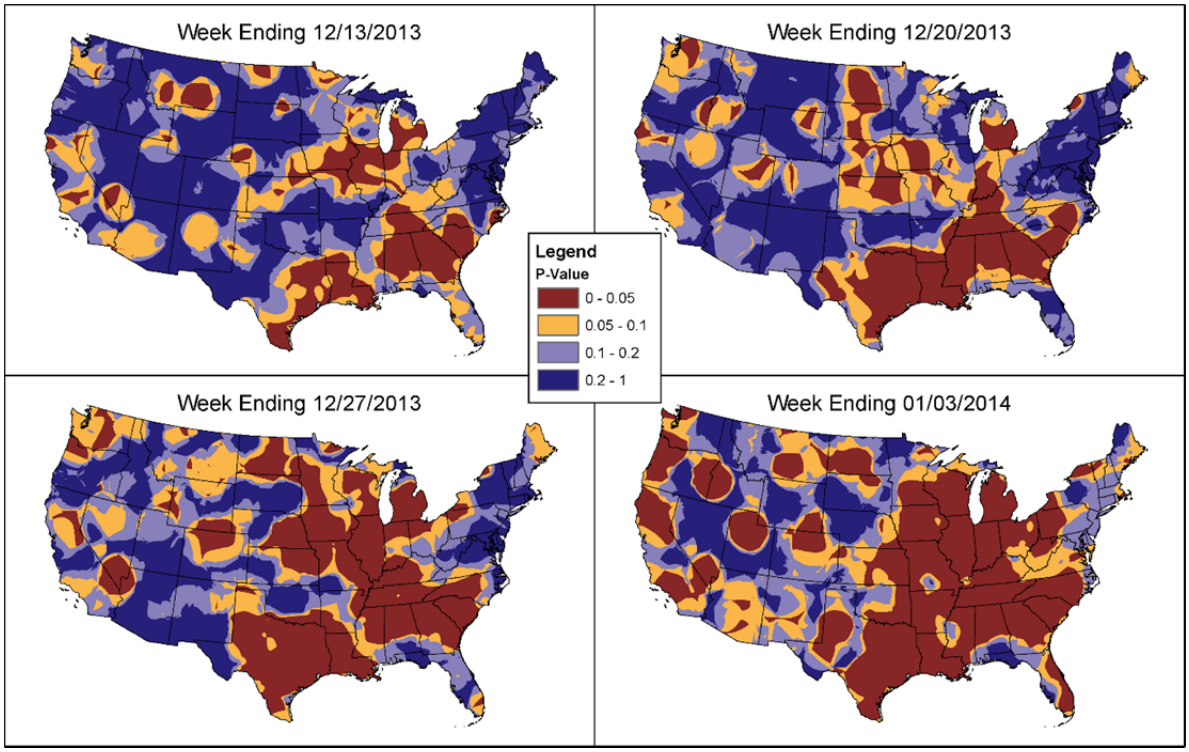

Figure 2.9: P-value maps from 12/07/2013 to 01/03/2014.

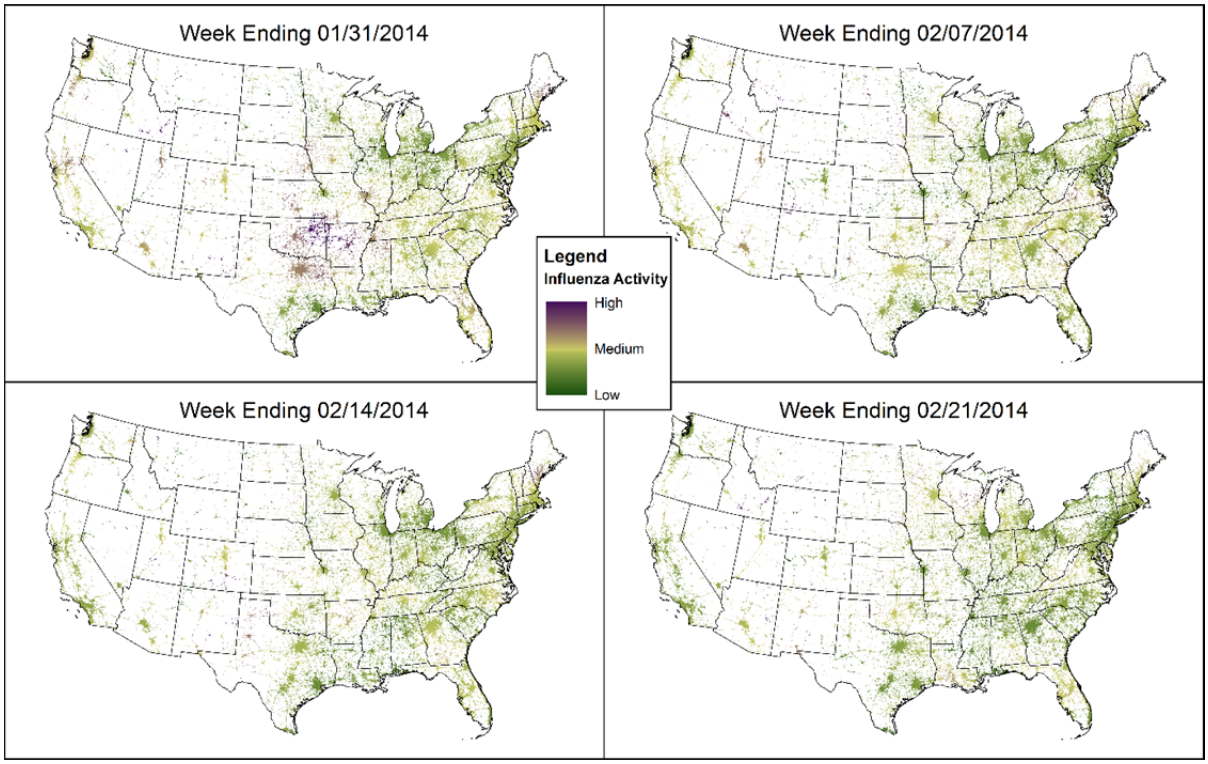

Figure 2.10: Influenza activity maps from 01/25/2014 to 02/21/2014.

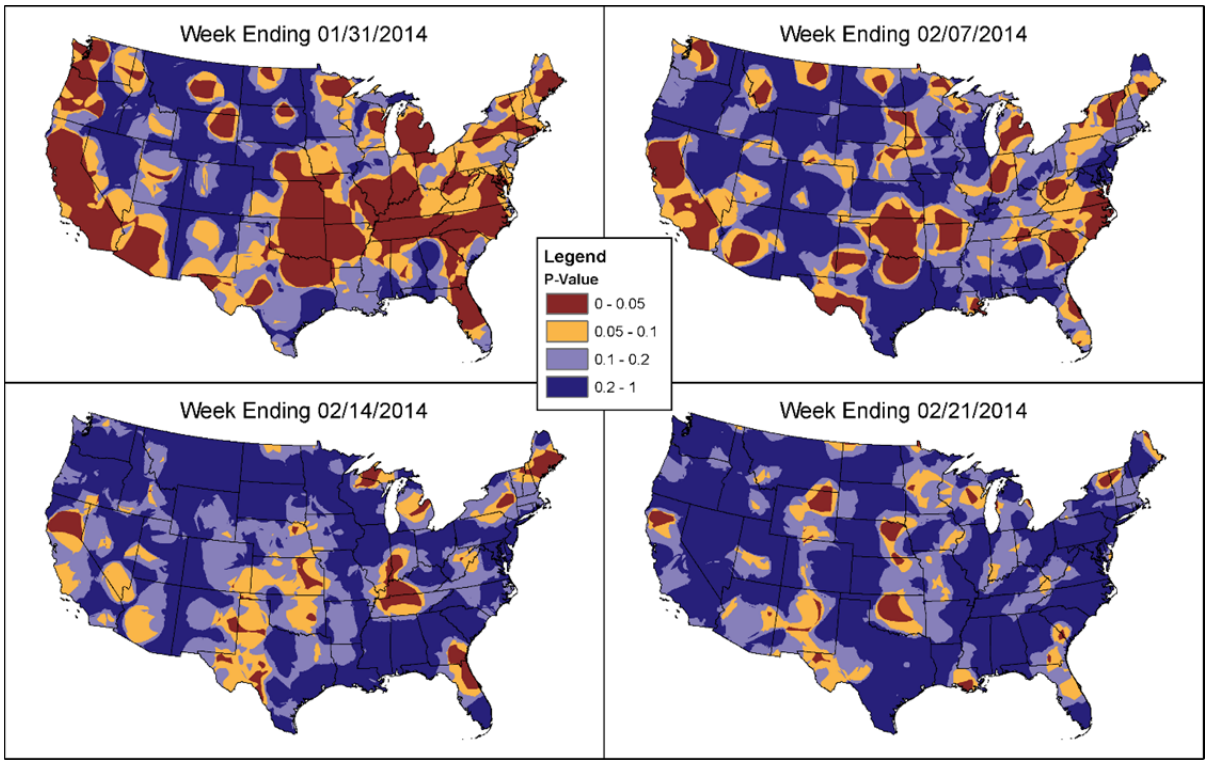

Figure 2.11: P-value maps from 01/25/2014 to 02/21/2014.

Besides being consistent with the official influenza reports, the influenza activity maps derived from social media data revealed potential fine-resolution spatiotemporal patterns of disease activities, which were not available from traditional surveillance systems. For example, as shown in Figure 2.8, the CDC and state Health Department only reported the overall influenza activities in Illinois and Missouri, while our results showed that the areas around St. Louis had a much higher influenza activity compared to the areas near Chicago. This finding was consistent with the result from Google Flu Trend's (Ginsberg et al. 2009) (GFT for short) results[4] (Table 2.4). For another example, based on our results, Dallas had the general trend of higher flu prevalence compared to Houston during the entire 2013-2014 influenza season (see Table 2.5 for comparison with GFT). Multiple limitations of the GFT are worth noting, however. One is that GFT's city-level results are not well justified compared to its state-level results. Another limitation is that the highest peak of the GFT has one or two weeks delay after the actual peak reported by CDC. While it is more desirable to compare our results with city-level ILI instead of GFT, such a comparison was not conducted mainly due to data limitations. Data in many cities during the study period are either not available (i.e. St. Louis, Dallas) or only cover influenza seasons (i.e. Houston).

Table 2.4: Influenza activity levels in Chicago and St. Louis (estimated using GFT's city-level result).

Week ending on	Chicago	St. Louis
12/13/2013	3.619462	7.110211
12/20/2013	7.015355	14.78967
12/27/2013	8.173206	16.89839
1/3/2014	5.742189	11.5174

[4] Data source: Google Flu Trends (http://www.google.org/flutrends)

Further, the estimated influenza activity maps were compared with a baseline method that simply calculated the normalized influenza activity levels for each bin without using KDE (Figure 2.12). The baseline result was generated using a bin size of 50km × 50km. The result was of lower quality and less informative when compared to the proposed methods (Figure 2.8). Since the number of influenza-related tweets in each bin was low, the estimated influenza activity level had a large variance. A bin's influenza activity level could easily shift between high and low values even in continuous weeks. In addition, neighboring bins could have very different influenza activity levels, and thus it was difficult to gain insights into spatiotemporal trends from these individual bin values.

Table 2.5: Influenza activity levels in Dallas and Houston (estimated using GFT's city-level result).

Week ending on	Dallas	Houston
11/24/2013	1.170433	2.044822
12/1/2013	2.335505	3.967126
12/8/2013	3.371937	3.687265
12/15/2013	7.524975	7.192494
12/22/2013	14.02424	9.659324
12/29/2013	12.44401	6.794863
1/5/2014	12.02885	5.115697
1/12/2014	8.706127	3.794904
1/19/2014	7.694546	3.189594
1/26/2014	4.651032	1.283752
2/2/2014	3.377784	0.893719
2/9/2014	1.373626	1.112796

2.6 CONCLUSIONS

This paper described a novel approach to mapping potentially unknown spatiotemporal patterns of events. It aims to provide informative hypotheses on event trends and distributions

worth following investigations based on social media data. Challenges that hinder the detection of informative patterns including the unstable data volume, spatial heterogeneity and data sparsity were identified. Interrelated strategies were employed in response to the challenges and produced consistent event activity indicators. The strategies include using KDE to generate smoothed social media intensity surfaces; utilizing event-unrelated posts to map the relative distribution of event-related posts; and normalizing event activity maps based on local baseline. The approach provides cartographic maps representing the estimated spatiotemporal variations of event prevalence as well as the statistical significance to identify space-time regions with potentially abnormal event activities. These results suggest probable event patterns that require additional investigations, and may be important for situation awareness of and timely responses to events.

The approach was applied to a real-world problem of detecting influenza trends in the conterminous US using geo-located *Twitter* data. Influenza-related tweets were extracted, and influenza activity maps were generated to capture fine-resolution spatiotemporal patterns of influenza activities. Many of the captured patterns could be explained by the best available ground truth. Furthermore, our approach provided a solid benchmark of influenza activity maps for future research based on social media or other new data sources. The case study also demonstrated that regression analysis methods only detect localized and superficial relationships between location-based social media and influenza activities. While these methods could be effective to predict influenza measures in designated regions, they may not be able to estimate the spatial extents of influenza activities. In addition, the case study provided an example of how potential unknown spatiotemporal patterns of an event could be estimated based on social media using the proposed approach. Except for the choice of influenza keywords and the training of

tweet classifiers, which are influenza-specific, the workflow can be generalized to other event mapping problems. The workflow first identifies both event-related and event-unrelated posts. Then, relative spatial distributions of event-related posts need to be estimated for each period. From these relative spatial distributions, localized baselines can be derived. Finally, the relative spatial distributions of event-related posts can be normalized to produce the final event activity maps and *p*-value maps.

This research showed that event patterns detected from social media have high consistency with ground truth data at available scales, in spite of the representativeness and the quality limitations of social media data. These limitations include users' demographics, their content preferences and the existence of robot or marketer accounts. While these limitations and the potential resulting biases cannot be ignored when using social media data, the advantages of social media data such as near real-time availability, large scale and individual level of granularity mean that social media may sometimes offer the only timely event indicator that helps to guide further investigations. The proposed approach leverages these advantages and mitigates limitations of social media data including unstable data volumes, unstable spatial distribution and spatial heterogeneity of event-related contents' prevalence. The results demonstrated that social media can, through scientific approaches, quickly inform us of unknown patterns that are likely to be important and valid.

Our approach has the following limitations. One limitation is due to the drawback of the normalized KDE method in eliminating extreme values and providing informative estimation in areas with low social media post density. For an area with few posts, occasional appearances of several event-related posts could make the area stand out as a high event activity area. Hence, event activity levels estimated in less populous areas are less reliable, and further improvement is

needed in the future. Second, the effectiveness of the approach may rely on the choice of baseline periods. While for influenza, a non-influenza season naturally serves as a valid baseline, it may not be the case for other events. Approaches or guidelines for baseline choices are thus another future research direction.

The quantitative evaluation of the estimated event activity maps is constrained by the situation that there is no gold standard or ground truth dataset to compare with. Thus, to better understand the uncertainty is an important future research direction. One strategy is to synthesize multiple social media data sources collected independently while resolving inconsistencies between patterns detected from these datasets. This strategy is also promising to mitigate the representativeness and quality issues of social media data.

CHAPTER 3: A MULTIDIMENSIONAL SPATIAL SCAN STATISTICS APPROACH TO MOVEMENT PATTERN COMPARISON[5]

Abstract. *This paper describes a multidimensional spatial scan statistics approach to comparing spatial movement patterns based on origin-destination (OD) representation. This approach aims to evaluate differences and similarities between the spatial distributions of a pair of OD movement datasets, and detect areas where the two spatial distributions differ the most. Specifically, two OD datasets being compared are modeled as a bivariate marked spatial point process in a multidimensional space, consisting of points representing individual OD movement records. Such multidimensional space is formed by the Cartesian product of the origins' and the destinations' geographic spaces. With this spatial data model, one can evaluate how two movement distributions differ from each other by testing against a random labeling null hypothesis. A multidimensional Bernoulli spatial scan statistics method is developed to detect OD region pairs with abnormally high concentrations of one movement dataset over the other. The existence and the spatial extents of these OD region pairs indicate whether and where the two movement distributions differ. Two case studies were conducted to evaluate the approach by comparing morning and afternoon taxi trips (individual movements), and county-to-county migration flows between age groups (aggregated movement flows), and demonstrated that areas with the most significant spatial distribution differences could be detected from large movement datasets.*

Keywords: CyberGIS, Movement Analysis, Spatial Analysis and Modeling, Spatial Scan Statistics

[5] Reprint, with permission, from Gao *et al*., 2018. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science*, doi: 10.1080/13658816.2018.1426859.

3.1 INTRODUCTION

The analysis of geographic mobility data and associated spatial interactions is of great importance to understanding complex geographic phenomena and their space-time dynamics (Dodge et al. 2016, González et al. 2008, Guo and Zhu 2014). Deep insights can be gained into geographic mobility data through the comparison of spatial patterns revealed from different sets of movements (Illian et al. 2008, McGuckin and Murakami 1999, Calabrese et al. 2011). Despite that movement pattern comparison has been examined in previous research through visual analytics (Guo and Zhu 2014) or correlation analysis (Calabrese et al. 2011), research on systematically evaluating whether the spatial patterns of two movement datasets differ is missing. As a consequence, given two movement datasets, it is challenging to measure whether they have different spatial hotspots or their pattern differences can be explained by random chances.

This paper presents a systematic approach to comparing any two spatial movement patterns based on OD representation. OD data are frequently used in many applications including for example human migration (Tobler 1981, Tubergen *et al*. 2004), and traffic analysis (Cascetta and Nguyen 1988, Bell 1991). Such data record the origin and destination of each movement but not any intermediate locations. The spatial patterns of OD movements are reflected in their spatial distributions - the arrangement of individual OD movement records in space. Two OD datasets have different spatial patterns if the two spatial distributions are different - there are pairs of origin regions and destination regions where one movement dataset has a much higher concentration than the other. Hence, the existence and the spatial extents of these OD pairs indicate whether and where the two movement distributions differ.

Comparing OD movement patterns requires a spatial data model that integrates origins and destinations of movements. OD data contain information about the spatial distributions of origins and destinations, and pairwise connections between them. The spatial patterns of two OD movement datasets can be different even if they have exactly the same origin and destination distributions. Separating origin and destination in movement analysis leads to fragmented views of the movement patterns. Hence our approach integrates both of them into a single analysis unit. Specifically, we construct a multidimensional space by fusing the origins' space and destinations' space through a Cartesian product, and model each OD record as a single point in the multidimensional space. Such a space has four spatial dimensions - two dimensions for origins and two for destinations. The spatial distribution of OD movements is reflected in the distribution of points in the multidimensional space. By giving different marks (labels) to points representing movements in different datasets, any two movement datasets being compared are modeled as points from a bivariate marked spatial point process. Hence, the differences between two movement distributions can be evaluated by detecting local point clusters of one dataset against the null hypothesis that the two datasets have the same spatial distribution.

In order to detect such point clusters, a multidimensional Bernoulli spatial scan statistics method is developed by extending from Kulldorff's (1997) spatial scan statistics. The method tests the spatial distributions of the bivariate marked spatial point process against a random labeling null hypothesis. The random labeling null hypothesis assumes that the label of a point is independent of its location, and thus the pair of OD datasets represented by these points has the same spatial distribution. The method recursively checks pairs of origin area and destination area in order to find the pairs that are the most likely clusters. It further uses a Monte Carlo simulation to evaluate the statistical significance of identified clusters. Finally, these clusters, with their

spatial extents and statistical significance, are summarized to describe the spatial pattern differences.

The proposed approach can help to answer questions that include but are not limited to:

- Understanding the similarities and differences between movement patterns of people with different genders, ages, ethnic groups or occupations;
- Comparing migration patterns during different years, or daily travel patterns during different hours; and
- Comparing movement patterns collected or estimated from different data sources, and evaluating the quality of new data sources (e.g. social media, mobile phones) for movement flow estimation.

There are additional advantages of the proposed approach. First, different from the convention of spatial interaction data analysis and modeling, where movement or interaction records are aggregated by areal units to derive OD matrix and follow-up analysis is mostly aspatial except using distances for deterrence functions (Fischer and Wang 2011), our approach takes movement datasets as point-based records through the entire analytical process. Furthermore, the approach detects spatial clusters by considering the spatial proximity between individual movement records, and does not rely on predefined zones for aggregation. Second, the spatial scan statistics method for detecting movement clusters can be easily parallelized to exploit cyberGIS (aka geographic information science and systems based on advanced computing and cyberinfrastructure) (Wang and Armstrong 2009; Wang 2010; Wang *et al*. 2016). Hence our approach can be applied to large movement datasets with hundreds of thousands of OD pairs and even movement records at the individual level.

The proposed approach was evaluated using two case studies: (1) comparing morning and afternoon taxi trips in New York City and (2) comparing migration patterns of young and senior people from county-to-county migration dataset in the US. The first case study provides an example of point-based OD data, where there are unique origin and destination for nearly every individual movement. The second case study uses area-based OD data, where movement records are already aggregated counts between counties. The experiment demonstrated that areas with the most significant spatial distribution differences could be detected from these large movement datasets.

3.2 RELATED WORK

3.2.1 Flow Mapping

One common approach to OD data presentation and visual-analytics is flow mapping based on representing origins and destinations using arrows or bands (Tobler 1981, Tobler 1987). When flow datasets are comprised of a large number of flows, traditional flow mapping approaches are no longer effective due to visual clutter (Cui et al. 2008, Holten and Van Wijk 2009). In order to resolve visual clutter, three different approaches have been proposed in the literature: location aggregation, flow crossing reduction and flow generalization. Location aggregation reduces the total number of flows by reducing the number of candidate origins and destinations (Tobler 1987, Andrienko and Andrienko 2008, Guo 2009, Andrienko and Andrienko 2011, Guo et al. 2012). It combines individual level movements or flows between small units into flows between larger regions. Flow crossing reduction reroutes and bundles (clusters) edges to minimize crossings between flows (Phan et al. 2005, Holten and Van Wijk 2009, Cui et al. 2008, Buchin et al. 2011). Flow generalization extracts representative flow samples to depict major flow patterns through flow sampling (Guo and Zhu 2014) or flow clustering (Zhu and Guo

2014). Furthermore, cyberGIS-based methods, which use a hybrid approach of all three techniques, have been developed (Padmanabhan et al. 2014, Wang et al. 2014).

A major limitation of flow mapping is that it is best suited for visualization rather than pattern analysis. While an informative visualization can help understand general trends in movement data, the exact spatial structures of movement patterns may not be apparent via visual inspections especially when patterns consist of overlapping flows. It is practically impossible to detect subtle differences and fine distinctions between movement patterns by only using visual analytics. Thus, statistical methods are necessary in order to provide adequate quantifications and standardized descriptions of patterns (Illian *et al*. 2008). Therefore, this paper focuses on developing statistical models and analytical procedures to systematically evaluate the spatial structure differences between movement patterns.

3.2.2 Spatial Scan Statistics

Scan statistic was originally developed to detect clusters in point processes in one-dimensional (Naus 1965b) or two-dimensional space (Naus 1965a). It was further extended into spatial scan statistics by Kulldorff (1997), which allows the area of scanning windows to vary and can detect clusters in spatiotemporal point processes. Spatial scan statistics have then been applied to a wide range of research domains, including epidemiology, public health, ecology, crime analysis, and astronomy, to identify clusters of events (Kulldorff 2015).

The original and most popular point process models in spatial scan statistics are (homogeneous or inhomogeneous) Poisson process and Bernoulli process (Kulldorff 1997). A Poisson model deals with the number of events occurring in a time interval and a spatial region. A Bernoulli model handles events that are in either one of two states, which is often used to compare the spatial distributions of two types of events, such as in a case-control study. Other

models include space-time permutation (Kulldorff *et al*. 2005), ordinal (Jung *et al*. 2007), exponential (Huang *et al*. 2007), normal (Kulldorff *et al*. 2009) and multinomial (Jung *et al*. 2010). The Bernoulli spatial scan statistic is adapted and extended in this paper to compare any two sets of OD movements.

The most widely used software program of spatial scan statistics is SaTScan™ developed by Kulldorff (2015) (the user guide listed hundreds of research papers in multiple domains using the software). However, this software only supports analyses in geographic space (with two spatial dimensions and optionally one temporal dimension). Furthermore, SaTScan™ is not scalable to large datasets. Finally, despite being free to use, SaTScan™ is not open source (Baker and Valleron 2014). Without open implementation details, it is hard to adapt or modify its method for our needs. Hence, a multidimensional spatial scan program is designed and developed in this research.

3.3 DATA MODEL

3.3.1 OD Data Representation

Let $M = \{M_1, M_2, \ldots, M_n\}$ be an OD movement dataset that has n records. $M_i =< S_{oi}, S_{di}, I_i >$ is one OD movement record that starts at location S_{oi} and ends at location S_{di}. Both $S_{oi} =< x_{oi}, y_{oi} >$ and $S_{di} =< x_{di}, y_{di} >$ are points in 2D geographic space. In order to distinguish origin and destination, the origin's 2D geographic space is referred to as $X_o \times Y_o$, and the destination's 2D geographic space is referred to as $X_d \times Y_d$. However, it is worth noting that $X_o \times Y_o$ and $X_d \times Y_d$ usually refer to the same study area. I_i indicates non-spatial attributes of the movement records. In the context of spatial movement pattern comparison, I_i is a binary indicator variable that specifies which of the two types this movement belongs to. For instance,

when comparing the spatial movement patterns of young and senior people, I_i is used to indicate whether it is a trip by a young or senior person.

In many applications, movement records are collected or reported in an aggregated form. Instead of recording the origin and the destination of each individual movement, only the numbers of movements between each pair of regions are tracked. For example, the US Census Bureau only reports the number of individuals migrating between each pair of counties. In this case, $M_{agg} = \{M_{agg\_i}\}$ is used to represent a set of aggregated movements that has one record for each origin-destination pair with at least one movement, where $M_{agg\_i} =< S_{oi}, S_{di}, A_i, B_i >$. A_i and B_i are the counts of trips from two datasets that start at S_{oi} and end at S_{di}. For simplicity without losing generality, we use points (e.g. areal centroids) to represent S_{oi} and S_{di}.

3.3.2. Multidimensional Spatial Point Data Model

In this paper, each OD record M_i is modeled as one spatial point $P_{Mi} =< x_{oi}, y_{oi}, x_{di}, y_{di} >$ in a 4D space $X_o \times Y_o \times X_d \times Y_d$. This 4D space results from the Cartesian product of the origins' 2D geographic space $X_o \times Y_o$ and the destinations' 2D geographic space $X_d \times Y_d$. We refer to this space as OD space in the rest of the paper. Hence, an OD movement dataset M is modeled as a set of 4D points $P_M = \{P_{Mi}\}$, which is a realization of its underlying spatial point process. When the binary indicator variable I_i is considered as the mark, the underlying spatial point process is a bivariate marked spatial point process (Diggle 2013) in the OD space, which contains points representing the two sets of movements being compared. In this OD space, a point $< a, b, c, d >$ represents a movement that originates from $< a, b >$ and ends at $< c, d >$. Both $< a, b >$ and $< c, d >$ represent points in the 2D geographic space. A 4D region contains a collection of movements from some geographic area A to some geographic area B, where A and B may have intersections.

The multidimensional data model integrates origins and destinations of movements into a single analysis unit, and thus preserves the pairwise connections between them. With this model, two movement records, $< a_1, b_1, c_1, d_1 >$ and $< a_2, b_2, c_2, d_2 >$ are near each other only if they have both near origins and near destinations. Spatial point patterns detected in the OD space describe the spatial arrangements of such integrated origin-destination units.

3.4 METHOD

3.4.1 Overview

An overview of the proposed approach is shown in Figure 3.1. The approach begins with two input OD datasets (α and β in Figure 3.1). Each record in these datasets has an origin point and a destination point, and is modeled as a point in multidimensional OD space. A multidimensional spatial scan statistic is then used to detect point clusters in the OD space. In Figure 3.1, one cluster is detected with a high density of points from dataset α. This point cluster has a region in the original space and one in the destination space, and represents a movement cluster with a higher concentration of dataset α over β. The existence and the spatial extents of these movement clusters indicate whether and where the two movement distributions differ.

3.4.2 Random Labeling Null Hypothesis and Pattern Difference Representations

With the multidimensional spatial point data model, the comparison of any two OD distributions can be achieved by analyzing the relationships between points from a bivariate marked spatial point process in an OD space. When analyzing such relationships, there are two straightforward benchmark (null) hypotheses: independent and random labeling (Diggle 2013). In an independent hypothesis, the two types of points are generated by two independent univariate point processes. In a random labeling hypothesis, each point of a univariate point

process is labeled based on independent Bernoulli trials. In other words, the two types of points have exactly the same spatial distributions.

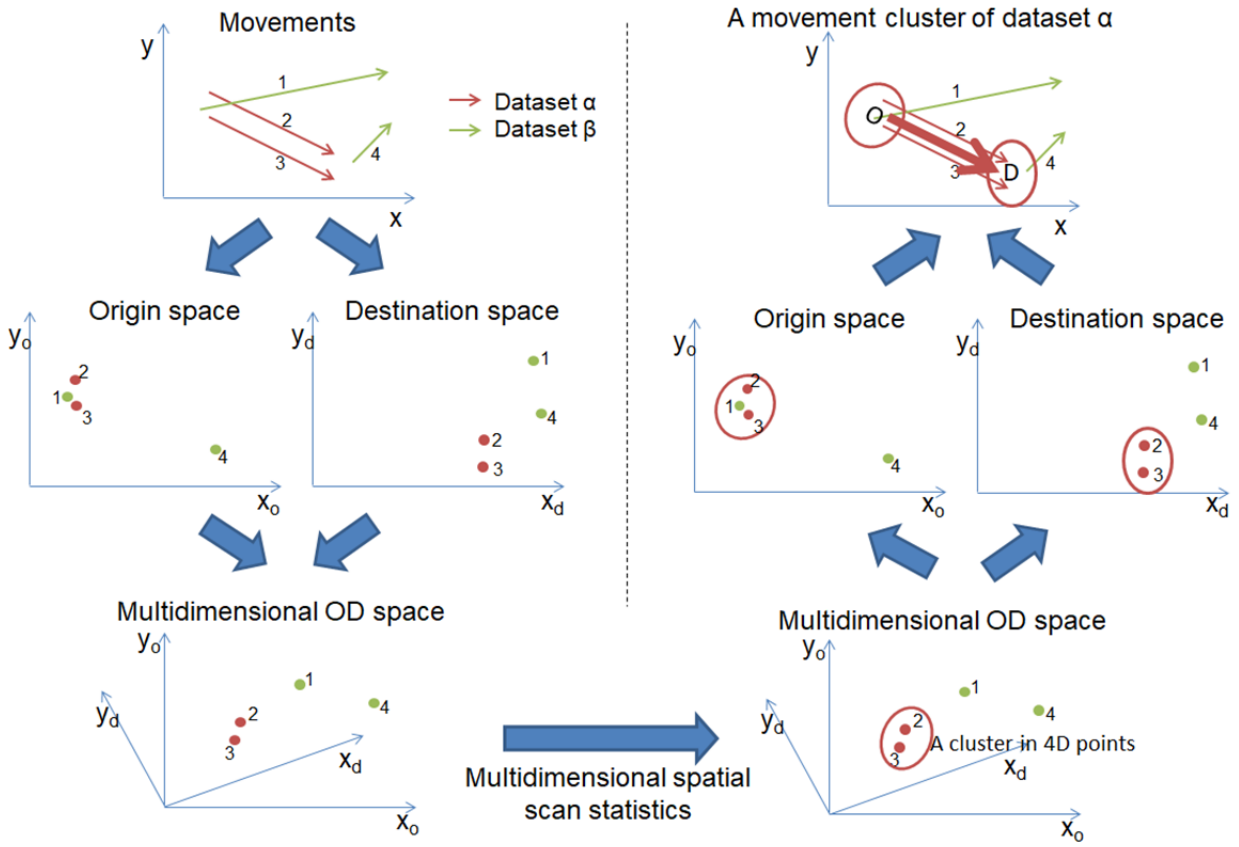

Figure 3.1: Conceptual illustration.

The independent null hypothesis is ill-suited to movement distribution comparison, because the two sets of movements are often dependent - both are more likely to be from and to more popular places. In contrast, the random labeling hypothesis is well suited to detect whether two movement distributions are different. Under the random labeling hypothesis, a randomly chosen movement record has a constant probability to be from a certain set, regardless of its origin or destination. This hypothesis also infers that the two types of movements have the exact same spatial distribution, and that the origin and the destination of a trip contain no information

about its type. Hence, by testing this hypothesis, the differences between two movement distributions can be evaluated.

When detecting the differences between two OD movement distributions against this null hypothesis, it is natural to find pairs of origin and destination regions where there are abnormally high concentrations of one movement dataset over the other. For example, if senior migrants are half as many as young migrants worldwide, an OD region pair with 1000 senior and 800 young migrants represents a higher concentration of senior migrants. As demonstrated previously, such a pair of areas is represented as a 4D region in the OD space. Hence the differences between two OD movement distributions can be represented as a list of such 4D regions (clusters). To this end, a multidimensional Bernoulli spatial scan statistic is used to find these 4D clusters.

3.4.3 Multidimensional Bernoulli Spatial Scan Statistics

3.4.3.1 Multidimensional spatial scan statistics

The multidimensional spatial scan statistics extend Kulldorff's spatial scan statistics (Kulldorff 1997). It aims to detect the locations and sizes of the most likely clusters of OD movements represented by points in the OD space. This is done by first putting a large number of 4D scanning windows over the study area. These scanning windows have different sizes, and represent different pairs of origin areas and destination areas. The maximum likelihood of each scanning window to be a cluster, L_W, is calculated based on the underlying point process models. The maximum likelihood if there is no cluster is represented as L_0. Then, the scan statistic λ of each scanning window is defined as the ratio of L_W to L_0 (Equation (3.1)). The scanning window with the highest likelihood ratio is identified as the primary cluster, and other secondary clusters may also be detected. Finally, the statistical significance (p-value) of each detected cluster is estimated using Monte Carlo simulation.

$$\lambda = L_W / L_0 \tag{3.1}$$

3.4.3.2 Bernoulli model

In this paper, a Bernoulli process model is used to compare distributions of two sets of movements. For intuitive reference, the two sets are referred to as α and β. A Bernoulli model aims to detect clusters in bivariate marked spatial point processes under a random labeling null hypothesis, conditioning on the origins and destinations (but not types or labels) of movements. Suppose that we have a scanning window W in the study area A. A Bernoulli model states that each point inside W has a Bernoulli distribution of probability p to be from α (and $1 - p$ to be from β), and each point outside has a Bernoulli distribution of probability q to be from α. Under the random labeling null hypothesis, any movement record has a constant possibility regardless of its origin and destination, and hence $p = q$. The likelihood of the observation under the null hypothesis L_0 is maximized when $p = q = N_\alpha/(N_\alpha + N_\beta)$, which is shown in Equation (3.2). Here, N_α is the total number of movements from α; N_β is the total number of movements from β respectively; $N_\alpha(W)$ and $N_\beta(W)$ are the numbers of movements in the scanning window W.

$$L_0 = p^{N_\alpha(W)}(1-p)^{N_\beta(W)} q^{N_\alpha - N_\alpha(W)}(1-q)^{N_\beta - N_\beta(W)} = \left(\frac{N_\alpha}{N_\alpha + N_\beta}\right)^{N_\alpha}\left(\frac{N_\beta}{N_\alpha + N_\beta}\right)^{N_\beta} \tag{3.2}$$

Under the alternative hypothesis, $p \neq q$, which means the scanning window W is either a (high) cluster of α movements $(p > q)$ or a (high) cluster of β movements $(p < q)$. When $p = N_\alpha(W)/(N_\alpha(W) + N_\beta(W))$ and $q = (N_\alpha - N_\alpha(W))/(N_\alpha + N_\beta - N_\alpha(W) - N_\beta(W))$, the likelihood L_w reaches its maximum as in Equation (3.3).

$$L_W = \left(\frac{N_\alpha(W)}{N_\alpha(W)+N_\beta(W)}\right)^{N_\alpha(W)}\left(\frac{N_\beta(W)}{N_\alpha(W)+N_\beta(W)}\right)^{N_\beta(W)}\left(\frac{N_\alpha - N_\alpha(W)}{N_\alpha+N_\beta-N_\alpha(W)-N_\beta(W)}\right)^{N_\alpha - N_\alpha(W)}\left(\frac{N_\beta - N_\beta(W)}{N_\alpha+N_\beta-N_\alpha(W)-N_\beta(W)}\right)^{N_\beta - N_\beta(W)} \tag{3.3}$$

From Equation (3.2), L_0 only depends on the total number of α movements and the total number of β movements, and thus is constant across all scanning windows once datasets are given. Thus only L_W needs to be considered to find clusters of movements and to test the statistical significance of each cluster.

3.4.3.3 Cluster detection

While the primary cluster is detected as the scanning window with the highest likelihood, finding secondary clusters is not straightforward. As pointed out by Kulldorff (2015), expanding or reducing the primary cluster's size only marginally often forms a secondary cluster that is almost identical to the primary one; however, this type of clusters usually provides little information about underlying spatial patterns. Hence in this paper, a non-intersection standard is used when finding secondary clusters. Specifically, no clusters should have intersections with the existing ones. This intersection infers that no 4D regions representing clusters can intersect with each other. However, two clusters can have intersecting origins or destinations if the clusters are separated in 4D space. With this standard, after each cluster is detected, all scanning windows that intersect with it are eliminated. The next cluster, if any, will be detected based on the remaining scanning windows with the highest likelihood.

There are two additional notes for the cluster detection process. First, a movement cluster with an origin area and a destination area that are intersecting or even identical is meaningful, as it represents movements within a region. However, these clusters may be eliminated by excluding scanning windows with intersecting OD pairs, if researchers only want to compare inter-region movement patterns. Second, a scanning window from a 2D region A to another 2D region B and another scanning window from B to A are two independent scanning windows. They represent two separate 4D regions in the OD space (if A and B are not intersecting). Hence,

it is possible to have them both detected as clusters of type α movements. For instances, if the task is to compare young and senior migration patterns, and both A and B are regions with more senior people, it is possible to have both A to B and B to A as clusters of senior migrants.

3.4.3.4. Monte Carlo simulation

In order to conduct the statistical inference, it is necessary to find the distribution of the highest likelihood that can be generated by random chance under the null hypothesis. Since it is difficult to find the analytical form of such distribution, spatial scan statistics use Monte Carlo simulation to estimate the p-value of each detected cluster (Kulldorff 1997). During the Monte Carlo simulation, many (N_{Rep}) random replications of movement datasets are generated under the null hypothesis. Following Kulldorff's (1997) spatial scan statistics, each simulated dataset is generated by randomly assigning N_α labels of α and N_β labels of β to all individual movements. For each replication, the same clustering detection procedure as described in section 4.3.1. is applied to find the cluster with the highest likelihood (the primary cluster). During the simulation, only the likelihood of the primary cluster, rather than the cluster's location, is necessary for each replication. The p-value of each original cluster is estimated using Equation (3.4), where $N_{Extreme}$ is the number of replications with a larger highest likelihood than the cluster's actual likelihood. This statistical inference approach is able to resolve the multiple testing problem, since every detected cluster, including secondary clusters, is tested against the highest likelihood of all scanning windows that can be generated by random chances (Kulldorff and Nagarwalla 1995).

$$P = \frac{N_{Extreme}+1}{N_{Rep}+1} \tag{3.4}$$

The steps of the Monte Carlo simulation are summarized below:

(1) Keep track of all the original clusters detected before the Monte Carlo simulation, including the likelihood value of each cluster;

(2) Generate a simulated dataset by randomly assigning labels;

(3) Detect the primary cluster in the simulated dataset and only record its likelihood value;

(4) Repeat step (2) and (3) N_{Rep} times;

(5) Estimate the p-value of each original cluster recorded in step (1) using Equation (3.4).

3.4.4 Implementation and Computation

The spatial scan statistics method to detect movement clusters can be parallelized in a straightforward way. This is because the most compute-intensive component - the calculation of scan statistics for each scan window - is independent and thus can be calculated in parallel by different computing processes. The algorithm for 4D Bernoulli spatial scan statistic is implemented using C and OpenMP (Open Multi-Processing). We plan to make the software open-sourced. OpenMP is an application programming interface for shared memory multiprocessing programming. It is used in our implementation to distribute the scan statistics calculation and cluster detection to multiple processor cores, while allowing them to simultaneously access all OD records stored in the shared memory.

Since there are an infinite number of potential scanning windows, it is only possible to limit clusters to be of certain shapes and find the best ones among them. Our software makes use of 4D spherical scanning windows, as circular (spherical) scanning windows are most widely used in spatial scan statistics, and are easy to implement (Kulldorff 1997, Kulldorff 2015). In this

research, scanning windows with centers at 4D points that represent input OD records are used. In order to determine whether a point is in a window or not, 4D Euclidean distance is used to define the distance between a 4D point and a cluster center. A point $< x_{oi}, y_{oi}, x_{di}, y_{di} >$ is considered in a window centered at $< x_{oc}, y_{oc}, x_{dc}, y_{dc} >$ if $\sqrt{(x_{oi} - x_{oc})^2 + (y_{oi} - y_{oc})^2 + (x_{di} - x_{dc})^2 + (y_{di} - y_{dc})^2}$ is smaller than or equal to the window radius. Numbers of an equal interval (e.g. 1km, 2km, ..., 100km) are used as scanning window radiuses. These values also define the maximum radius of clusters, which controls the spatial scale of pattern analysis. If there are n point OD records and a sequence of m numbers as window radiuses, the software will check a total of n × m scanning windows to detect clusters. An alternative way to define spherical scanning windows is to use a regular grid as centers. An advantage of this strategy is that a cluster does not need to have a point at its center. However, the computing intensity of this approach is too high in 4D space. For example, in order to have a 100 × 100 grid for both origins and destinations (which is a coarse resolution), there will be 100 million potential cluster centers in 4D space to test, which poses a significant computational challenge.

3.5 CASE STUDIES

To evaluate the effectiveness of the proposed approach to movement pattern comparisons, two case studies are carried out in this paper. The first case study compares the taxi trips in New York City within different time periods during a day. The taxi trip data provide an example of point-based OD data, where each trip has unique pick-up (origin) and drop-off (destination) location recorded by GPS devices. The second case study compares the migration patterns among different age groups from a US county-to-county migration dataset. This dataset is area-based - migration records are aggregated by county and are recorded as counts of

migrating individuals between counties. The experiments were conducted using a computing node with two Intel Xeon E5-2660 processors (10 cores each) and an RAM of 256G.

3.5.1 New York City Taxi Trip Analysis

3.5.1.1 Data source

All the New York City taxi trips (of both Yellow and Green taxis) on a typical workday (01/21/2015) with no major events were retrieved from the official website of New York City6. The data include information of pick-up and drop-off timestamps and locations of each taxi trip that originates on that day. In total, there are 440,373 valid taxi trips.

This case study compares the taxi trip distributions during the morning rush hours (7:00 AM to 10:00 AM) and the distributions during afternoon rush hours (5:00 PM to 8:00 PM). For this purpose, taxi trips that start in these two time periods are extracted respectively. These two sets of trips are called morning trips and afternoon trips respectively. There are 72,144 morning trips and 83,524 afternoon trips. Hence, under the null hypothesis of random labeling, when the two sets of trips are combined, each randomly chosen trip has a constant probability of $72{,}144/(72{,}144 + 83{,}524) = 0.463$ to be a morning trip, regardless of its origin and destination.

3.5.1.2 Results

The centers of top 30 clusters are shown in Figure 3.2. These clusters are detected with a maximum scan window radius of 2.5km. Since the center of each cluster is a 4D point in OD space that corresponds to an origin-destination pair, it is plotted as an arrow from the center of its origin to the center of its destination. The color of an arrow indicates whether the cluster represents a concentration of morning or afternoon trips in NYC. These clusters depict the major

[6] Data source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

distribution differences between morning and afternoon taxi trips. On the one hand, trips to Manhattan, especially to the Midtown and the Lower Manhattan CBD, consist of significantly more morning trips than afternoon trips. On the other hand, there is a much higher probability for afternoon trips to leave from Manhattan and especially the Midtown than morning trips.

Figure 3.2: New York City taxi trips. Centers of top 30 clusters with a maximum radius of 2.5km.

For a closer look at the pattern differences, the spatial extents of top five clusters are shown in Figure 3.3. Table 3.1 shows the summary statistics of these five clusters. In Table 3.1, the type indicates whether it is a high cluster of morning or afternoon trips; the morning trips and the afternoon trips represent actual OD trip counts in the cluster; the expected morning trips and the expected afternoon trips represent the expected trip counts if morning trips and afternoon trips have the same spatial distribution. The p-value is estimated using Monte Carlo simulation with 999 runs. The top 5 clusters are mostly in Manhattan, except for the 5th cluster whose destination region contains part of Brooklyn. This result is not surprising since Manhattan has the most taxi trips and the most dramatic traffic flow changes throughout a day. Among the 5 clusters, the 1st, the 4th and the 5th are the clusters of afternoon trips; the 2nd and the 3rd are the clusters of morning trips. The 1st cluster originates from the Midtown (commercial areas) to the Upper West Side and the Upper East Side (residential areas). This cluster has afternoon trips almost three times as morning trips, which is significantly different from the expectation of the same-distribution null hypothesis. Hence, the probability of a morning trip to be in this cluster is much higher than that of an afternoon trip. The 2nd cluster, which has a slightly low likelihood, is a cluster of morning trips in the opposite direction: from the Upper West Side and the Upper East Side to the Midtown. These two clusters, together, likely depict a home-work travel pattern, that people travel to workplaces in the morning and back home in the afternoon. The 3rd cluster has a similar destination extent as the 2nd cluster, but the origin is in the south where there is a larger proportion of residential area comparing to its destination. The 4th and 5th clusters are both afternoon clusters, indicating that people are more likely to travel in the afternoon from the Midtown to the Lower Manhattan and from Lower Manhattan to Brooklyn, respectively. To evaluate how the choice of maximum cluster radius influences clustering results, the experiment

was conducted again by limiting the maximum cluster radius to 1km. The resulting top five clusters and summary statistics are shown in Figure 3.4 and Table 3.2. Among these new clusters, the 3rd and the 5th are afternoon clusters; the 1st, 2nd and 4th are morning clusters. These clusters depict morning and afternoon trip pattern differences with finer spatial details. The original cluster 1 is split into new cluster 3 and cluster 5, which depict trips from the Midtown to Upper East Side and Upper West Side respectively. New cluster 1 is part of the original cluster 2, and only contains trips from Upper West Side. Similarly, new cluster 2 and 4 are subsets of original cluster 3.

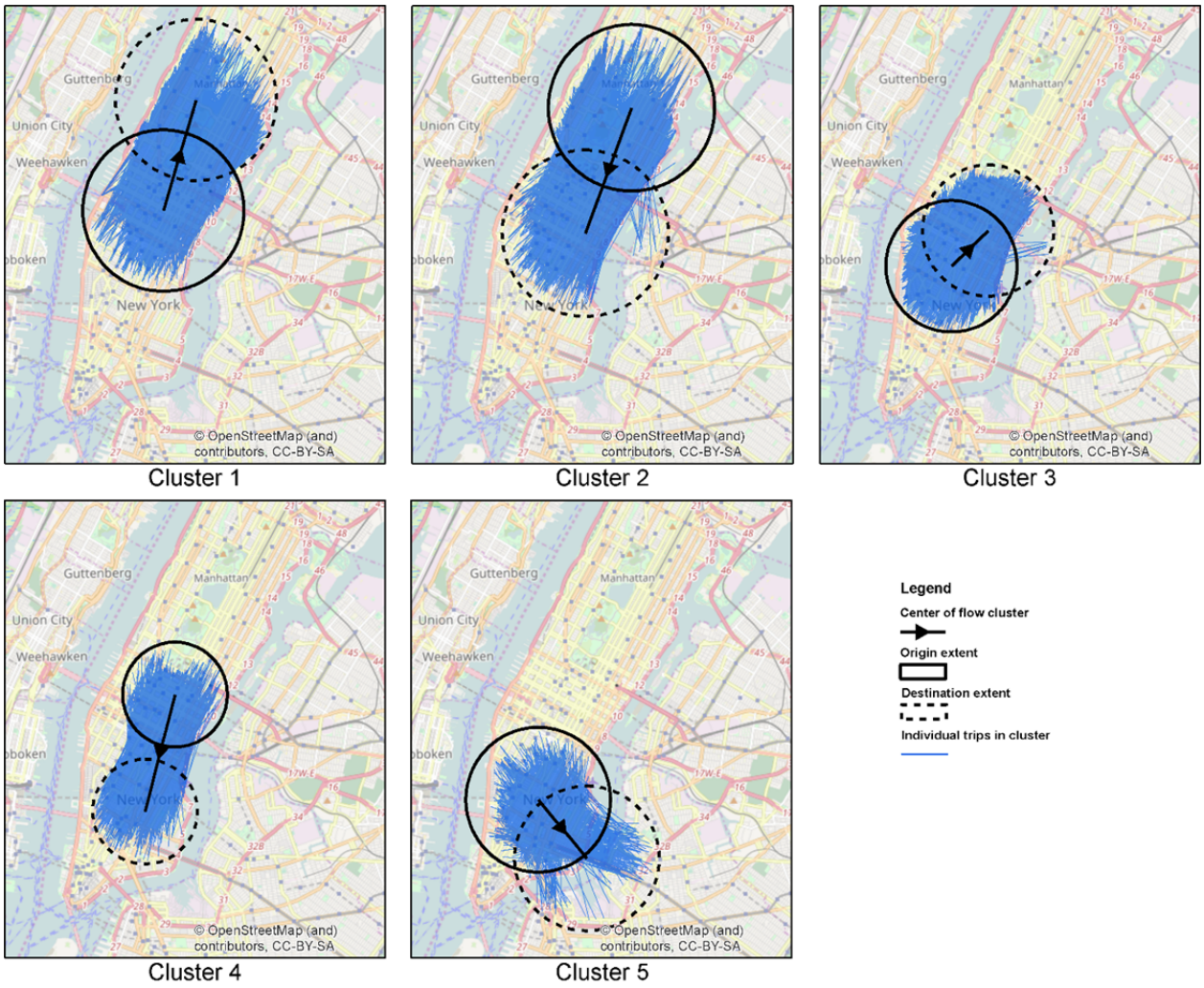

Figure 3.3: New York City taxi trips. Centers and spatial extents of top five clusters with a maximum radius of 2.5km.

The execution time depends on the number of simulations, and on average each simulation takes less than 1 minute in our testing environment.

Table 3.1: New York City taxi trips. Summary statistics of the top 5 clusters with a maximum radius of 2.5km.

Cluster ID	Type	Morning trips	Afternoon trips	Expected morning trips	Expected afternoon trips	Log-likelihood	*P*-value
1	Afternoon	2960	8672	5390.8	6241.2	-106324.2	0.001
2	Morning	8030	3912	5534.5	6407.5	-106338.8	0.001
3	Morning	11120	8112	8913.0	10319.0	-106904.6	0.001
4	Afternoon	628	2507	1452.9	1682.1	-107001.9	0.001
5	Afternoon	435	2084	1167.4	1351.6	-107004.5	0.001

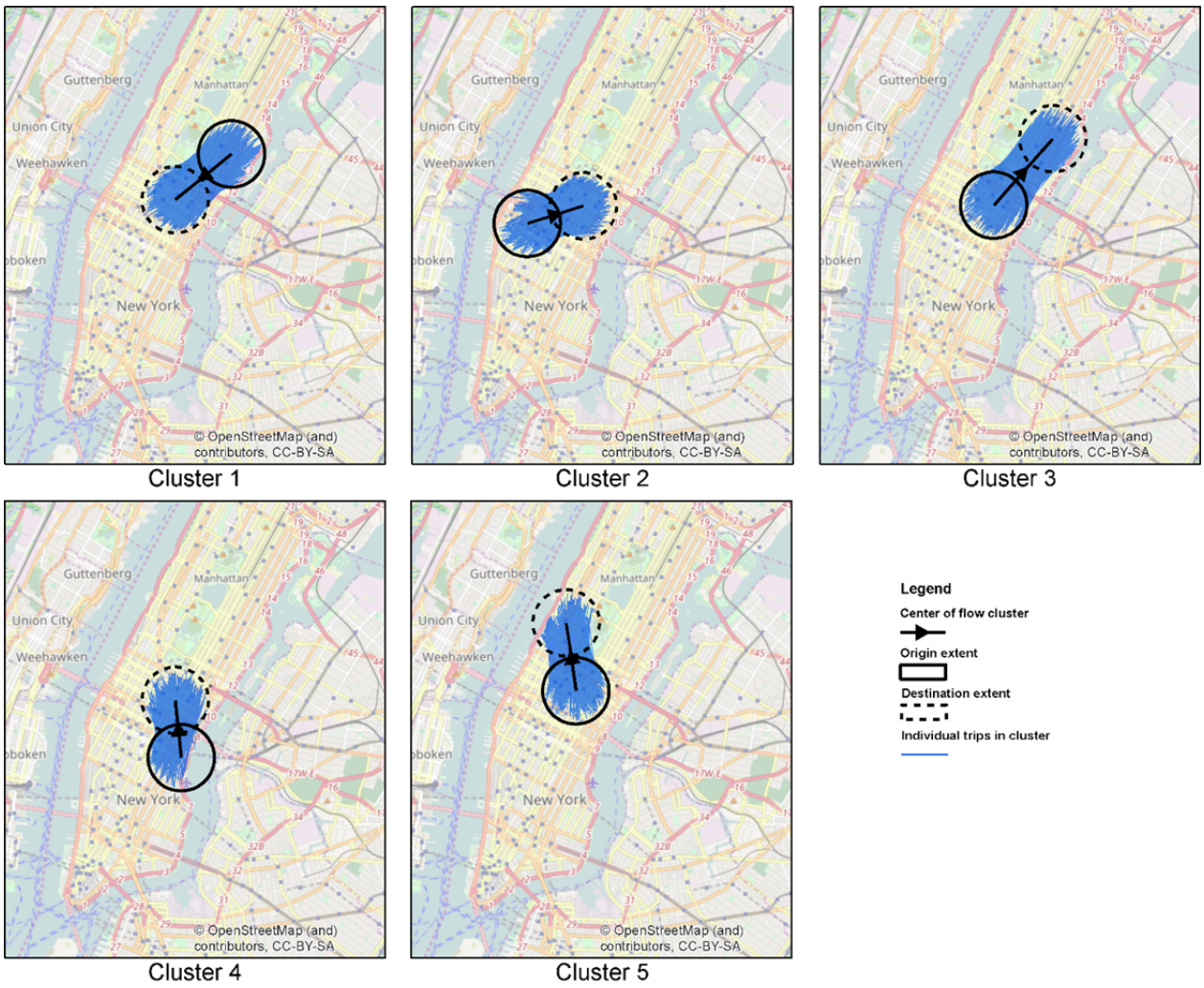

Figure 3.4: New York City taxi trips. Centers and spatial extents of top five clusters with a maximum radius of 1km.

Table 3.2: New York City taxi trips. Summary statistics of the top 5 clusters with a maximum radius of 1km.

Cluster ID	Type	Morning trips	Afternoon trips	Expected morning trips	Expected afternoon trips	Log-likelihood	*P*-value
1	Morning	2371	623	1387.6	1606.4	-106791.6	0.001
2	Morning	2144	806	1367.2	1582.8	-107055.9	0.001
3	Afternoon	270	1542	839.8	972.2	-107075.4	0.001
4	Morning	1343	349	784.2	907.8	-107091.5	0.001
5	Afternoon	179	1108	596.5	690.5	-107173.8	0.001

3.5.2. US county-to-county migration flow analysis

3.5.2.1. Data source

US county-to-county migration flow data derived from 2006-2010 American Community Survey were retrieved from the US Census Bureau[7]. This dataset contains the number of residents moving among US counties by age groups. This case study compares the differences between migration patterns of age group 25-29 (young migrants) and age group 65-69 (senior migrants). After data preprocessing, there are 33,626 separate pairs of origin-destination counties with young or senior migrants. In total, there are 1,788,892 young migrants and 175,515 senior migrants. This case study aims to find OD region pairs with a high concentration of senior migrates relative to young migrants. Our approach and software support this type of query by only detecting high clusters of senior migrants. To achieve this, all scanning windows with the count of senior migrants below expectation are eliminated in the clustering detection process.

3.5.2.2 Results

The top 10 clusters with high concentrations of senior migrants relative to young migrants are shown in Figure 3.5. The results are generated with the maximum cluster radius to be 1000km. In this figure, the origin extent (solid line) and the destination extent (dashed line) for each cluster are displaced in the same color. Arrows labeled by cluster ID represent centers of clusters. Individual county-to-county movement flows in these clusters are shown in Figure 3.6, and the summary statistics of these clusters are shown in Table 3.3. In Table 3.3, the senior migrants and the young migrants represent actual migrant counts in each cluster; the expected senior migrants and the expected young migrants represent the expected migrant counts if senior and young migrants have the same spatial distribution.

[7] Data source: http://www.census.gov/hhes/migration/data/acs/county-to-county.html

The most likely cluster (cluster 1) originates from the Northeast parts of the US to mostly Florida. This cluster has senior migrants more than three times of the expectation if senior migrants and young migrants have the same spatial distribution. This result matches well with common sense that Florida has a much higher attraction for senior than for young migrants from the Northeast. The other top clusters also depict major regions that attract senior migrants (e.g. Florida, Arizona), and regions where senior people are moving out (e.g. California). Most of these patterns meet with common knowledge. The most unexpected cluster is from Florida to Midwest (cluster 8), which describes that Midwest is more attractive to senior than young migrants from the South Atlantic.

Among the top 10 clusters, both cluster 4 and 7 have the almost identical origin and destination. They represent migrations between nearby counties and thus provide little information about inter-region migration patterns. In addition, the overlapping origin-destination also occurs in cluster 2 and cluster 9. In order to eliminate these situations, the clustering approach is executed again without allowing intersecting origin-destination. The results are shown in Figure 3.7 and Figure 3.8, and the summary statistics are shown in Table 3.4. In the ten new top clusters, six are identical to the original clusters, which are (new) cluster 1, 3, 4, 5, 6 and 7. The new cluster 2 is similar to the original cluster 2, but avoids overlapping origin-destination. A similar situation happens between new cluster 10 and original cluster 4. The original cluster 7 and 9 disappear, and two new clusters (8 and 9) originate from the west coast to nearby inland regions.

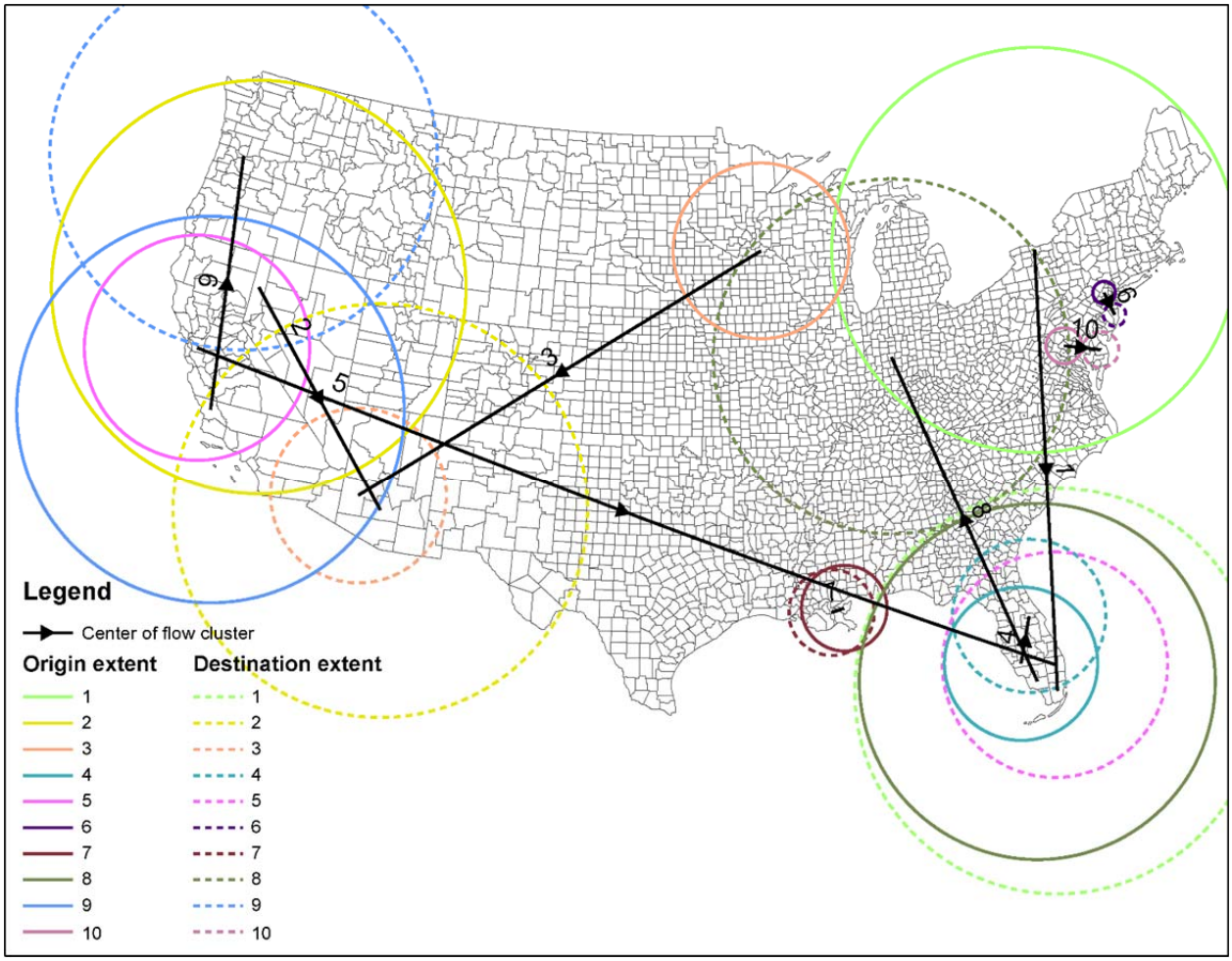

Figure 3.5: US county-to-county migration. Centers and spatial extents of top ten clusters allowing intersecting OD.

The execution time depends on the number of simulations, and on average each simulation takes less than 10 seconds in our testing environment.

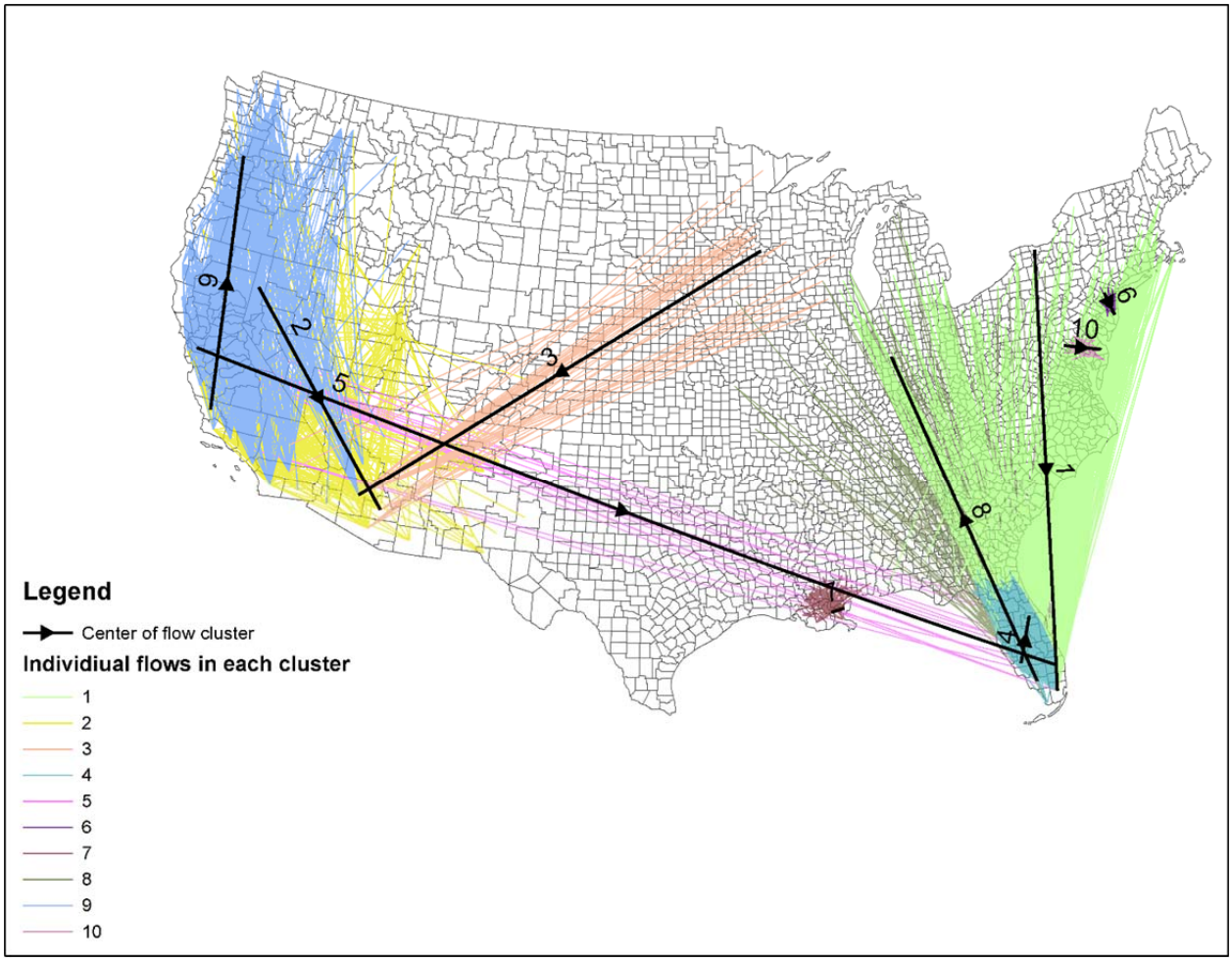

Figure 3.6: US county-to-county migration. County-to-county flows in top ten clusters allowing intersecting OD.

3.6 CONCLUSIONS AND FUTURE WORK

This paper presents a multidimensional spatial scan statistics approach to comparing the spatial distributions of OD movement datasets. It aims to measure whether two OD datasets have different spatial hotspots or their pattern differences can be explained by random chances. This approach is based on a multidimensional spatial data model that integrates each OD record into a single point with multiple spatial dimensions. A multidimensional spatial scan statistics method is developed to detect OD region pairs (point clusters) to indicate whether and where two movement distributions differ. Software for the proposed approach was developed. Case studies

demonstrated that areas with the most significant pattern differences could be detected from large movement datasets. They also showed the effects of parameters (i.e. maximum scanning windows, whether to allow intersecting origin-destination) on the clustering results.

Table 3.3: US county-to-county migration. Summary statistics of the top 10 clusters allowing intersecting OD.

Cluster ID	Senior migrants	Young migrants	Expected senior migrants	Expected young migrants	Log-likelihood	*P*-value
1	7133	14180	1904.3	19408.7	-586280.7	0.001
2	9566	56880	5936.8	60509.2	-590248.9	0.001
3	961	1130	186.8	1904.2	-590350.9	0.001
4	7446	43689	4568.8	46566.2	-590459.9	0.001
5	675	840	135.4	1379.6	-590668.4	0.001
6	742	1095	164.1	1672.9	-590680.6	0.001
7	2137	8420	943.2	9613.8	-590701.2	0.001
8	2042	8097	905.9	9233.1	-590736.5	0.001
9	5331	31749	3313.0	33767.0	-590744.6	0.001
10	373	243	55.0	561.0	-590826.3	0.001

One limitation of our approach is that 4D spherical scanning windows, which are currently used, are insufficient to accurately depict the sizes and the shapes of clusters. The origin and the destination of a cluster may not necessarily be of the same size. For instance, a cluster may represent flows from a small place to a large area. In addition, the origin and the destination of a cluster may be in shapes other than a circle. Hence, it is essential to improve the clustering approach in order to detect clusters of irregular shapes other than spheres, which is one future research direction.

Another critical future research direction is to expand our approach to include temporal dimensions. While temporal information can be treated as the indicator variable as in the first

case study to compare morning with afternoon taxi trips, the approach itself is purely spatial. It cannot automatically detect spatiotemporal clusters. Hence, it is important to include temporal dimensions in multidimensional spatiotemporal models for OD movements, and to develop spatiotemporal scan statistics approaches. There may be one (a timestamp for each OD record) or two (origin and destination with separate timestamps) temporal dimensions that need to be modeled. As a consequence, 5D or 6D spatiotemporal data models and pattern comparison approaches are needed.

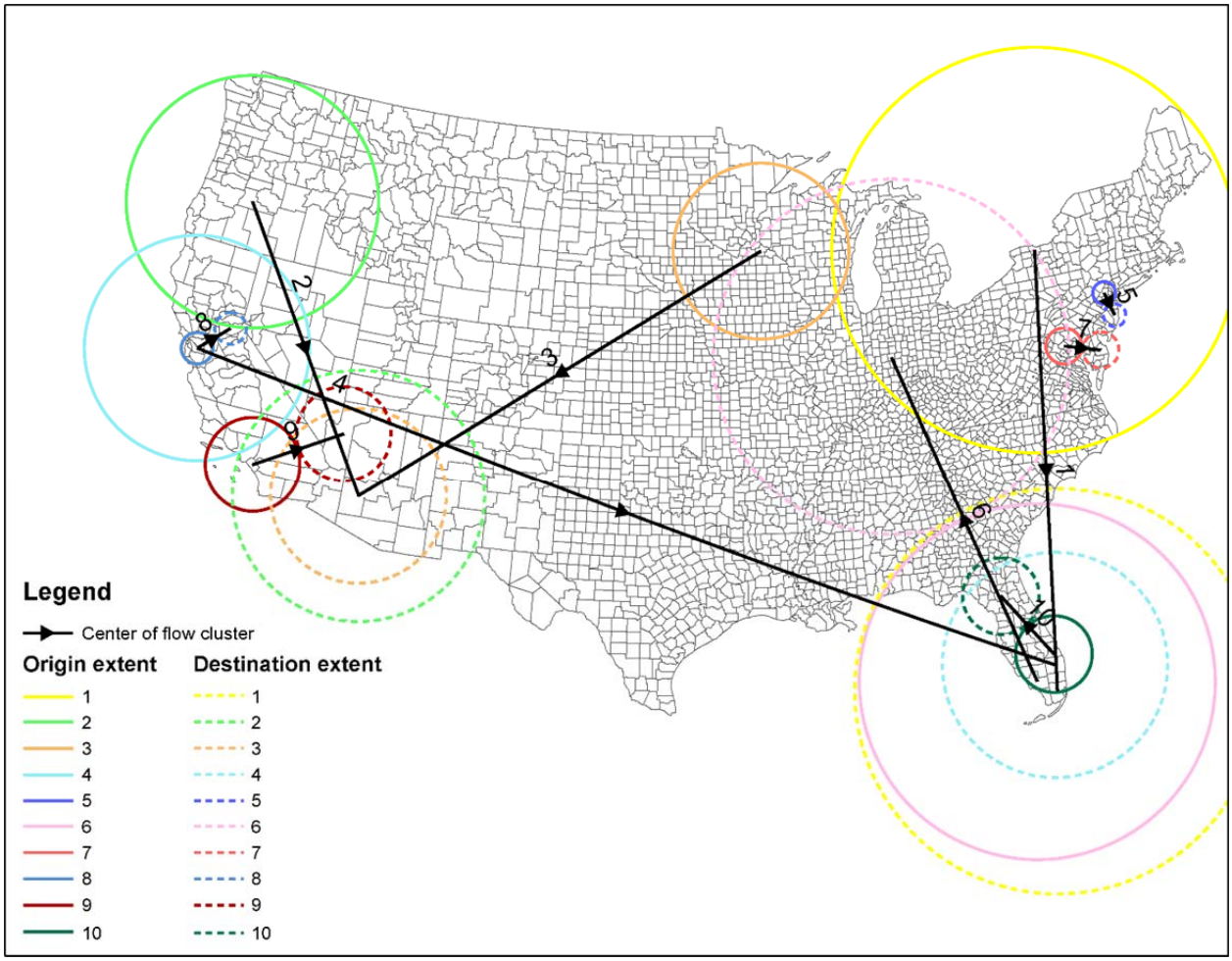

Figure 3.7: US county-to-county migration. Centers and spatial extents of top ten clusters without intersecting OD.

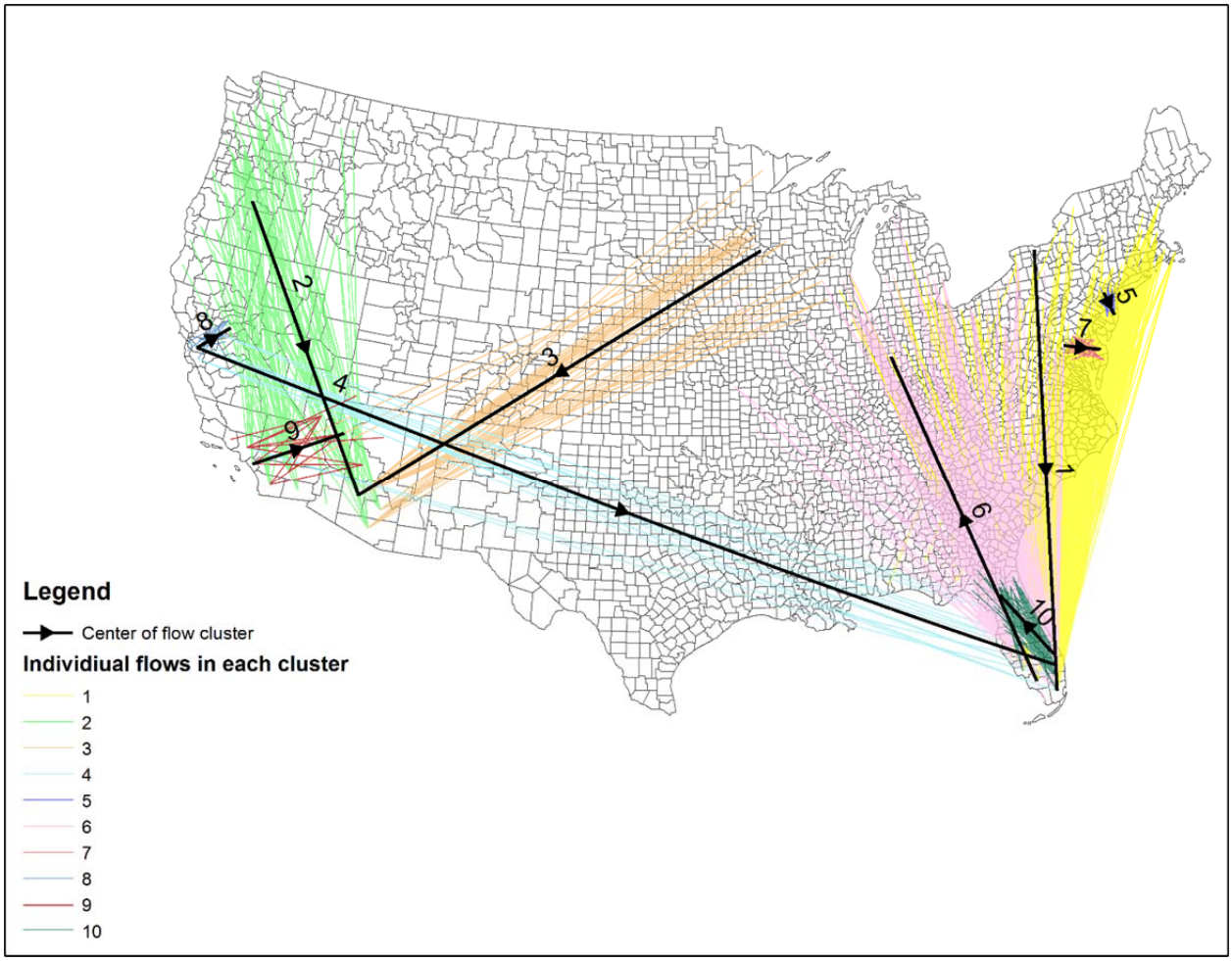

Figure 3.8: US county-to-county migration. County-to-county flows in top ten clusters without intersecting OD.

In addition to OD movement data, the approach may be extended to spatial interaction data that involve a pair of locations. Examples of such data include the follower-followee relationships between geo-located social media users, telephone records with geo-located callers and callees, and spatial patterns of journal article citations. Through the comparison of these interaction patterns, the approach may help researchers from multiple domains to gain insights into these spatial interaction phenomena.

More broadly, this paper provides an example of analyzing complex geographic phenomena by modeling them as simple objects (e.g. points) in conceptual spaces. The

conceptual spaces may have more spatial dimensions than the physical space, and thus require new analytical approaches. The conceptual space model may lead to unique findings that are hardly detectable in the physical space. For instances, it may be possible to model spatial associations (Anselin 1995) in planar space as a 4D field that represents the association between each pair of geographic features. Spatial association patterns might be better understood, by modeling not only the associations between neighboring features but also the entire 4D space that integrates all pairwise associations.

Table 3.4: US county-to-county migration. Summary statistics of the top 10 clusters without intersecting OD.

Cluster ID	Senior migrants	young migrants	Expected Senior migrants	Expected young migrants	Log-likelihood	*P*-value
1	7133	14180	1904.3	19408.7	-586280.7	0.001
2	1523	3098	412.9	4208.1	-590294.0	0.001
3	961	1130	186.8	1904.2	-590350.9	0.001
4	675	840	135.4	1379.6	-590668.4	0.001
5	742	1095	164.1	1672.9	-590680.6	0.001
6	2042	8097	905.9	9233.1	-590736.5	0.001
7	373	243	55.0	561.0	-590826.3	0.001
8	364	282	57.7	588.3	-590873.8	0.001
9	1017	3283	384.2	3915.8	-590924.5	0.001
10	581	1237	162.4	1655.6	-590956.6	0.001

CHAPTER 4: SCANNING WINDOW DESIGNS FOR MULTIDIMENSIONAL SCAN STATISTICS APPROACH TO MOVEMENT PATTERN ANALYSIS

Abstract. *Multidimensional spatial scan statistics have recently been developed to analyze spatial movement patterns and detect movement clusters based on origin-destination (OD) representations. As one of the most important aspects of scan statistics, the scanning window designs suitable for movement analysis need to be better understood. The 4D hyperspherical windows used in the previous research are not the only valid option and has clear limitations including that they cannot effectively detect clusters with an origin and a destination of different areal sizes. This chapter explores a variety of 4D scanning window designs for movement pattern analysis. From shape, location, and size aspects, the chapter investigates valid options for 4D scanning windows. For window shapes, three 4D extensions to 2D circles are identified and used, namely duocylinders with the same or different origin and destination radii, and 4D hyperspheres. In terms of scanning window centers, this chapter explores both an input record approach and a regular grid approach. Six scanning window designs are proposed by combining these window shapes and center allocation strategies. Efficient algorithms and parallel computing approaches are then developed for the six scanning window designs in order to resolve the computational challenges of movement pattern analysis. These scanning window designs are evaluated and compared both analytically and using two large real-world OD datasets: New York City taxi trips and US county-to-county migration flows. The results provide insights into how to select scanning window designs and the trade-off between computing cost and the ability to detect high-quality movement clusters.*

Keywords: CyberGIS, Movement analysis, Scan statistics, Spatial analysis and modeling, Algorithms, Geocomputation

4.1 INTRODUCTION

Understanding the spatial patterns of geographic mobility and spatial interactions is of great importance (Dodge et al. 2016, Gonzalez et al. 2008, Guo and Zhu 2014). Many research efforts have been pursued to analyze the spatial patterns of origin-destination (OD) movement flows and to detect spatial movement clusters (Berglund and Karlstrom 1999, Lu and Thrill 2008, Liu et al. 2015, Tao and Thrill 2016). Multidimensional scan statistics have been recently developed to analyze spatial movement patterns (Gao et al. 2018). It is able to detect movement clusters based on point process models and can be used to compare the spatial patterns of different movement datasets. However, it is yet not clear what scanning window designs are best-suited for OD movement pattern analysis. For instance, the 4D spherical scanning windows used in the existing research have clear limitations since they cannot effectively detect clusters with an origin and a destination of different areal sizes (Gao et al. 2018). The aim of this research is to explore a variety of scanning window designs for movement analysis.

The choice of scanning windows is of fundamental importance to any scan statistics approach (Kulldorff 1997, Kulldorff 1999, Kulldorff et al. 2006). This is because scanning windows define the search space for clusters. Specifically, scan statistics calculate the likelihood value for each scanning window for cluster detection, and detected clusters can only be chosen from these scanning windows. Since potentially there are an infinite number of scanning windows with different shapes and sizes, it is only possible to check a small number of representative ones for cluster detection. In conventional 2D spatial point analysis, the most commonly used shape of scanning windows is circle since it is the most compact shape and one of the easiest to calculate (Kulldorff 1999, Tango and Takahashi 2005, Kulldorff et al. 2006). When conducting space-time data analysis, the circular shaped scanning windows naturally

extend to cylindrical windows – 2D spatial circles each with a perpendicular line segment serving as the height of the cylinder to represent its temporal period (Kulldorff et al. 1998, Kulldorff 2001, Kulldorff et al. 2005). However, extending 2D circles into 4D space for movement analysis is more challenging. This chapter explores and compares different 4D extensions of 2D circles and identifies those that are valid for movement analysis. Specifically, this chapter analyzes 4D scanning window designs from the shape, location, and size aspects, and explores how 2D windows can be extended into 4D in all these aspects. Three 4D shapes (duocylinders with the same or different origin and destination radii, or 4D hyperspheres) and two scanning window allocation approaches (input points or 4D regular grids) are found to be valid for 4D movement analysis. The resulting six scanning window designs which combine these shapes and centering allocation approaches are proposed and evaluated.

Computational performance has long been limiting factor of spatial scan statistics (Agarwal et al. 2006, Pei et al. 2011, Li et al. 2018). Several computational challenges exist. First, a large number of scanning windows need to be tested in order to accurately depict spatial clusters. The total number of scanning windows can be orders of magnitude larger than the input data volume. Second, for each input point location and each scanning window, it needs to judge whether the location is within the scanning window or not. Third, scan statistics require Monte Carlo (MC) simulation to estimate the statistical significance of the detected clusters. As a consequence, when the data volume is large, the computing cost of spatial scan statistics increases dramatically. With a higher dimensionality, these challenges are exacerbated since a representative set of scanning windows to cover a 4D region is much larger than one to cover a 2D region. Hence, computational performance is an essential factor when evaluating scanning window designs. In this research, efficient algorithms are developed for all six scanning window

designs, and parallel computing approaches are also developed to advance cyberGIS and exploit HPC.

The proposed scanning window designs and algorithms are evaluated using both the New York taxi trip and US county-to-county migration data. The analysis of scanning windows and the experiment results demonstrated that using scanning windows with origin and destination of different areal sizes can detect better movement clusters at higher computational cost, and that using regular grids as cluster centers tremendously increases computational intensity but is only useful when analyzing aggregated movement. These results provide insights into the choice of scanning windows and the trade-off between computing intensity and the result quality of movement pattern analysis.

4.2 MULTIDIMENSIONAL SPATIAL SCAN STATISTICS AND OD DATA MODEL

Originally developed by Naus (1965a and 1965b) and popularized by Kulldorff (1997), spatial scan statistics have been applied to a wide range of research domains to detect clusters in a spatial point process Kulldorff (2015). By modeling individual movement records as spatial points in a multidimensional OD space, Gao et al. (2018) extended spatial scan statistics to analyze spatial patterns of OD movements and to evaluate the spatial structural differences between movement patterns. Detailed descriptions of the multidimensional data model and the multidimensional spatial scan statistics can be found in chapter 3.

The general procedure of the spatial scan statistics is as follows (Kulldorff 1997, Gao et al. 2018). First, a large number of scanning windows are put in the study area. In OD space, each of the scanning windows is a 4D region. Second, for each scanning window, the number of point observations in it is counted. Third, the maximum likelihood of each scanning window to be a cluster L_w, and the maximum likelihood if there is no cluster L_{w0} are calculated. The scan

statistics (likelihood ratio) λ is calculated as L_w/L_{w0}. Fourth, the scanning windows with the highest likelihood are identified as clusters. Finally, Monte Carlo simulation is conducted to estimate the *p*-value of each detected cluster for statistical inferences.

Another important aspect of spatial scan statistics is the point process model that describes the probability distribution generating events under the null hypothesis (Kulldorff 1999). Poisson process and Bernoulli process are two most popular point process models in spatial scan statistics (Kulldorff 1997). A Poisson model is often used to analyze the spatial distributions of events over a spatially varying population density. It assumes that the number of events occurring during a time interval and in a spatial region follows a Poisson distribution. A Bernoulli model is used to compare the spatial distributions of two different types of events (e.g., people with or without a disease, morning or afternoon traffic) and is popular in case-control studies. There are other point process models including space-time permutation (Kulldorff et al. 2005), ordinal (Jung et al. 2007), exponential (Huang et al. 2007), normal (Kulldorff et al. 2009) and multinomial (Jung et al. 2010). For multidimensional spatial scan statistics, Bernoulli models have been used to compare the spatial patterns of movements in Chapter 3. This chapter will follow Chapter 3 and conduct experiments using a Bernoulli model.

4.3 SCANNING WINDOW DESIGNS

For each of the three parameters of scanning window design, this chapter reviews commonly used approaches in 2D spatial scan statistics and 3D space-time scan statistics, analyzes their advantages and disadvantages, and proposes and evaluates 4D extensions that are valid for movement pattern analysis. It first explores a variety of scanning window shapes in section 4.3.1 and identifies three shapes that are appropriate for movement analysis. Two scanning window center allocation approaches are selected and explored in section 4.3.2. The

choice of scanning window sizes is discussed in 4.3.3. A total of six scanning window designs are proposed as a combination of the three shapes and the two center allocation approaches, and are summarized in section 4.3.4. These six designs are implemented and evaluated later in the remaining part of this chapter. Other possible alternative designs are discussed in section 4.3.5.

4.3.1 Scanning Window Shapes

4.3.1.1 Scanning windows in lower dimensions

Before designing scanning window shapes in 4D space, it is necessary to first understand 2D spatial scanning windows and 3D space-time scanning windows from a geometric point of view. It would also be necessary to understand the concept of the ball in mathematics, which is the space bounded by a sphere. An *n*-ball means a ball in *n* dimensions. For instance, a 1-ball is a line segment, a 2-ball is a circle, a 3-ball is a sphere, and a 4-ball is a 4D hypersphere.

For 2D scanning window shapes, the most popular one, circle, is a 2-ball. Some other studies such as Naus 1965b, Loader 1991, and Chen and Glaz 1996, used rectangular 2D scanning windows. A rectangle is the Cartesian product (×) of two line segments and thus 1-ball × 1-ball. There are no other shapes in 2D that are the results of the Cartesian product of balls and that can be considered as intermediate between a rectangle and a circle. A circle is usually a preferred choice of scanning windows in spatial point analysis since circular windows are more compact, more visually appealing, and easier to define (Kulldorff et al. 2005). For rectangular scanning windows, if the edges are constrained to be perpendicular to axes as in Naus 1965b, the clustering results will be arbitral defined by the axes' directions; if, on the other hand, the edges are allowed in any directions, it would be difficult to both define and calculate these scanning windows.

In 3D space-time analysis, there are potentially three shapes that are the results of the Cartesian product of balls: cuboid, cylinder, and sphere. A cuboid is the Cartesian product of three line segments (1-ball × 1-ball × 1-ball), a sphere is a 3-ball, and a cylinder is the only intermediate between these two, which is the Cartesian product of a circle and a line segment (2-ball × 1-ball). Cuboid scanning windows are not popular in space-time scan statistics for the same reasons why rectangular windows are not popular in 2D spatial scan statistics. To the best of our knowledge, there are no studies using cuboid scanning windows. Spherical scanning windows are also not appropriate for space-time analysis since temporal dimension is usually considered to be different from spatial dimensions with different linear units, and thus it is difficult to define a space-time sphere. Hence among the three shapes, only cylindrical scanning windows are valid options for space-time scan statistics and are widely used in existing research (Kulldorff et al. 2005, Kulldorff 2015).

4.3.1.2 4D scanning window shapes

In 4D space, five shapes are the results of the Cartesian product of balls. In addition to Tesseract (4D cuboid) and 4D hypersphere, three shapes exit between these two, namely cubinder, duocylinder, and spherinder. The information of all these five shapes is shown below:

- Tesseract: 1-ball×1-ball×1-ball×1-ball, the Cartesian product of four line segment;
- Cubinder: 2-ball×1-ball×1-ball, the Cartesian product of one circle and two line segment;
- Duocylinder: 2-ball×2-ball, the Cartesian product of two circles;
- Spherinder: 3-ball×1-ball, the Cartesian product of one 3D sphere and one line segment; and
- 4D hypersphere: 4-ball, a sphere in 4D.

Not all five scanning window shapes are appropriate for movement pattern analysis. First, a scanning window should not include one origin dimension into one ball with destination dimension(s) but leave the other origin dimension separately. Hence, the spherinder is not a valid option. Second, the scanning window should treat the 2 origin's dimensions and the 2 destination's dimensions symmetrically. It is thus not appropriate to use squares for origin dimensions and circles for destination dimensions. As a consequence, cubinder can also be ruled out. Finally, due to the same reason that rectangular windows are not used in spatial analysis and cuboid windows are not used in space-time analysis, the tesseract window solution is also less preferred. Hence, only duocylinders (with one circle in the two origin dimensions and another circle in the two destination dimensions) and 4D hyperspheres remain valid for 4D scanning window options, which will be used in this chapter.

4D hyperspherical scanning windows represent a straightforward approach to evaluate the distance or the similarity between two OD movements – the distance between two OD movements is measured by their 4D Euclidian distance that incorporated both their origins' distance and their destinations' distance. A 4D hyperspherical scanning window can be uniquely defined by a 4D point $< x_{co}, y_{co}, x_{cd}, y_{cd} >$ serving as the window center, and window radius. With 4D hyperspherical scanning windows, an OD movement is considered to be in a scanning window if the Euclidean distance between the movement and the scanning window center is smaller than the scanning window radius. This scanning window shape is used by Gao et al. (2018). This scanning window shape will be referred to Hypersphere for short in the remaining part of this chapter. The major limitation of the hypersphere window is that it cannot be used to detect a movement cluster with an origin and a destination of different areal size. However, in movement pattern analysis, a movement cluster may need to have an origin and a destination

with different radii in order to detect more realistic and meaningful movement clusters, i.e., from a small transportation hub to a larger and sparser residential area. Hence, hypersphere windows may be incompetent to handle these situations.

A duocylindrical scanning window represents a 4D region that is formed by the Cartesian product of a 2D circular origin region and a 2D circular destination region. It can be uniquely defined by a 4D window center, an origin radius r_o, and a destination radius r_d. An OD movement is considered to be in a duocylindrical scanning window when its origin is in the window's origin circular (2D) region and its destination is in the window's (2D) destination region. Hence, the distances of the origins and the destinations between an OD movement and the cluster center need to be calculated separately and compared with the origin and the destination radii respectively. In this chapter, two kinds of duocylindrical scanning window shapes are explored based on whether r_o can be different from r_d. The first scanning window design requires $r_o = r_d$, and thus only one radius is necessary to define a scanning window. This scanning window shape is referred to as Duocylinder-Same for short. The second design, on the other hand, does not have such a requirement. This design is referred to as Duocylinder-Diff for short.

Duocylinder-Same windows can be considered as a special case and a subset of Duocylinder-Diff windows. For instance, when there are three designated scanning window radii 1km, 2km and 3km, Duocylinder-Diff can have 9 different sizes whose <origin, destination> radius pairs are <1km, 1km>, <1km, 2km>, <1km, 3km>, <2km, 1km>, <2km, 2km>, <2km, 3km>, <3km, 1km>, <3km, 2km> and <3km, 3km> respectively. Among the 9 scanning window sizes, only three windows (i.e. <1km, 1km>, <2km, 2km> and <3km, 3km>) are applicable when using Duocylinder-Same windows. As a consequence, while movement clusters with an origin

and a destination with different radii can be detected with Duocylinder-Diff windows but cannot with Duocylinder-Same, Duocylinder-Diff requires checking a much larger set of scanning windows and thus having significantly higher computational cost. An illustration of these two scanning window shapes is shown in Figure 4.1, where three different scanning windows are displayed. All three scanning windows are duocylindrical, each of which is formed by the Cartesian product of a circular origin region (shown at the bottom) and a circular destination region (shown on the left). All three scanning windows belong to Duocylinder-Diff, but only the red one in the middle belongs to Duocylinder-Same.

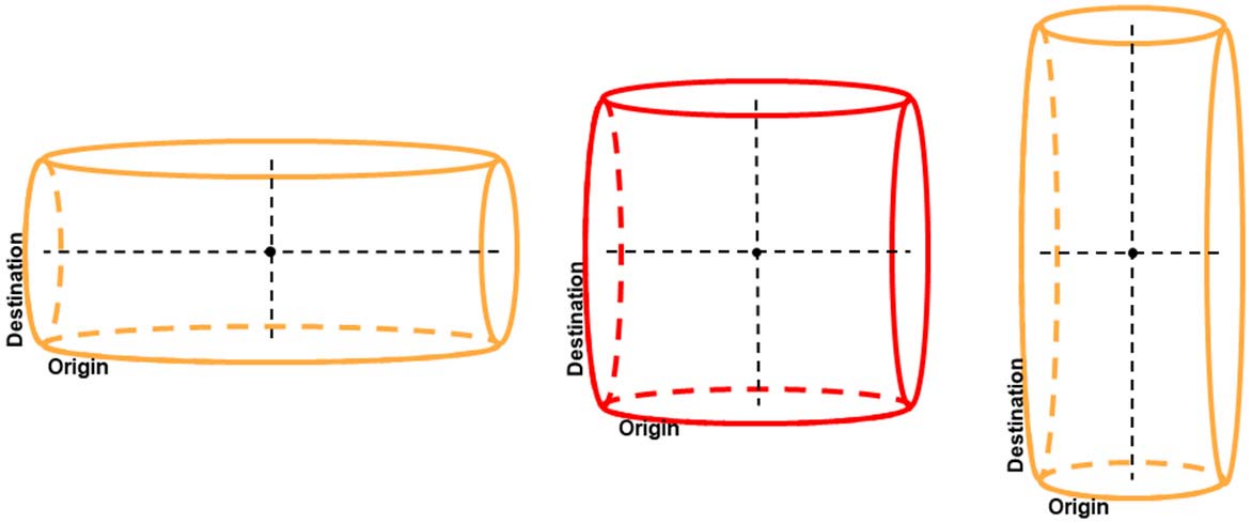

Figure 4.1: Illustration of Duocylinder-Same and Duocylinder-Diff.

Hypersphere and Duocylinder-Same windows have some similarities since they both can only detect movement clusters with an origin and a destination of the same size. However, clear differences exist between these two scanning window designs. The differences between them are illustrated in Figure 4.2. For a simplified explanation, a 2D analogy is first used (Figure 4.2 left). If we consider both the origin and the destination as lines (1D), a Hypersphere (now 2D) window is a circle (in blue) and a Duocylinder-Same window is a square (in red). The circle is the inscribed circle of the square. The orange dot is an observation that is inside the square but

outside the circle. The real shapes of these two type of scanning windows can be understood by expanding the origin and the destination from 1D line segments in Figure 4.2 left to 2D circles in Figure 4.2 right. The blue circle becomes a 4D hypersphere, and the red square becomes a duocylinder that contains the 4D sphere. The hypersphere is still inscribed in the duocylinder. The orange observation is inside the Duocylinder-Same scanning window but not the Hypersphere one. In terms of the 4D volume, suppose both scanning windows have a radius r, the Hypersphere has a volume of $\pi^2 r^4/2$, and the Duocylinder-Same has a volume of $\pi^2 r^4$, which is twice as that of the Hypersphere. The aforementioned shape differences also showed the design principle differences between these two scanning windows. A Duocylinder-Same window simply designates an origin area and a destination area, and all the movements from the origin area to the destination area belong to this scanning window. On the other hand, a Hypersphere window cares more about the similarity between an OD movement and the window center and would exclude movements whose both origin and destination are far away from the window center.

4.3.2 Scanning Window Center Locations

Scanning window centers define where scanning window of shapes mentioned in section 4.3.1 are placed in the study area. In 2D spatial scan statistics, the most commonly used center allocation approach for circular and elliptical scanning windows is using the locations of input point records. This strategy is also the default choice of SaTScan™ (Kulldorff 2015). With this approach, if the data being analyzed contain records at *n* unique locations, scanning windows with the designated shapes will be placed at all these *n* locations. The second approach is to use a list of user-specified grid points as scanning window centers. This approach is also supported by SaTScan™ and its user guide suggests the usage of grid files to limit the number of center

locations to reduce computing time if necessary (Kulldorff 2015). For instance, in Kulldorff and Nagarwalla 1995, the centroids of census tracts and census groups are used as window centers. One approach that is seldomly mentioned in the literature is to use a regular grid as scanning window centers. This approach uses equally spaced 2D regular grid points that are aligned into rows and columns covering the entire study area as center locations. This approach has a clear advantage that it checks for potential clusters at any location, and is able to identify clusters without necessarily an input record at its very center. Figure 3.4 demonstrated the input record approach (left) and the regular grid points approach (right) for 2D spatial analysis. The input point records are shown as black dots, and the circular scanning windows are shown in blue and red circles respectively.

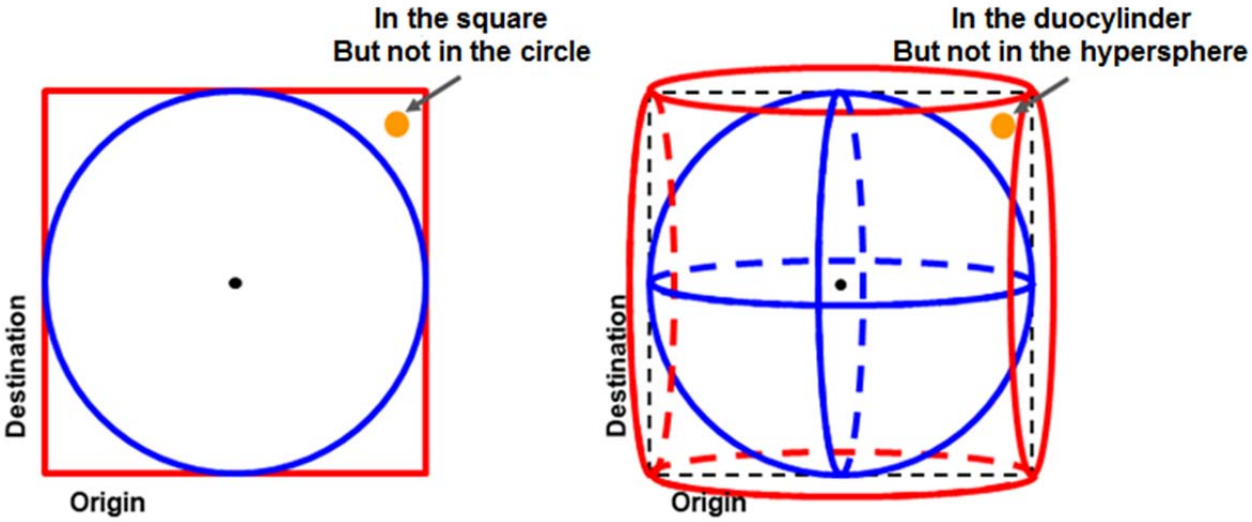

Figure 4.2: Comparison between Hypersphere and Duocylinder-Same.

In this chapter, we explore and compare both the input record approach and the regular grid approach in 4D space for the multidimensional scan statistics. When using input records as window centers, 4D scanning windows are centered at each input OD movement record that is represented as a 4D point, and hence the total number of scanning window centers equals to the total number of unique OD movement records. When the regular grid approach is used, a 2D grid

with n_{row} rows and n_{col} columns is first used to cover the (2D) study area. Then, the Cartesian product of this 2D grid with itself is used as the 4D grid for window centers: every 4D scanning window center can be formed by selecting a 2D grid point as the origin center and another 2D grid point as the destination center to get a 4D center point. As a result, using this approach, the total number of scanning window centers is $(n_{row}n_{col})^2$.

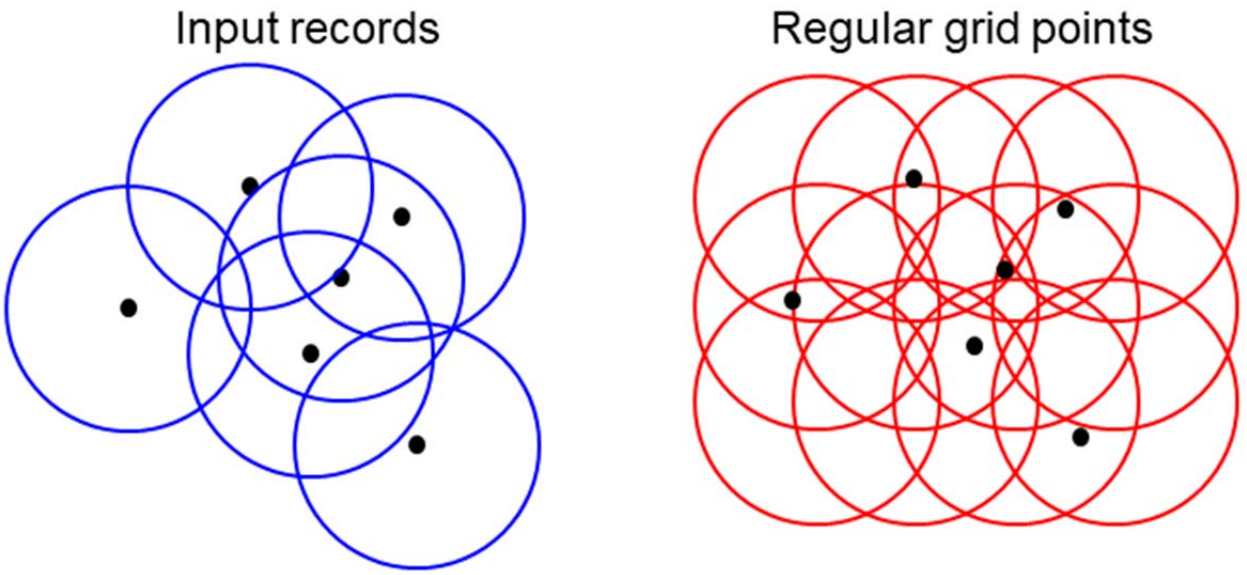

Figure 4.3: 2D circular scanning window center allocations

The major advantage of using regular grid points over input records is that it allows the possibility to detect clusters without an actual movement at its center, which can be preferred in some real-word scenarios. For instance, an airport may have two taxi pick-up areas on its two ends, and the center of these two areas is inside the airport not accessible by taxis. Hence, the best cluster to describe taxi trips from the airport to another place (e.g. the downtown) may not have a trip at the cluster center. This strategy is also beneficial in regions with relatively low point density since it increases the potential locations of clusters in these regions. Furthermore, this strategy is especially useful when analyzing aggregated movement datasets, since only using the aggregated OD pairs as window centers may result in too small a search space for clusters.

An illustrative example of such an issue of aggregated datasets in 2D space is shown in Figure 4.4, where observations are aggregated into four regions. Each region is represented by its centroid. If input records are used as scanning window centers, it is impossible to have a circular scanning window that only includes region B and C. However, such a scanning window, for instance, the dashed circle, may be possible when a regular grid is used as window centers.

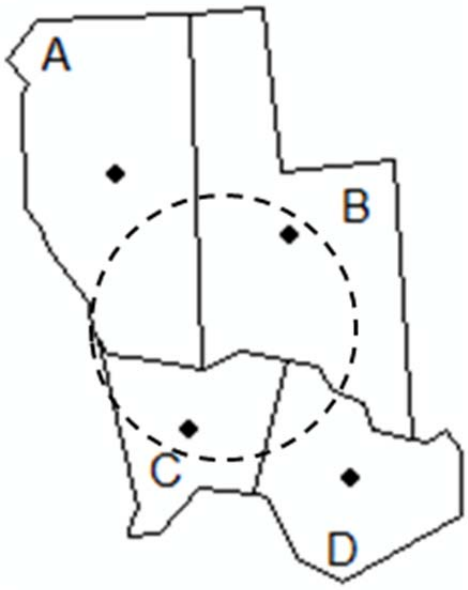

Figure 4.4: An illustrative example of an aggregated dataset.

There are two major disadvantages of using regular grid points, though. One disadvantage of using regular grid points is its computational intensity due to the increased number of scanning windows. In multidimensional spaces, the number of grid points can be much larger than the total number of input observations. For instance, if a 100 rows by 100 columns grid is used, which is at quite a coarse scale, there will be a hundred million window centers in 4D space. Since the volume of input records is not likely to be at this hundred-million scale, the total number of scanning windows will be larger when using the regular grid approach. Another disadvantage is that the regular grid approach evenly places window centers in the study area, and thus cannot adjust for point density differences. In contrast, the input observation

approach self-selects more scanning windows in high-density areas where a small change to a scanning window's location has a greater impact and fewer windows in low-density areas where a scanning window is less sensitive to small changes of its location.

4.3.3 Scanning Window Sizes

The choice of scanning window sizes defines how many scanning windows are put at each centering location and how large are these windows. In the context of 4D hyperspherical and duocylindrical scanning window, it sets a list of window radii (combination of origin radius and destination radius for the Duocylinder-Diff case). In the early years of spatial scan statistics, it is a common practice to use only one fixed size (e.g. Naus 1965b, Turnbull et al. 1989). While such an approach is easy to execute, it cannot be used to efficiently identify the spatial extents of clusters and may miss important clusters if the chosen size is different from the true cluster size.

As a result, modern scan statistics approaches unusually uses scanning window with various sizes. At present, the most commonly used approach for window sizes is to use the distance between scanning window centers to every input records, which is used in SaTScan™. Using this approach, for each scanning window, its distance to all input records are calculated, and all these distances will serve as scanning window radii. The rationale for this approach is that if we put a circular scanning window at a location and changes its radius from 0 to infinity, the point count in (and thus likelihood of) this cluster will only be changed every time a new observations is encountered (Kulldorff 1999). As a result, a new scanning window is needed every time there is a new observation. In practice, an upper limit is usually set based on either maximum radius or the population inside it in order to eliminate clusters that are too large or to control the scale of analysis (Kulidorff and Nagarwalla 1995, Kulldorff 2015). While this approach is effective when analyzing small datasets, it is highly inefficient when the data volume

is large. This is because the number of scanning windows at each center is proportional to the total number of inputs, and when the data volume is large, many of these scanning windows are very similar to each other and is redundant to check all of them. This approach is also difficult to be extended to 4D spaces specifically when using Duocylinder-Diff scanning windows since the combination of all origin distances and all destination distances needs to be used and hence the number of scanning windows at each center is proportional to the square of input count. The total number of scanning windows thus will be too large to handle.

Hence in this chapter, a list of numbers with an equal interval is used as scanning window radii for all our scanning window designs. An example of this with 1km interval and 100 different radii is 1km, 2km, …, 100km. The maximum cluster radius specified as the last number in the list can be used to control the scale of analysis, which is demonstrated in Gao et al. 2018. For Hypersphere windows, each number in the list will be used as the radius for a 4D hypersphere at each center. For Duocylinder-Same windows, each number in the list will be used as both the origin circle radius and the destination circle radius for a duocylindrical window at each center. Thus, for the above two cases, the total number of scanning windows at each center location equals to the number of designated search radii n_R. When using Duocylinder-Diff windows, one radius from the list needs to be picked as the origin radius and another as the destination radius. Hence there are n_R^2 different scanning windows at each center. This list-with-equal-interval approach has a clear computational advantage over the distance to input approach since the size of the designated radii list is usually much smaller than the input count. In addition, this list-with-equal-interval approach provides a nice sampling strategy that covers scanning windows with different sizes, and avoids redundant calculations if many observations are at almost the same distance from a window center. Furthermore, it provides more options for

users to control the computational cost by changing the interval size. For instance, using 2km, 4km, …, 100km instead of 1km, 2km, …, 100km will reduce the total number of scanning windows by half.

4.3.4 Six Scanning Window Designs

In section 4.3.1, three 4D scanning window shapes are identified to be valid for movement pattern analysis, which are Hypersphere, Duocylinder-Same and Duocylinder-Diff. In section 4.3.2, two scanning window center allocation approaches are proposed, which uses input records (input for short) and 4D regular grid points (grid for short). Hence, this research explores and compares six scanning window designs that are the combination of the three shape types and two centering allocation approaches. For easy reference, these six designs will be called <shape>- for short. For instance, Duocylinder-Same-Input refers to the scanning design uses Duocylinder-Same as the shape and input records as centers. All six designs use the list-with-equal-interval approach for window sizes for the reasons discussed in section 4.3.3. The algorithms for these designs and computational intensity of these designs will be discussed in section 4.4.

4.3.5 Alternatives

There are some less popular alternative approaches for scanning window design mostly for 2D space that are used by existing literature. This section is dedicated to describe these alternative approaches and discuss whether they could be extended for movement pattern analysis.

There are some other approaches to define circular 2D scanning windows other than the center-and-radius approach as mentioned previously. Anderson and Titterington (1997) proposed a circular scanning window design with a fixed size *d*. In their approach, for any pair of input

points, two circular windows with radius *d* are used such that both points are on the circles' circumference. Kulldorff (1999) extends this method to variable sizes by using any three input points to define a circular window such that all three points are on the circle's circumference. It could also be possible to define variable-sized circles by using any two input points as the diameter. These kinds of scanning window designs, however, only work with very small datasets since the total number of scanning windows is the square or cube of input data size. Furthermore, while three points can uniquely define a 2D circle, five points are necessary to uniquely define a 4D hypersphere. As a consequence, the number of 4D windows using similar approach is the fifth power of input size, which is not computationally feasible.

In addition to circular windows, elliptical scanning windows have been developed in order to detect clusters with more flexible shapes (Kulldorff et al. 2006). An elliptical scanning window can be defined by a center (x,y), a major axis length, a minor axis length (or the eccentricity), and its major axis' direction. As a result of more parameters to define a scanning window, the total number of scanning windows is much larger than when using circles. Furthermore, checking whether a point is in an ellipse, which involves trigonometric calculations, is more complicated and slower than checking points in circles. As a result, elliptical scanning windows have a higher computational requirement than circular ones. Nevertheless, elliptical scanning windows have the potentials to be extended into 4D spaces for movement pattern analysis. For instance, it might be possible to extend 4D hyperspheres into 4D ellipsoid in order to detect movement clusters with an origin and a destination of different areal sizes. It would also be possible to extend duocylindrical windows such that its origin region and destination region are ellipses in order to detect more flexibility shaped movement clusters.

These extensions require much more efforts and are beyond the scope of this chapter, but are valid future work directions.

Finally, flexible shaped scan statistics have been developed since neither circle nor ellipse can capture spatial clusters of sufficiently complexed shapes. Tango and Takahashi 2005 is one of the most popular flexible spatial scan statistics, with a standalone software, FlexScan (Takahashi et al. 2010). FlexScan combines nearby regions to form irregularly shaped clusters, which has high computational requirement. Many approaches have been developed for irregularly shaped cluster detection, including simulated annealing (Duczmal and Assuncao 2004), genetic optimization algorithm (Duczmal et al. 2007) and colony optimization (Pei et al. 2011). Furthermore, flexibly shaped space-time scan statistics have also been developed (Takahashi et al. 2008). The major problem of irregularly shaped movement clusters is that they are too difficult to interpret. An irregularly shaped scanning window expanding in the 4D space may not be easily interpreted as a collection of movement from an origin region to a destination region, and thus lacks necessary geographic meanings. Furthermore, these methods can only handle small datasets and assume that the data are aggregated into few regions. As a result, they cannot be applied to large individual-level data. They are highly computationally expensive even in 2D space, and hence are almost impossible to be applied to 4D space at present.

4.4 ALGORITHM AND COMPUTATION

4.4.1 Algorithms

Multidimensional scan statistics algorithm consists of four major phases. The first phase is to count the number of observations in each scanning window. It is more efficient to check all scanning windows with the same center simultaneously in order to reduce the computing cost for

distance calculations. The algorithm in this phase starts by checking each scanning window center and input record pair and updating the observation counts of all concentric scanning windows at the center location accordingly. Specifically, the algorithm loops through all designated scanning window center. For each center, it calculates its distance to every input observation. Based on the distance of each input observation, the event counts of all scanning windows at this center that the observation falls in are increased by one. The second phase calculates the likelihood of each scanning window to be a cluster, based on the observation counts from the first phase.

The third phase detects the locations and the sizes of spatial clusters as the scanning windows with the highest likelihood. Both the likelihood values and the geometries of scanning windows need to be considered in this phase. This phase starts by first identifying the primary cluster as the scanning window within the highest likelihood value. When detecting secondary clusters, it is necessary to ensure new clusters have no intersections with existing clusters since clusters intersecting with existing ones provide little information about the underlying spatial patterns (Kulldorff 2015, Gao et al. 2018). Hence, once a cluster is detected, the algorithms need to check all scanning windows and exclude the ones that intersect with the newly detected cluster. The algorithm iteratively goes through the detection-elimination procedure until a designated number of spatial clusters are identified.

The final phase of the algorithm estimates the *p*-value of each cluster detected in the third phase for statistical inferences through MC simulation. In each iteration of the MC simulation, a random dataset is first simulated under the null hypothesis of no spatial clustering. The clustering detection procedure described in phase one to three is applied to the simulated dataset, with the only difference that only the primary cluster needs to be detected in each replication. Finally, the

p-value of each original spatial cluster is estimated by comparing its likelihood with the primary clusters' likelihoods from all simulated datasets.

The algorithms from six scanning window designs mainly differ in the first phase and the intersection elimination in the third phase. In the first stage, for Hypersphere windows, one 4D Euclidean distance calculation is necessary for each window center and observation pair; for both duocylindrical window shapes, two 2D Euclidean distances need to be calculated for origins and destinations respectively. Once the distances are calculated, different procedures are used to update the observation counts. For Hypersphere windows, the 4D Euclidean distance between an observation and the window center is compared with the spherical radii. After calculating a 4D center-to-observation distance, the observation counts of all scanning windows at the that center with a larger radius are increased by one. For instance, if scanning window radii range from 10m to 80m with an increment of 10m and an observation is 37m from a window center, this observation will contribute to the scanning windows with radii 40m, 50m, 60m, 70m and 80m.

For Duocylinder-Diff windows, the 2D distance between an observation's origin and the window center's origin is compared with the origin radii of clusters, and the 2D distance between destinations is compared with the destination radii. Using Figure 4.5 as an example, if scanning window radii range from 10m to 80m with an increment of 10m, there are 64 different scanning windows (represented as 64 squares in Figure 4.5) centered at each location. For an observation and a window center, if the observation's origin is 32m from the center's origin and the observation's destination is 56m from the center's destination, the observation will contribute to all the 15 scanning windows marked with "+1" in Figure 4.5 left. Duocylinder-Same windows only consist of a subset of Duocylinder-Diff windows. Only the 8 scanning windows (8 diagonal elements as shown in Figure 4.5 right) are valid for Duocylinder-Same, and hence only the 8

scanning windows need to be stored. Given the same observation and window center, the observation only contributes to 3 scanning windows marked with "+1" in Figure 4.5 right.

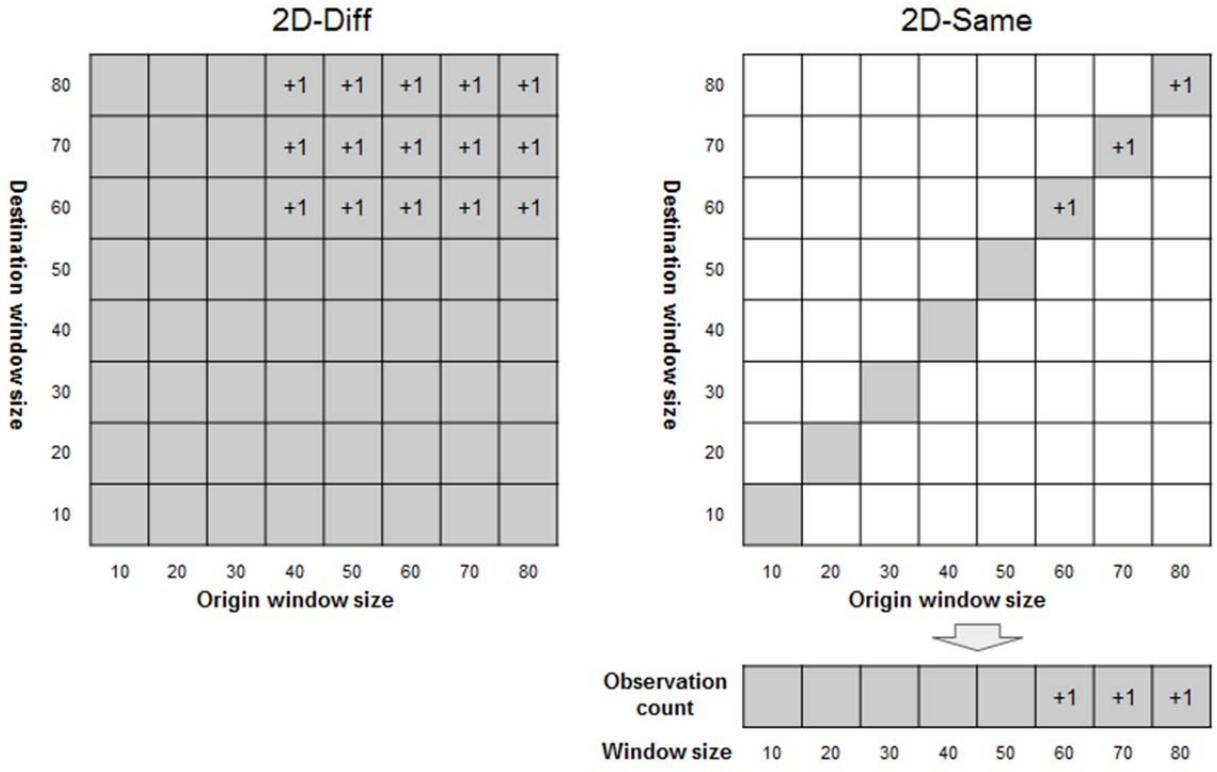

Figure 4.5: Updating Observation Counts for Duocylinder-Diff and Duocylinder-Same Windows.

If a regular grid is used as scanning window centers, the computational efficiency could potentially benefit from a change of processing order in the first stage. Rather than starting with each scanning window center and checking whether each observation is in scanning windows located there, it may be more efficient to start with each observation and identify all scanning windows that this observation is in. This is because it is more efficient to find grid points near an observation by calculating a grid index range than to find observations near a grid point. However, such efficiency improvement is limited since the maximum scanning window size is comparable with the extent of the study area and as a consequence, each observation may contribute to scanning windows centered at a large percent of all grid points. Furthermore, while

this approach may increase the efficiency of the sequential algorithm, it is not preferred when the algorithm is parallelized since multiple subtasks handling different observations might need to update the observation counts for the same scanning window simultaneously and hence cause concurrency issues. Resolving the concurrency issues would require either buffering space or communication between subtasks and thus introduces additional cost.

4.4.2 Computational Intensity

The four most computationally intensive components in aforementioned multidimensional scan statistics algorithm are the distance calculation between window centers and input points, counting input points in each window, likelihood calculation of each window, and cluster detection. The final computational intensity is a sum of these four components and depends on both the analysis requirements (data and parameters) and the computing environment (e.g., the speed of logarithm calculations). The computational intensities of the four components are shown in Table 4.1, Table 4.2, Table 4.3 and Table 4.4 respectively. Each row of these tables represents a scanning window shape and each column represents one window center allocation strategies. For instance, the value of row "Hypersphere" and column "Input" represents the computational intensity of the algorithm for Hypersphere-Input windows. In these tables, n_P is the number of input point observations, n_R is the number of different 4D sphere or 2D circle radii for scanning windows, n_{row} and n_{col} are the number of rows and columns of the regular grid, n_{MC} is the number of replications in MC simulation and n_{Cl} is number of clusters to detect.

Table 4.1 shows the number of distance calculation that is necessary for each of the scanning window design. For instance, when using Hypersphere-Grid windows, a total of $n_P(n_{row}n_{col})^2 n_{MC}$ Euclidian distance calculations are necessary. Table 4.2 provides the upper bounds for the number of count updates when counting input points in windows. The actual

computational intensity will depend on the distribution of input data and the scanning window size, and the actual intensity will be smaller but comparable with the number in Table 4.2. While numbers in Table 4.2 has the largest value, it is not the most computational intensive component since it only requires trivial count updates, and no mathematical operations (such as square root or logarithm) are involved. Table 4.3 shows the number of total log-likelihood calculations, which equals to the number of scanning windows times the number of MC simulations. The calculation of a single log-likelihood value is the slowest among the four components. Table 4.4 shows the number of log-likelihood checks for the final cluster detection. The $(2n_{Cl}+n_{MC})$ term in the computational intensity of cluster detection results from the sum of $2n_{Cl}$ in regular cluster detection and n_{MC} for MC simulation where only the top cluster needs to be detected in each replication.

Table 4.1: Computational intensity of distance calculation

	Input	Grid
Hypersphere	$n_P^2 n_{MC}$	$n_P(n_{row}n_{col})^2 n_{MC}$
Duocylinder-Same	$2n_P^2 n_{MC}$	$2n_P(n_{row}n_{col})^2 n_{MC}$
Duocylinder-Diff	$2n_P^2 n_{MC}$	$2n_P(n_{row}n_{col})^2 n_{MC}$

Table 4.2: Computational intensity of counting points in windows

	Input	Grid
Hypersphere	$n_P^2 n_R n_{MC}$	$n_P(n_{row}n_{col})^2 n_R n_{MC}$
Duocylinder-Same	$n_P^2 n_R n_{MC}$	$n_P(n_{row}n_{col})^2 n_R n_{MC}$
Duocylinder-Diff	$n_P^2 n_R^2 n_{MC}$	$n_P(n_{row}n_{col})^2 n_R^2 n_{MC}$

As demonstrated in section 4.3.2, $(n_{row}n_{col})^2$ is usually much larger than n_P given a realistic dataset. Hence, using a regular grid as window centers is usually more computationally

intensive. In all the four components, the algorithm using Hypersphere-Input windows have the lowest computational intensity, while Duocylinder-Diff-Grid windows have the highest.

Table 4.3: Computational intensity of likelihood calculation

	Input	Grid
Hypersphere	$n_P n_R n_{MC}$	$(n_{row} n_{col})^2 n_R n_{MC}$
Duocylinder-Same	$n_P n_R n_{MC}$	$(n_{row} n_{col})^2 n_R n_{MC}$
Duocylinder-Diff	$n_P n_R^2 n_{MC}$	$(n_{row} n_{col})^2 n_R^2 n_{MC}$

Table 4.4: Computational intensity of cluster detection

	Input	Grid
Hypersphere	$n_P n_R (2n_{Cl}+n_{MC})$	$(n_{row} n_{col})^2 * n_R (2n_{Cl}+n_{MC})$
Duocylinder-Same	$n_P n_R (2n_{Cl}+n_{MC})$	$(n_{row} n_{col})^2 * n_R (2n_{Cl}+n_{MC})$
Duocylinder-Diff	$n_P n_R^2 (2n_{Cl}+n_{MC})$	$(n_{row} n_{col})^2 * n_R^2 (2n_{Cl}+n_{MC})$

Since the total number of scanning windows is much larger than the number of input records or the number of output clusters, the memory requirement of multidimensional scan statistics algorithms is decided by the number of scanning windows. Specifically, for each scanning window, the numbers of points (e.g., case and control counts when using a Bernoulli model, two integers) in it and its log-likelihood value (one floating point) need to be stored. Other memory usages, such as storages the input coordinates and final cluster information, are much smaller compared to the memory usage for storing the scanning windows.

4.4.3 Parallel Computing

The algorithm of multidimensional scan statistics can be parallelized to leverage high-performance computing resources and cyberGIS (Wang and Armstrong 2009, Wang 2010). First, the computation of stage 1 (counting observations in each scanning window) and stage 2

(likelihood calculation), are independent for different scanning window centers, and thus can be parallelized in a straightforward manner by assigning each subtask to process a share of all scanning window centers. Second, stage 3 (cluster detection) can be parallelization using a slightly more complicated approach. In this stage, each subtask will still be assigned to the same subset of window centers. In order to detect each cluster, each subtask first detects the scanning window with the highest likelihood among its subset, and the global scanning window with the highest likelihood is identified as the cluster, among these local best ones. The information of this detected cluster is then distributed to all subtasks and, in preparation for detecting the next cluster, these subtasks invalidate all the remaining scanning windows that intersect with the just-detected cluster. Third, in stage 4 (MC simulation), since only the maximum likelihood of each MC replication is necessary, different replications can be calculated in parallel and only their maximum likelihoods need to be collected. In summary, there are three major approaches to parallelize the multidimensional scan statistics algorithm: (1) to conduct replications of MC in parallel, (2) to parallelize the cluster detection procedure within each replication and (3) to do both (1) and (2).

The choice of parallel approaches and parallel computing models depends on the memory constraints. In a typical application of multidimensional scan statistics, the memory requirement is mostly defined by the total number of scanning windows, since it is usually several orders of magnitude larger than the number of input observations. If approach (2) is used, MC replications are conducted sequentially, and hence only one replication of all scanning windows need to be kept in memory. Both shared- and distributed- memory parallel computing are valid options. When using shared-memory parallelization, all input observations are accessible from every parallel process, and each process processes its own share of scanning windows that are stored in

the shared memory. If distributed-memory parallelization is used, each process will have an entire copy of input observations but only its own share of scanning windows in its memory. Messages need to be passed between processes in order to identify scanning windows with the highest likelihood.

If MC simulation is parallelized (using approach (1) or (3)), the memory requirement is greatly increased because each replication being processed needs to store the events count and likelihood values of all scanning windows. The event counts and likelihood values vary among replications, and thus have to be stored independently. Hence, when using approach (1), distributed memory parallelization becomes a better option due to both greatly increased total memory requirement and limited communication between iterations. A master-slave can be used here - the master process that detects clusters from original datasets can simply collect the maximum likelihood values from slave processes each processing different MC iterations, and then calculate the *p*-value of each detected cluster. When using approach (3), a hybrid parallel computing model can be used. While iterations of MC simulation can be parallelized using a distributed-memory model as in approach (1), the cluster detection procedure within each iteration can be parallelized using a shared memory model for efficient usage of memory.

4.5 EXPERIMENTS AND RESULTS

Two experiments are used to compare the proposed scanning window designs and the associated algorithms, in terms of both computational efficiency and the ability to identify meaningful clusters. In the first experiment, the New York City taxi trip patterns in the morning rush hour and afternoon rush hour are compared. Since almost every taxi trip has a pick-up location and a drop-off location that are unique from each other, the dataset contains individual OD movement records. In the second experiment, the migration patterns of age group 25-29 and

age group 65-69 in the US are compared. The migration dataset is aggregated to county level and is presented as the count of individuals migrating between each pair of US counties. The datasets for both of these two experiments were previously used by Gao et al. (2018). The algorithms for all six scanning window designs were implemented using the second strategy that is mentioned in the Parallel Computing section. The software code was written in C and uses OpenMP for shared-memory parallel computing. Experiments in this chapter were conducted using a computing node with Intel Xeon E5-2680v3 processors (24 cores) and 128 GB DDR4 DRAM.

4.5.1 New York City Taxi Trip Analysis

4.5.1.1 Data source and experiment settings

This experiment uses taxi trip records in New York City on a typical workday (01/21/2015). The data were downloaded from the New York City's official website[8]. It contains both Yellow and Green taxi trips. For each taxi trip, its pick-up (origin) and drop-off (destination) locations and timestamps are recorded in the dataset. Taxi trips that start during the morning rush hours (7:00 AM to 10:00 AM) and afternoon rush hours (5:00 PM to 8:00 PM) are extracted respectively, which are referred to as morning trips and afternoon trips respectively. The dataset contains 72,144 morning trips and 83,524 afternoon trips.

Multidimensional scan statistics with a Bernoulli model were used to compare the spatial patterns between morning and afternoon taxi trips, and detect spatial clusters against a null hypothesis that the probability of each trip in the combined-morning-and-afternoon dataset to be a morning trip is, regardless of its origin and destination, constant at 72144/(72144+83524)=0.463. The clustering detection procedure detects both clusters with a significantly high number of morning taxi trips (morning clusters) and the ones with a

[8] Data source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

significantly high number of afternoon taxi trips (afternoon clusters) simultaneously. Algorithms for all six scanning window designs are applied to the taxi trip dataset. For all six designs, scanning window radii vary from 100 m to 2500m with a 100m interval. When using regular grid points as scanning window centers, a 400m by 400m grid covering the major part of New York City (Figure 4.6) is used. This grid contains 46 rows and 36 columns for a total of 2,742,336 scanning window centers. Monte Carlo simulation with 99 replications is carried out for statistical inferences.

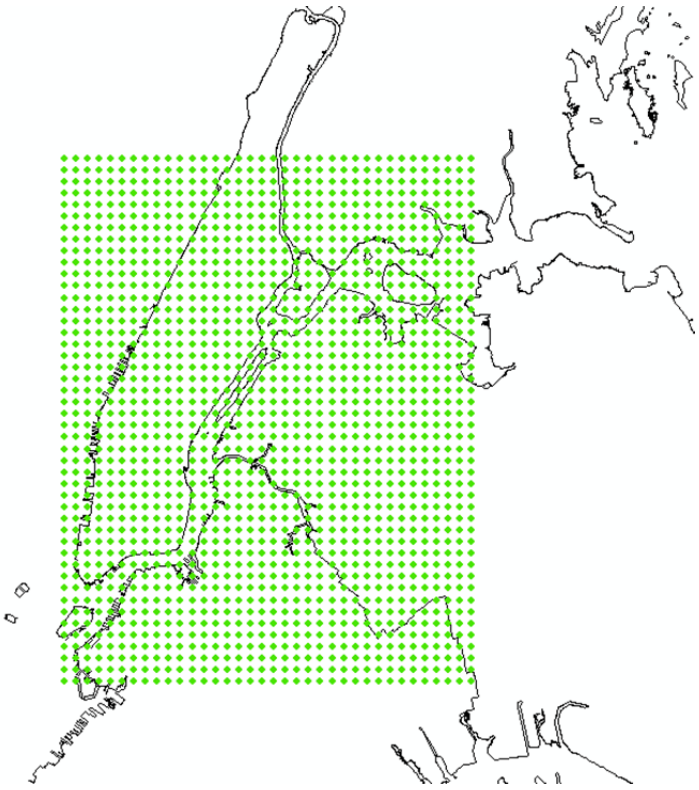

Figure 4.6: New York taxi trips. Regular grid points used as scanning window centers.

4.5.1.2 Results

The computing times of the six scanning window designs without MC simulation are shown in Table 4.5, and results with MC of 99 replications are shown in Table 4.6. For all scanning window designs, the computing time with MC (Table 4.6) is nearly 100 times than

without MC (Table 4.5). This is because while the most time-consuming clustering procedure needs to be repeated in each MC replication, the input, the output and the detection of secondary clusters are only conducted once. Among all these six scanning windows, the Hypersphere-Input is the fastest and Duocylinder-Diff-Grid is the slowest. For all three scanning window shapes, using input records as scanning window centers has a much higher computational efficiency than using regular grid points. For both two scanning window centering approaches, the Hypersphere has the highest efficiency while Duocylinder-Diff has the lowest. Such computational efficiency differences can be mostly explained by the differences of the total scanning windows checked in these six algorithms (Table 4.7). Hypersphere-Input and Duocylinder-Same-Input need to check the smallest number of scanning windows and are thus faster. Using grid points as centers requires checking a much larger number of scanning windows and is thus slower. For the slowest case, Duocylinder-Diff-Grid algorithm needs to check almost 2 billion scanning windows and is much slower than any other algorithms. Hypersphere approaches are usually slightly faster than Duocylinder-Same approaches, although they check the same number of scanning windows. This is because Duocylinder-Same approaches need to calculate origin distances and destination distances respectively (Table 4.1), and thus is slightly slower in checking whether an observation is in a window or not.

The quality of clustering results from the six scanning window designs are evaluated by comparing the log likelihood of the top cluster detected by each algorithm. The results are shown in Table 4.8. In scan statistics, given the same dataset, clusters with higher likelihoods are considered to be more likely and thus better. Hence, methods that can detect clusters with higher likelihood are considered to be more effective. As shown in Table 4.8, in terms of scanning window shapes, Hypersphere detects clusters with the lowest likelihood, Duocylinder-Diff

detects ones with the highest likelihood, and Duocylinder-Same lies in between. The reason for the superiority of Duocylinder-Diff over Duocylinder-Same is clear since the Duocylinder-Same windows are only a subset of Duocylinder-Diff windows. In terms of scanning window centers, although using regular grid points greatly increased the total computing cost, the results is not necessarily better. The scanning window design that has the highest maximum log-likelihood is Duocylinder-Diff-Input.

Table 4.5: New York taxi trips. Total computing time in seconds without MC simulation.

	Input	Grid
Hypersphere	59.095	533.791
Duocylinder-Same	64.992	705.962
Duocylinder-Diff	300.623	1408.949

Table 4.6: New York taxi trips. Total computing time in seconds with MC simulation.

	Input	Grid
Hypersphere	4997.852	51399.69
Duocylinder-Same	6394.398	68357.48
Duocylinder-Diff	28544.11	107661.9

Table 4.7: New York taxi trips. Total number of scanning windows tested.

	Input	Grid
Hypersphere	3,891,700	68,558,400
Duocylinder-Same	3,891,700	68,558,400
Duocylinder-Diff	97,292,500	1,713,960,000

To further compare the clustering results of the six algorithms, their top five clusters are visualized in Figure 4.7 and Figure 4.8. Since each cluster represents a collection of movement flows from some region to another region, it is represented as an arrow indicating the cluster

center, a solid circle representing its origin extent, and a dashed circle representing its destination extent. Figure 4.7 compares scanning window centering allocations, where windows are centered at input records in Figure 4.7 (a) and at regular grid points in Figure 4.7 (b). Both of these results are generated using Hypersphere scanning windows. In Figure 4.7 (a), cluster 2 and 3 are morning clusters (e.g., clusters with significantly more morning trips than expected), and cluster 1, 4, 5, are afternoon clusters. More detailed descriptions of results using the same experiment settings as Figure 4.7 (a) are provided in Gao et al. 2018. In Figure 4.7 (b), cluster 1 and 4 are morning clusters, and cluster 2, 3, 5 are afternoon clusters. High consistency exists between the results in Figure 4.7 (a) and (b). Although the exact spatial extents and the orders are slightly different, the top five clusters in these two figures are almost the same, and each cluster in Figure 4.7 (a) has its counterpart in Figure 4.7 (b) (Table 4.9).

Table 4.8: New York taxi trips. Log-likelihood of the top cluster.

	Input	Grid
Hypersphere	-106336.7914	-106330.3512
Duocylinder-Same	-106067.5939	-106051.0918
Duocylinder-Diff	-105560.6824	-105704.4888

Figure 4.8 compares the results of Duocylinder-Same (Figure 4.8 (a)) and Duocylinder-Diff (Figure 4.8 (b)) scanning windows. Both of these results use scanning windows that are centered at input records. In Figure 4.8 (a), cluster 2 and 4 are morning clusters, and cluster 1, 3 and 5 are afternoon clusters. The clustering patterns (Figure 4.8 (a)) share many similarities with the patterns using Hypersphere windows (Figure 4.7 (a)). Three out of the five top clusters have highly similar counterparts in Figure 4.7 (a): 1st to 1st in Figure 4.7 (a), 2nd to 2nd in Figure 4.7 (a), and 4th to 3rd in Figure 4.7 (a). Cluster 3 is similar to cluster 5 combined with some flows in cluster 4 in Figure 4.8 (a). (A Duocylinder-same cluster contains a much larger area than a

Hypersphere with the same radius as explained in Figure 4.1.). A cluster similar to the cluster 5 in Figure 4.8 (a) is the 30th cluster when using Hypersphere windows.

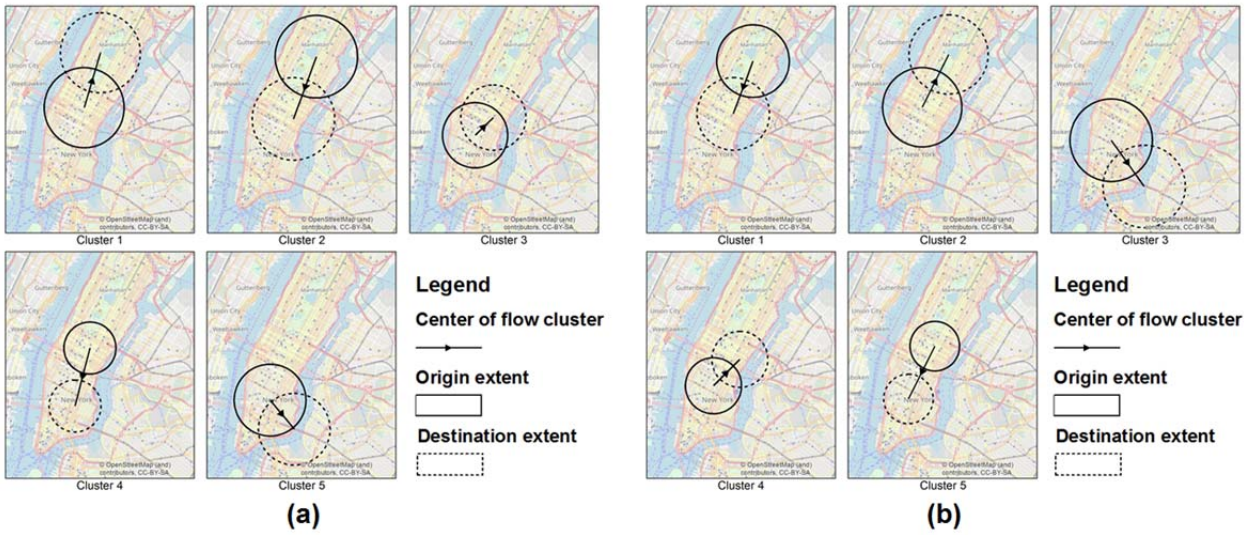

Figure 4.7: New York taxi trips. Result comparison between scanning window center allocation approaches. 4.6a. Hypersphere-Input. 4.6b. Hypersphere-Grid.

Table 4.9: New York taxi trips. Top clusters in Figure 4.6 (a) and (b), and their relationship.

Figure 4.7 (a)			Figure 4.7 (b)		
Cluster ID	Morning trips	Afternoon trips	Cluster ID	Morning trips	Afternoon trips
1	2960	8672	2	2855	8434
2	8030	3912	1	7645	3598
3	11120	8112	4	5428	3258
4	628	2507	5	801	2697
5	435	2084	3	427	2396

In Figure 4.8 (b), cluster 1, 4 and 5 are morning clusters, and cluster 2 and 3 are afternoon clusters. The top clusters shown in Figure 4.8 (b) demonstrate clear differences from the top clusters in Figure 4.8 (a). Although cluster 2 and 3 in Figure 4.8 (b) are similar to cluster 1 and cluster 3 in Figure 4.8 (a), each of the remaining three clusters in Figure 4.8 (b) has an origin

extent and a destination extent of clear different areal sizes. Both cluster 1 and 2 have higher likelihoods than the top cluster when using Duocylinder-Same scanning windows (Figure 4.8 (a)). Cluster 1 in Figure 4.8 (b) is a cluster of morning trips from a large area covering Midtown, Upper West Side and Upper East Side to a much smaller area around the Times Square, Rockefeller Center and the Midtown section of Fifth Avenue. This cluster likely depicts the commuting trips from affluent residential areas in Manhattan to work places in Midtown. Since this cluster intersects with cluster 2 in Figure 4.8 (a), a cluster similar to that one is unable to be detected using Duocylinder-Diff scanning windows. It is also worth noting that this cluster has destination zone located fully within its origin zone. It demonstrates that it is entirely legitimate and meaningful to have a cluster's destination intersect with its origin, which is also discussed in Gao et al. 2018. Cluster 4 in Figure 4.8 (b) originates from Lower Manhattan and ends at an area similar to the destination of cluster 1. Cluster 5 represents the cluster of morning trips from Cluster 1's origin to the southernmost part of Manhattan. The destination of this cluster is also a history, culture and business center of the New York City with famous landmarks such as the Statue of Liberty, Wall Street and the 9/11 Memorial, which are the destinations of many morning commuting and business trips.

4.5.2 US Migration Flow Analysis

4.5.2.1 Data source and experiment settings

This experiment uses the US county-to-county migration flow data derived from 2006-2010 American Community Survey that is available from the US Census Bureau[9]. This dataset contains the estimated annual migrants from each US county to any other US county. The centroid of each county is used to represent a county's location, and hence the migration from

[9] Data source: http://www.census.gov/hhes/migration/data/acs/county-to-county.html

county A to county B is simplified as a flow from A's centroid to B's centroid. The migration flows of age group 25-29 and age group 65-69 are extracted, and the migrants in these two age groups are referred as young migrants and senior migrants respectively. There are 1,788,892 young migrants and 175,515 senior migrants between 155,668 different county pairs.

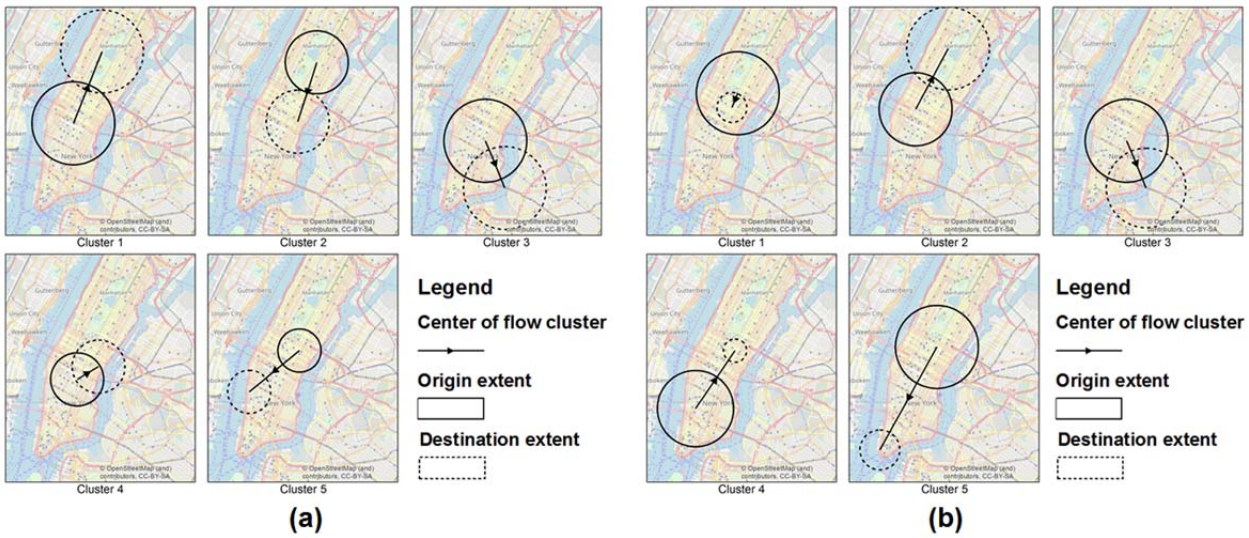

Figure 4.8: New York taxi trips. Result comparison between Duocylinder-Same and Duocylinder-Diff scanning window shapes. 4.8a. Duocylinder-Same-Input. 4.8b. Duocylinder-Diff-Input.

This experiment also uses a Bernoulli model to compare the spatial migration patterns between the two age groups. However, different from the previous experiment, it aims to find OD region pairs with a significantly high number of senior migrants against the null hypothesis that 0.0893 of migrants are senior anywhere. As a consequence, only clusters of senior migrants are identified in the cluster detection procedure. The scanning window radii used in this experiment are 40km, 80km, … , 1000km. When using regular grid points as scanning window centers, a 100km by 100km grid covering the conterminous United States is used. This grid contains 28 rows and 47 columns for a total of 1,731,856 scanning window centers. The exclusion of Alaska and Hawaii in the grid is because top detected clusters using input records are all in the conterminous United States and that using a single grid that includes non-

contiguous areas is too computationally expensive. This experiment also uses Monte Carlo simulation with 99 replications.

4.5.2.2 Results

The computing times of the six scanning window designs are shown in Table 4.10 (without MC) and Table 4.11 (with MC) respectively. The log likelihoods of the top cluster detected by each algorithm are shown in Table 4.12. These performance results demonstrate several similar patterns to the New York City taxi trip experiment: using grid points as window centers greatly increases the computing cost; Duocylinder-Diff windows are the slowest to process but the results have the highest likelihood; using Hypersphere window is slightly faster than using Duocylinder-Same windows. However, two patterns different from the previous experiment. First, for all three scanning window shapes, using regular grid points as window centers consistently generate results with a higher likelihood. A likely reason for such a difference is that when aggregated movement data are being analyzed, possible choices for scanning window location are limited to these aggregated units. By using regular grid points as cluster centers, a larger search space for clusters could be provided and thus better clusters are more likely to be found. Second, different from the previous experiment where Duocylinder-Same windows detect clusters with higher likelihood than Hypersphere for both center allocation approaches, in this experiment, Duocylinder-Same-Grid has better results than Hypersphere-Grid.

Table 4.10: US county-to-county migration. Total computing time in seconds without MC simulation.

	Input	Grid
Hypersphere	2.921	83.915
Duocylinder-Same	3.696	111.587
Duocylinder-Diff	22.653	400.833

Table 4.11: US county-to-county migration. Total computing time in seconds with MC simulation.

	Input	Grid
Hypersphere	267.016	7770.997
Duocylinder-Same	336.347	10493.93
Duocylinder-Diff	2044.59	28533.61

Table 4.12: US county-to-county migration. Log-likelihood of the top cluster.

	Input Records	Grid
Hypersphere	-586355.0098	-586097.5452
Duocylinder-Same	-586569.8941	-585806.6395
Duocylinder-Diff	-586046.6544	-585705.0636

To visually compare the clustering results, the top 10 clusters using Duocylinder-Same-Input, Duocylinder-Diff-Input, Duocylinder-Same-Grid, and Duocylinder-Diff-Grid are shown in Figure 4.9, 4.10, 4.11 and 4.12 respectively. These figures show that many movement clusters are between regions of different areal sizes, and that Duocylinder-Diff can detect these more informative clusters than Duocylinder-Same. Migration clusters from a large sparse area to a dense urban area (e.g. cluster 6 in Figure 4.10), from a subcontinental region to a specific state (e.g. cluster 1 in Figure 4.10) or from a dense urban area to a larger region (e.g. cluster 8 and 10 in Figure 4.10) can only be detected with scanning windows that allow for origin and destination with areal size differences (i.e. Duocylinder-Diff, Figure 4.10 and Figure 4.12). By limiting

clusters to have an origin and a destination of the same size, Duocylinder-Same's results (Figure 4.9 and 4.11) lack such explanatory power of these spatial patterns, although the major migration trends can still be captured and the detected clusters are still significant. In terms of scanning window allocation approaches, when using a regular grid approach (Figure 4.11 and 4.12), more clusters from or to low density areas can be detected. These clusters are difficult to be identified when using the input location approach, since there have to be actual flows from one county to another county in order for this county pair to be considered as a candidate cluster center, which can be difficult in low density areas. Furthermore, when using the grid approach, it is possible to put the cluster at a location without a nearby county centroid. A noticeable example of this is cluster 1 in both Figure 4.11 and 4.12, which describes the migration flows from the Northeast and East North Central to South Atlantic region. The best cluster center to describe the origin (combined Northeast and East North Central area) is in outside of US boundary in Canada, which is only possible when using the regular grid approach. With a input location approach, compromise needs to be made such that the origin center of this cluster is set as close to the US-Canada boundary as possible.

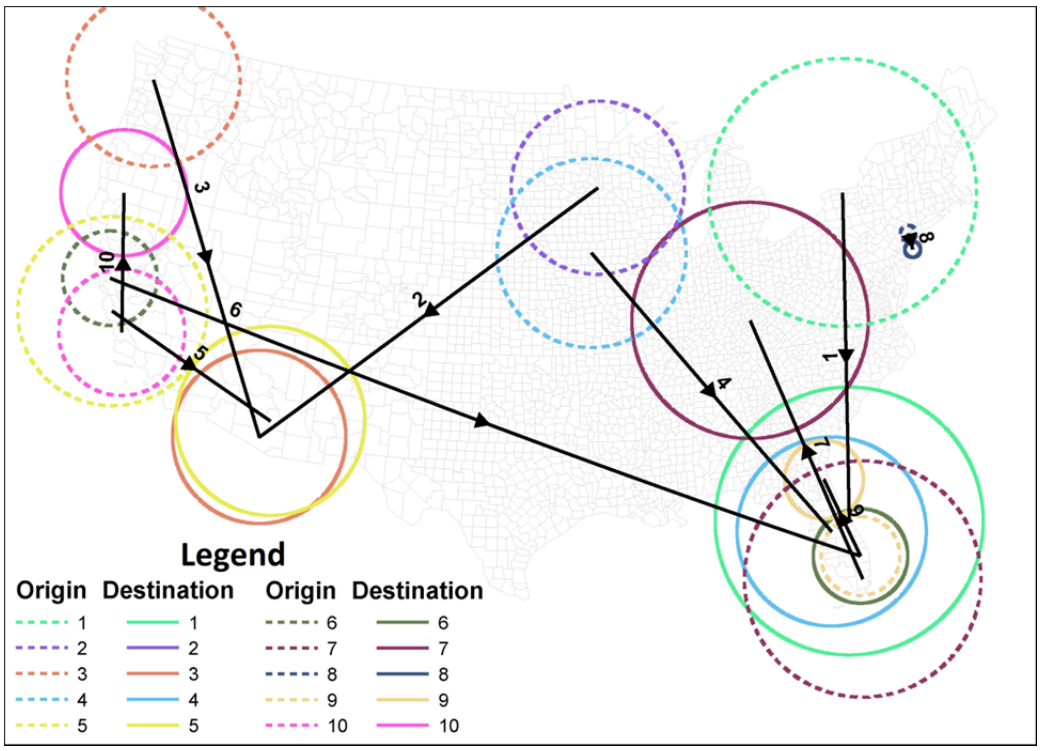

Figure 4.9: US county-to-county migration. Top 10 clusters using Duocylinder-Same-Input

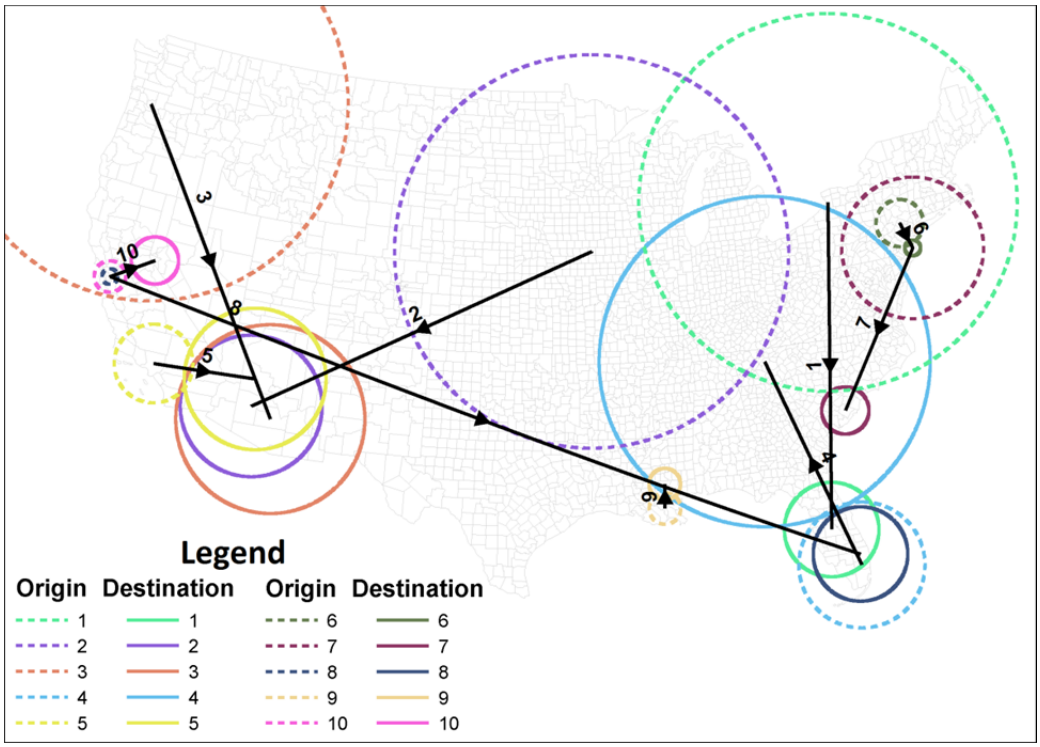

Figure 4.10: US county-to-county migration. Top 10 clusters using Duocylinder-Diff-Input

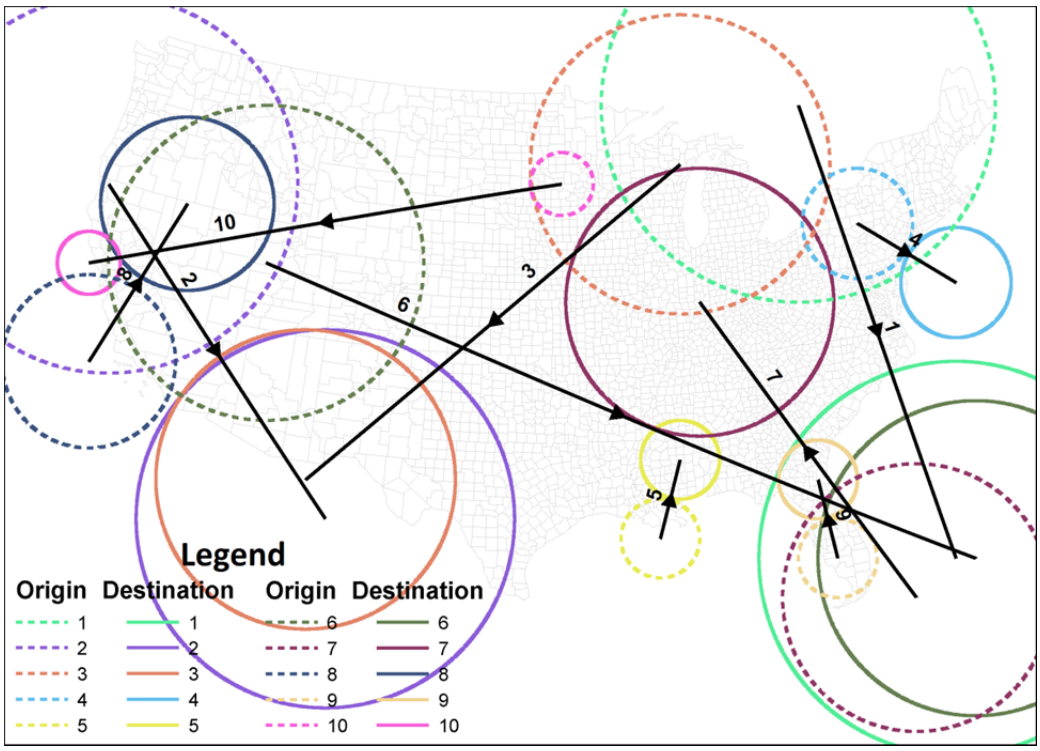

Figure 4.11: US county-to-county migration. Top 10 clusters using Duocylinder-Same-Grid

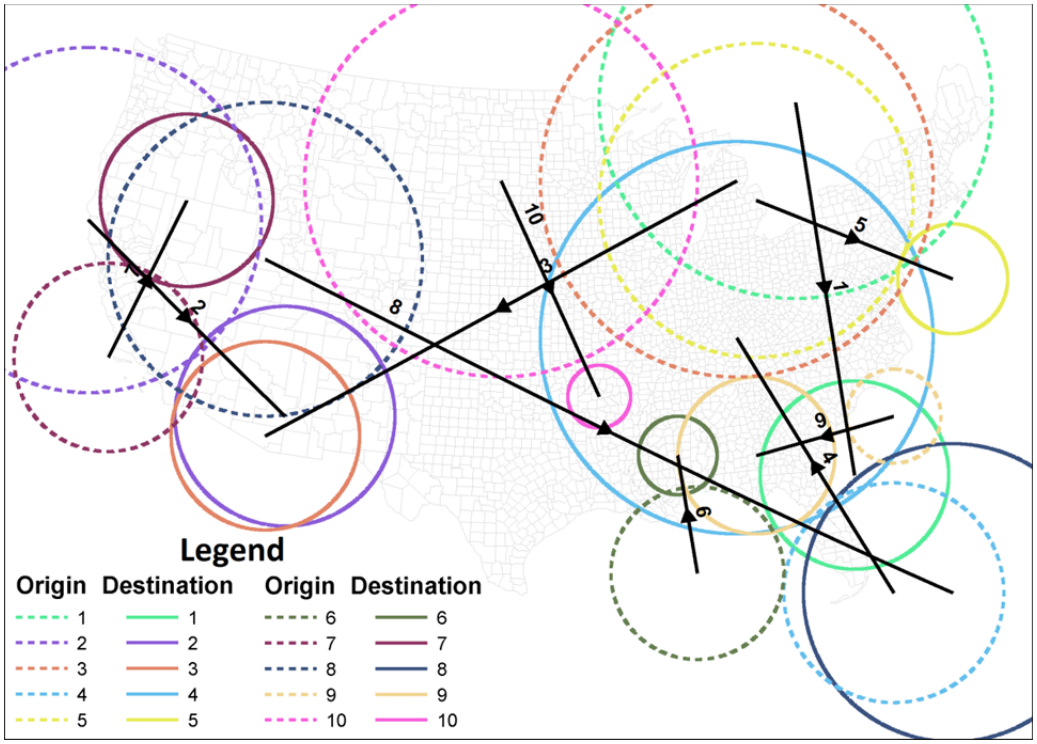

Figure 4.12: US county-to-county migration. Top 10 clusters using Duocylinder-Diff-Grid

4.6 CONCLUSION

This chapter explores and evaluates scanning window designs for movement pattern analysis with multidimensional scan statistics. For each of the three important aspects of scanning window design – shape, location, and size, this chapter reviews commonly used approaches in 2D spatial scan statistics and 3D space-time scan statistics, analyzes their advantages and disadvantages, and proposes 4D extensions that are necessary and valid for movement pattern analysis. Three 4D shapes, Hypersphere, Duocylinder-Same, and Duocylinder-Diff, are identified as valid scanning window shapes. Combined with two centering allocation strategies – input locations and 4D regular grid points – six scanning window designs are proposed. The algorithms for the six scanning window designs are then developed, and parallel computing approaches are developed to advance cyberGIS and exploit HPC.

The six scanning window designs are evaluated and compared both analytically and through two real-world applications, in terms of computational performance and the quality of clustering results. Analytical and experimental results provide insights into these scanning window designs and the trade-off between computing intensity and the ability to detect high-quality clusters for movement pattern analysis. First, many real movement clusters are from a larger origin to a smaller destination or *vice versa*, and such clusters can only be accurately detected when such differences are allowed (i.e. using Duocylinder-Diff windows). However, Duocylinder-Diff requires a much higher computing cost. Hypershpere and Duocylinder-Same still provide valid options if users need a quick summary of movement patterns. Second, using 4D regular grid points as window centers greatly increased the computing cost, and it is generally not recommended when analyzing individual-level movement records since the results will not be better and may even be worse. However, it has a clear edge when analyzing

aggregated movement data since it increases the search space for clusters, and hence better clusters are more likely to be identified.

The two experiments described typical and realistic parameter settings. The actual computing requirement and clustering results also depend on the choice of other parameters, such as the number of window sizes at each location or the grid size when using regular grid centers. For instance, when using regular grid points as centers, it is theoretically possible to use a very fine grid such that each input observation can be accurately approximated by one grid point. Under this setting, the clusters detected using regular grid points as centers will always be better than using input records as centers. However, such a fine grid is practically impossible due to its high computing cost and not worth exploring. Furthermore, if users intend to use Duocylinder-Diff windows but have limited computing resources, they can use window radii with a larger interval to reduce the total number of scanning windows and hence the computing requirement. By doing so, they can sacrifice cluster radii's accuracy for the ability to capture the origin-destination radius differences.

Future research can be pursued in multiple directions. First, as mentioned in section 4.3.5, it is worthy to explore more flexibly shaped scanning windows such as 4D ellipsoid or extensions of duocylindrical windows from circular to elliptical origins and destinations. These flexibly shaped scanning windows may be necessary to describe complicated movement patterns such as the migration from a coastal area to a river basin, both of which are not circular. It would be challenging both methodologically and computationally and hence requires extensive investigations. Second, reducing the computational intensity of multidimensional scan statistics represents another research direction to improve this methodology. Counting the number of observations in each scanning window and testing each scanning window for clusters have high

computing requirement. Such requirement could be greatly reduced if the amount of cluster testing can be reduced through approaches such as approximations (Agarwal et al. 2006), sampling (Matheny et al. 2016), or expansion (Li et al. 2018). Third, the multidimensional scan statistics method and different scanning window designs need to be evaluated in a variety of applications, using different datasets, and at different spatial scales. In addition to movement or migration, it would be worthwhile to explore the application of this method to other spatial interaction datasets, such as friendship relations, industry relocations, citations or even the hyperlinks on the internet. The experiences gained from these applications will drive methodological and computational innovation of the multidimensional scan statistics.

CHAPTER 5: CONCLUSION AND FUTURE DIRECTIONS

5.1 SUMMARY OF CONTRIBUTIONS

This dissertation established several novel approaches to analyzing geospatial big data and to detecting spatial patterns based on generalized spatial point representations. These approaches are presented in three pieces of work to address real-world data-intensive applications of event detection from social media data and movement pattern analysis. They together provide the methodology and techniques for analyzing geospatial big data, and examples of how they can be applied to real-world problems to resolve application-specific challenges.

The first piece of work (chapter 2) starts with analyzing conventional spatial points. This chapter describes a systematic approach to detecting spatiotemporal patterns of events from social media data. Individual geo-located social media posts are perceived as points in space-time, each representing the location where and the time when a post is generated. Text analysis approaches are then employed to identify whether a post is related to a particular event. Finally, the collective patterns of massive posts are used to map the spatiotemporal variations of event prevalence and to identify space-time regions with potentially abnormal event activities. Challenges of mapping spatiotemporal patterns of events based on social media data are identified, which include: large but unstable volumes of available data with a potentially changing spatial distribution; spatial heterogeneity of both underlying event activities and the popularity of event-related topics; and data sparsity when analyzing fine-scale spatial patterns. A scientific approach is thus proposed in order to find a consistent event indicator from social media data by tackling these challenges. The approach employs several interrelated strategies including: using a KDE to generate smoothed social media intensity surfaces; using event-

unrelated social media posts as controls to map the relative spatial distribution (ratio map) of event-related posts through KDE; and uses local historical values as the null hypothesis to identify areas with abnormal event activities. The approach was applied to detecting influenza trends in the conterminous US using geo-located *Twitter* data. Results showed that event patterns detected from social media data have high consistency with ground truth data at available scales. It also provides a solid benchmark of influenza activity maps for future research based on social media or other new data sources.

Starting from the second piece of work (chapter 3), this dissertation goes beyond conventional geographical points into generalized points and uses multidimensional conceptual points to analyze more complicated geographic phenomena such as spatial movements and spatial interactions. Each OD spatial movement trip or spatial interaction record is a directional connection between an origin and a destination, both of which are conventional spatial points. Chapter 3 of this dissertation proposes a multidimensional point model for OD movements. If both the origin and the destination are represented as 2D points, each movement record is modeled as a spatial point with four spatial dimensions in a 4D OD space. With this multidimensional point data model, the spatial patterns of OD movements are reflected in the 4D point patterns. 2D spatial point process models and analytical approaches can thus be extended into the 4D space to analyze OD movement patterns.

In addition to the multidimensional point model, the second piece of work (chapter 3) describes a multidimensional scan statistics approach to comparing OD movement patterns based upon the data model. This scan statistics approach is able to evaluate the differences and similarities between two OD movement patterns by detecting areas where the two spatial patterns differ the most. To achieve this goal, a bivariate marked spatial point process model is

used to describe OD movements represented by 4D points. A multidimensional Bernoulli spatial scan statistics method is developed to detect OD region pairs with abnormally high concentrations of one movement dataset over the other, using random labeling as the null hypothesis. The existence and the spatial extents of these OD region pairs indicate whether and where the two movement distributions differ. Case studies demonstrated that the approach can effectively detect spatial patterns from large movement datasets, and is applicable to both individual-level and aggregated movement data.

The third piece of work (chapter 4) follows chapter 3 and improves the multidimensional scan statistics by investigating the most essential element of it - the design and selection of scanning windows. Based on the methodological basis of multidimensional point data model and scan statistics, this chapter explores what scanning window designs should be used for movement pattern analysis. From shape, location, and size perspectives, this chapter analyzes commonly used approaches in 2D spatial analysis and 3D space-time analysis and investigates how these approaches can be extended into 4D spaces. Six scanning window designs are proposed by combining three shapes and two center allocation approaches. This chapter also designs efficient algorithms and parallel computing approaches for efficient movement pattern analysis from large movement datasets by leveraging cyberGIS and HPC. These scanning window designs are evaluated using both individual-level and aggregated movement datasets. The analysis of scanning windows and the experiment results demonstrate the advantages and disadvantages of these scanning window designs. Recommendations are given for the choice of scanning window designs to achieve the trade-off between computing cost and the ability to detect high-quality clusters.

In summary, the contributions of dissertation can be viewed in methodological, and technological, and empirical perspectives. Methodologically, this dissertation presents models, methods, and approaches to analyze complex geographic phenomena from spatial big data through generalized point analysis approaches. They help to answer questions such as how to reveal meaningful spatial and spatiotemporal patterns from large datasets, and how to compare these patterns for gaining insights into the complexity of such patterns. From the technological point of view, this dissertation describes scientific workflows, computational approaches, and software tools that combine spatial analytical approaches, cyberGIS and geospatial big data. Computational challenges of the analytical processes are analyzed and HPC approaches are provided. This dissertation also leverages cyberGIS resources and platform to retrieve, store, and process geospatial big data. On the empirical side, this dissertation applies the aforementioned approaches to real-world research questions such as influenza activity mapping, traffic pattern analysis, and migration studies. Valuable spatial patterns are detected from these large real datasets, which can help to gain insight into the space-time dynamics of the questions being studied. The dissertation also provides detailed end-to-end descriptions to analysis procedures, the characteristics of the data, and how the choice of experiment settings many influence analytical results. These results can also serve as solid benchmarks for future research.

5.1.1 Open-sourced Software Tools

Algorithms and software tools are increasingly essential components of scientific research. Providing efficient open-sourced codes for the scientific community is another crucial component and major contribution of this dissertation work. While many pieces of software codes or scripts have been developed during this dissertation research, here the two most important pieces are presented. They are GPU-based KDE, and scan statistics (for both 2D and

4D spaces). Both of these two software tools are parallel codes that are designed for cyberGIS environment to take advantage of HPC resources.

5.1.1.1 Multi-GPU KDE

Multi-GPU KDE is a tool to calculate the spatial density of massive points through a distance decay kernel function. The computation of KDE is challenging when the input point data volume is large or the output map resolution is high. In data-intensive analytical tasks, it is common to have millions or even billions of input points (e.g., there are millions of geo-located tweets in the US alone daily). Such computational intensity cannot be well handled by conventional sequential KDE algorithms and toolkits, and thus high-performance parallel KDE solutions are developed as part of CyberGIS toolkit, taking advantage of HPC resources. The Multi-GPU KDE leverages a cluster of graphics processing units (GPUs) for efficient KDE calculation. The spatial characteristics of KDE are well-suited for GPU algorithms:

- First, the calculation for each grid pixel is independent of each other. Thus, each pixel can be processed by one thread and the final KDE results can be calculated by a large number of parallel threads simultaneously.
- Second, the process of calculating density values is similar between pixels: sequentially reading each input data point, calculating the distance and then accumulating the contribution of input data points. This process can be handled efficiently using a Single Instruction Multiple Data architecture.
- Third, a pixel is only influenced by its nearby input data points and nearby grids need to access similar input data points. Thus, it is efficient to group the threads for nearby grids into blocks with shared memories to be processed together.

The GPU KDE algorithm can be further improved through a grid-based spatial indexing. Grid pixels are decomposed into hyperrectangle blocks whose sides are the same as kernel bandwidth. All input points are decomposed using the same blocks. Since input points more than one bandwidth away from a grid pixel cannot contribute to the density value at a specific pixel, only input points in one and its direct-neighboring blocks are necessary to calculate the density value for a pixel in that block. In a 2D case, as shown in Figure 5.1, only the input points in the 9 blocks surrounded by the red square are necessary for any of the blue pixels in Block(2,2). Furthermore, the computational tasks of KDE can be decomposed into multiple GPU nodes to further increase the speed-up. Such decomposition is conducted based on the estimated spatial computational domain to achieve load balancing (Wang and Armstrong 2009). It utilizes blocks as the smallest spatial unit for decomposition, and adaptive partitioning approach (Ding and Densham 1996) are utilized to decompose the computational domain into subdomains with similar computing load.

The multi-GPU KDE code was implemented using C and Compute Unified Device Architecture (CUDA). Message Passing Interface (MPI) is used to communicate and distribute tasks between GPU nodes. It is part of CyberGIS toolkit public available at https://github.com/cybergis/cybergis-toolkit. A different single GPU version was utilized by the Quantum Population Geo-Analytics projects at Construction Engineering Research Laboratory (Ehlschlaeger et al. 2016).

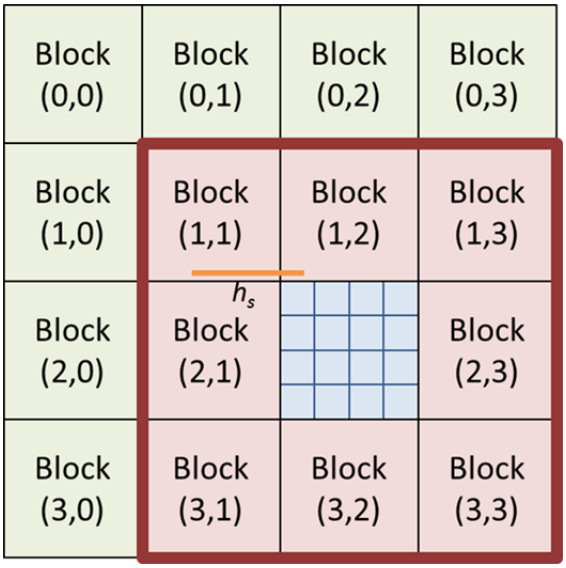

Figure 5.1: An illustration of spatial indexing and spatial computational domain decomposition for KDE.

5.1.1.2 Parallel spatial scan statistics

While SaTScan™ developed by Kulldorff has been the most well-known software for scan statistics, it has some noticeable drop backs as discussed in chapter 3: only for analysis of conventional point in geographic space and designed for the single desktop environment not for geospatial big data. Furthermore, SaTScan™ does not open source its code, and thus it is difficult to tune or modify its method. Hence, open-sourced cyberGIS solutions to data-intensive scan statistics are both crucial to this dissertation research, and may potentially benefit the GIS and spatial analysis community. Hence, open-sourced parallel spatial scan statistics are developed as part of this dissertation research.

In this dissertation research, mainly three versions of spatial scan statistics are produced. Although all three versions are based on the same scan statistics methodology, they are designed for different purposes. The first version is designed for analyzing 2D spatial points and supports both Poisson and Bernoulli point process models. Different from SaTScan™ where point to

point distances are used as scanning window radius, our codes use numbers of an equal interval (e.g. 1km, 2km, ..., 100km) that can be specified by users. With this specification, the number of total scanning windows checked can be greatly reduced when analyzing large datasets. The sources code and instructions for it is available at https://github.com/tsccsj/SpatialScan. The second version is 4D spatial scan statistics that is used in chapter 3. This version uses 4D spherical scanning windows that centered at input observations for movement pattern analysis. This version is available at https://github.com/tsccsj/4DSpatialScan. Finally, the third version extends from the third version, and provides the support to all six scanning window designs that are mentioned in chapter 4. The algorithm details for all these scanning windows are described in depth in section 4.4. The source code is available at https://github.com/tsccsj/Adv4DSpaScan.

All three versions are developed in C and OpenMP. OpenMP (Open Multi-Processing) is an application programming interface (API) for shared-memory multiprocessing programming. It is used in our implementation to distribute the scan statistics calculation and cluster detection procedure to multiple processor cores while allowing them to simultaneously access all input records and scanning window info stored in the shared memory.

5.2 FUTURE WORK

Topics in this dissertation can be expanded in multiple related directions as the future work.

First, we plan to apply the proposed models, methods, and approaches to other application areas. For instance, in addition to influenza activity mapping, the methods described in chapter 2 might be useful to map other events such as earthquake or tornado to help with the real-time awareness of and timely response to natural disasters. The 4D scan statistics approach described in chapter 3 can be used to evaluate whether social media can be a solid indicator of

real migration patterns, or to compare the moving patterns of poor and rich families within a city to gain insight into urban landform changes. These applications are of similar nature as the case study used in this dissertation and should be straightforward to apply the developed approaches. The results and feedback from these applications can be helpful to improve the proposed approach and suggest new analytical approaches that could be added to our methodological framework.

Second, the scientific communities will benefit from scientific software or toolkits that are robust, easy-to-use, and generalizable. While open-sourced software codes have been produced as part of this dissertation, more research and technical work is necessary in order to improve the computational efficiency and usability of these tools. It is worthy of investigating additional performance improvement strategies, necessary parameters for better generalizability, user-friendly input and output, and software stability. A long-term goal for this would be to produce standalone software tools that can be used on different platforms, have a graphic user interface, and contain more related functionalities. Good examples of these tools include SatScan™, GeoDa, and CyberGIS Gateway. Another direction for the software tool is to make these analytical functionalities as library that can be called in other programming languages such as R or Python.

Third, it would be essential to explore additional spatial point process models and analytical methods for generalized spatial points. While conventional 2D point process models such as Poisson or Bernoulli, and conventional 2D spatial analysis approaches such as KDE or scan statistics can be extended into multidimensional space, some applications may have research questions that are not easily approachable using these existing models or tools. For instance, when analyzing spatial interaction models, it is possible to have a spatial point process

models that are derived from a gravity model where distance decay influences the spatial interaction strength. Then we could possibly detect strong interacts, possibly caused by strong socioeconomical ties, that cannot be explained by distance decay alone.

The fourth future work direction is to further develop the generalized point model and associated analytical approaches to other types of geographic phenomena. While this dissertation only covers conventional point observations (e.g. social media, 2 spatial dimensions) and OD movement flows (or more generally bivariate spatial interactions, 4D points), the generalized spatial point models may be applied to analyze other types of data. One example of it is to include temporal dimensions in movement analysis, where there may be one (a timestamp for each OD record) or two (origin and destination with separate timestamps) temporal dimensions. Adding these additional temporal dimensions requires a higher dimensional (e.g., 5D or 6D) spatiotemporal point data models. The generalized point model may also help to push spatial interactions research from binary spatial interactions (e.g. friendship, director trips) to ternary spatial interactions (e.g. A, B are both friends of C, trips with intermediate stops, home-work-shopping/childcare patterns) or even n-ary spatial interactions. By doing so, more complicated interaction dynamics might potentially be revealed.

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B., 2011. Predicting Flu Trends using Twitter data. In: *2011 IEEE Conference on Computer Communications Workshops*, 10-15 April 2011, Shanghai, China. Washington, DC: IEEE, 702–707.

Agarwal, D., McGregor, A., Phillips, J. M., Venkatasubramanian, S., and Zhu, Z., 2006. Spatial scan statistics: approximations and performance study." In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 24-33.

Albuquerque, J.P. de, Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science, 29* (4), 667-689.

Anderson, N.H., Titterington, D.M., 1997. Some methods for investigating spatial clustering, with epidemiological applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(1), 87-105.

Andrienko, N., Andrienko, G., 2011. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, 17 (2), 205–219.

Andrienko, G., Andrienko, N., 2008. Spatio-temporal aggregation for visual analysis of movements. In: D. Ebert and T. Ertl eds. 2008 *IEEE Symposium on Visual Analytics Science and Technology*, 19-24 Oct. 2008, Columbus, OH, USA. Washington, D.C., USA: IEEE, 51–58.

Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data. ACE*, pp. 49-60.

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27 (2), 93–115.

Aramaki, E., Maskawa, S., Morita, M., 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *In*: *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 27-31 July 2011, Edinburgh, United Kingdom. Stroudsburg, PA, USA: Association for Computational Linguistics, 1568–1576.

Aslam, A.A., Tsou, M.-H., Spitzberg, B.H., An, L., Gawron, J.M., Gupta, D.K., Peddecord, K.M., Nagel, A.C., Allen, C., Yang, J.-A., Lindsay, S., 2014. The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance. *Journal of medical Internet research*. 16(11).

Asur, S., Huberman, B.A., 2010. Predicting the Future with Social Media. *In*: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 31 Aug.-3 Sept. 2010, Toronto, ON, Canada. Washington, D.C., USA: IEEE, 492–499.

Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data. ACE*, pp. 49-60.

Bailey, T.C., Gatrell, A.C., 1995. *Interactive spatial data analysis*. Harlow: Longman Scientific & Technical.

Baker, D.M., Valleron, A.-J., 2014. An open source software for fast grid-based data-mining in spatial epidemiology (FGBASE). *International Journal of Health Geographics*, 13 (1), 46.

Barwick, H., 2011. The" four vs" of big data. Implementing information infrastructure symposium. *2012-10-02]. http://www. computerworld. com. au/article/396198/iiis_four_vs_big_data*.

Bell, M.G.H., 1991. The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, 25 (1), 13–22.

Berglund, S., Karlström, A., 1999. Identifying local spatial association in flow data. *Journal of Geographical Systems*, 1 (3), 219-236.

Bithell, J.F., 1991. Estimation of relative risk functions. *Statistics in Medicine, 10* (11), 1745-1751.

Bithell, J.F., 1990. An application of density estimation to geographical epidemiology. *Statistics in medicine, 9*(6), 691-701.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of computational science, 2* (1), 1-8.

Buchin, K., Speckmann, B., Verbeek, K., 2011. Flow Map Layout via Spiral Trees. *IEEE Transactions on Visualization and Computer Graphics*, 17 (12), 2536–2544.

Calabrese, F., Ratti, C., Lorenzo, G.D., Liu, L., 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing*, 10 (4), 36–44.

Cascetta, E., Nguyen, S., 1988. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodologica*l, 22 (6), 437–455.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., Soltani, K., 2015. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*, 70-82.

CDC, 2011. *Cold Versus Flu* [online]. Available from: http://www.cdc.gov/flu/about/qa/coldflu.htm [Accessed March 03, 2015]

CDC, 2014. *Key Facts about Influenza (Flu) & Flu Vaccine* [online]. Available from: http://www.cdc.gov/flu/keyfacts.htm [Accessed March 03, 2015]

CDC, 2015. *Overview of Influenza Surveillance in the United States* [online]. Available from: http://www.cdc.gov/flu/weekly/overview.htm [Accessed March 03, 2015]

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

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S., Ertl, T., 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. *In: 2012 IEEE Conference on Visual Analytics Science and Technology*, 14-19 Oct. 2012, Seattle, WA, USA. Washington, D.C., USA: IEEE, 143–152.

Corley, C., Mikler, A.R., Singh, K.P., Cook, D.J., 2009. Monitoring Influenza Trends through Mining Social Media. *In: International Conference on Bioinformatics and Computational Biology*, 13-16 July 2009, Las Vegas, NV, USA. Athens, GA, USA: CSREA, 340–346.

Cromley, E.K., McLafferty, S.L., 2011. *GIS and public health*. New York, NY, USA: Guilford Press.

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

Cui, W., Zhou, H., Qu, H., Wong, P.C., Li, X., 2008. Geometry-Based Edge Clustering for Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14 (6), 1277–1284.

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

Culotta, A., 2010. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. *In*: *Proceedings of the First Workshop on Social Media Analytics*, 25 – 28 July 2010, Washington, D.C., USA. New York, NY, USA: ACM, 115–122.

Culotta, A., 2010b. Detecting influenza outbreaks by analyzing Twitter messages. *arXiv preprint arXiv:1007.4748.*

Davies, T.M., Hazelton, M.L., 2010. Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine, 29* (23), 2423-2437.

Ehlschlaeger, C.R., Gao, Y., Westervelt, J.D., Lozar, R.C., Drigo, M.V., Burkhalter, J.A., Baxter, C.L., Hiett, M.D., Myers, N.R., Hartman, E.R., 2016. Mapping neighborhood scale survey responses with uncertainty metrics. *Journal of Spatial Information Science*, 2016(13). 103-130.

Diggle, P.J., 2013. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, Third Edition. Boca Raton, FL, USA: CRC Press.

Ding, Y., Densham, P.J., 1996. Spatial strategies for parallel spatial modelling. *International Journal of Geographical Information Systems*, 10(6), 669-698.

Doan, S., Ohno-Machado, L., Collier, N., 2012. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. *In: 2012 IEEE Second*

International Conference on Healthcare Informatics, Imaging and Systems Biology, 27-28 Sept. 2012, San Diego, CA, USA. Washington, D.C., USA: IEEE, 62–71.

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

Duczmal, L., and Assunção, R., 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45 (2), 269-86.

Duczmal, L., Cançado, A. L., Takahashi, R. H., and Bessegato, L. F. 2007. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, 52 (1), 43-52.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., Vaughan, A., 2010. OMG earthquake! Can Twitter improve earthquake response?. *Seismological Research Letters, 81* (2), 246-251.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: AAAI Press, 226–231.

Fischer, M.M., Wang, J., 2011. *Spatial Data Analysis: Models, Methods and Techniques*. Berlin, Germany: Springer Science & Business Media.

Gao, Y., Ting, L., Wang, S., Jeong, MH and Soltani, K., 2018. A multidimensional spatial scan statistics approach t movement pattern comparison. *International Journal of Geographical Information Science*, doi: 10.1080/13658816.2018.1426859.

Gilbert, P.D., 1995. Combining var estimation and state space model reduction for simple good predictions. *Journal of Forecasting, 14* (3), 229-250.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014.

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

Guo, D., 2009. Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. *IEEE Transactions on Visualization and Computer Graphics*, 15 (6), 1041–1048.

Guo, D., Zhu, X., 2014. Origin-Destination Flow Data Smoothing and Mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20 (12), 2043–2052.

Guo, D., Zhu, X., Jin, H., Gao, P., Andris, C., 2012. Discovering Spatial Patterns in Origin-Destination Mobility Data. *Transactions in GIS*, 16 (3), 411–429.

Hasan, S., Zhan, X., Ukkusuri, S.V., 2013. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. *In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing,* 11 – 11 August 2013, Chicago, IL, USA. New York, NY, USA: ACM.

Helwig, N.E., Gao, Y., Wang, S., Ma, P., 2015. Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance. *Spatial Statistics, 14*, 491-504.

Hilbert, M., 2016. Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1), 135-174.

Hinneburg, A., Keim, D.A., 1998. An efficient approach to clustering in large multimedia databases with noise. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. ACM, 58-65.

Holten, D., Van Wijk, J.J., 2009. Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum*, 28 (3), 983–990.

Hsu, C.W., Chang, C.C., and Lin, C. J., 2003. *A Practical Guide to Support Vector Classification* [online]. National Taiwan University. Available from: http://zoro.ee.ncku.edu.tw/mlb2007/res/MLB08_1.pdf [Accessed 03 March 2015].

Huang, L., Kulldorff, M., Gregorio, D., 2007. A Spatial Scan Statistic for Survival Data. *Biometrics*, 63 (1), 109–118.

Hwang, M.-H., Wang, S., Cao, G., Padmanabhan, A., Zhang, Z., 2013. Spatiotemporal Transformation of Social Media Geostreams: A Case Study of Twitter for Flu Risk Analysis. *In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, 05 – 05 November 2013, Orlando, FL, USA. New York, NY, USA: ACM, 12-21.

Illian, D.J., Penttinen, P.A., Stoyan, D.H., Stoyan, D., 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*. Hoboken, NJ, USA: John Wiley & Sons.

Jung, I., Kulldorff, M., Klassen, A.C., 2007. A spatial scan statistic for ordinal data. *Statistics in medicine*, 26 (7), 1594–1607.

Jung, I., Kulldorff, M., Richard, O.J., 2010. A spatial scan statistic for multinomial data. *Statistics in medicine*, 29 (18), 1910–1918.

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

Kelsall, J.E., Diggle, P.J., 1995a. Non-parametric estimation of spatial variation in relative risk. *Statistics in medicine, 14* (21-22), 2335-2342.

Kelsall, J.E., Diggle, P.J., 1995b. Kernel Estimation of Relative Risk. *Bernoulli*, 3-16.

Kulldorff, M., 2015. *SaTScan™ User Guide for version 9.4* [online]. Available from: https://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download [Accessed 20 December 2017].

Kulldorff, M. and Nagarwalla, N., 1995. Spatial disease clusters: detection and inference. *Statistics in medicine*, 14 (8), 799-810.

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

Kulldorff, M., 1999. Spatial scan statistics: models, calculations, and applications. In: J. Glaz. and N. Balakrishnan eds. *Scan statistics and applications*. Boston, MA: Springer, 303-322.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., Mostashari, F., 2005. A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLOS Medicine,* 2 (3), e59.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. 2006. An elliptic spatial scan statistic. *Statistics in medicine*, 25 (22): 3929-3943.

Kulldorff, M., Huang, L., Konty, K., 2009. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics,* 8 (1), 58.

Kurashima, T., Iwata, T., Irie, G., Fujimura, K., 2010. Travel Route Recommendation Using Geotags in Photo Sharing Sites. *In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 26 – 30 October 2010, Toronto, ON, Canada. New York, NY, USA: ACM, 579-588.

Kulldorff, M., Athas, W.F., Feurer, E.J., Miller, B.A., Key, C.R., 1998. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American journal of public health*, 88 (9), 1377-1380.

Kulldorff, M., 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164 (1), 61-72.

Lamb, A., Paul, M.J., Dredze, M., 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 09-14 June 2013, Atlanta, GA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 789-795.

Lampos, V., 2012. Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods. Thesis (PhD). University of Bristol.

Lampos, V., Cristianini, N., 2012. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology, 3* (4), 72.

Lampos, V., Cristianini, N., 2010. Tracking the flu pandemic by monitoring the social web. *In*: *2010 2nd International Workshop on Cognitive Information Processing*, 14-16 June 2010, Elba, Italy. Washington, D.C., USA: IEEE, 411–416.

Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Lee, R., Sumiya, K., 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. *In: Proceedings of the 2nd ACM SIGSPATIAL*

International Workshop on Location Based Social Networks, 02 – 02 November 2010, San Jose, CA, USA. New York, NY, USA: ACM, 1–10.

Li, L., Goodchild, M.F., 2012. Constructing Places from Spatial Footprints. *In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, 06 - 06 November 2012, Redondo Beach, CA, USA. New York, NY, USA: ACM, 15–21.

Li, L., Goodchild, M.F., Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *cartography and geographic information science, 40* (2), 61-77.

Li, T., Gao, Y., Wang. S, 2018. ESCIP: An Expansion-based Spatial Clustering Method for Inhomogeneous Point Processes, in revision.

Liu, Y., Tong, D., Liu, X., 2015. Measuring spatial autocorrelation of vectors. *Geographical Analysis*, 47 (3), 300-319.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), .512-530.

Lichman, M., Smyth, P., 2014. Modeling Human Location Data with Mixtures of Kernel Densities. *In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 24 – 27 August 2014, New York, NY, USA. New York, NY, USA: ACM, 35–44.

Lloyd, A., Cheshire, J., 2017. Deriving retail centre locations and catchments from geo-tagged Twitter data. *Computers, Environment and Urban Systems 61*, 108–118.

Loader, C.R., 1991. Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, 23 (4), 751-771.

Lu, Y., Thill, J.C., 2003. Assessing the cluster correspondence between paired point locations. *Geographical Analysis*, 35 (4), 290-309.

MacEachren, A.M., Jaiswal, A., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., Blanford, J., 2011. Senseplace2: Geotwitter analytics support for situational awareness. *In: 2011 IEEE Conference on Visual Analytics Science and Technology*, 23 – 28 October 2011, Providence, RI, USA. Washington, D.C., USA: IEEE, 181–190.

Matheny, M., Singh, R., Zhang, L., Wang, K., and Phillips, J. M., 2016. Scalable spatial scan statistics through sampling. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 20.

McGuckin, N., Murakami, E., 1999. Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women. *Transportation Research Record: Journal of the Transportation Research Board*, (1693), 79–85.

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

McIntosh, J., Yuan, M., 2005. Assessing similarity of geographic processes and events. *Transactions in GIS, 9* (2), 223-245.

Nagel, A.C., Tsou, M.-H., Spitzberg, B.H., An, L., Gawron, J.M., Gupta, D.K., Yang, J.-A., Han, S., Peddecord, K.M., Lindsay, S., Sawyer, M.H., 2013. The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets. *Journal of medical Internet research, 15*(10).

Naus, J.I., 1965a. Clustering of Random Points in Two Dimensions. *Biometrika,* 52(102), 263–267.

Naus, J.I., 1965b. The Distribution of the Size of the Maximum Cluster of Points on a Line. *Journal of the American Statistical Association,* 60 (310), 532–538.

Openshaw, S. and Openshaw, S., 1984. The modifiable areal unit problem. Geo Abstracts University of East Anglia.

Padmanabhan, A., Wang, S., Cao, G., Hwang, M., Zhang, Z., Gao, Y., Soltani, K., Liu, Y., 2014. FluMapper: A cyberGIS application for interactive analysis of massive location-based social media. *Concurrency and Computation: Practice and Experience, 26* (13), 2253-2265.

Paul, M.J., Dredze, M., Broniatowski, D., 2014. Twitter Improves Influenza Forecasting. *PLOS Currents Outbreaks [online]*. Available from: http://currents.plos.org/outbreaks/article/twitter-improves-influenza-forecasting/ [Accessed 17 December 2017].

Pei, T., Wan, Y., Jiang, Y., Qu, C., Zhou, C., and Qiao, Y., 2011. Detecting arbitrarily shaped clusters using ant colony optimization. *International Journal of Geographical Information Science*, 25 (10), 1575-1595.

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., 2005. Flow map layout. In: J. Stasko and M. Ward eds. *IEEE Symposium on Information Visualization*, 23-25 Oct. 2005, Minneapolis, MN, USA. Washington, D.C., USA: IEEE, 219–224.

Pozdnoukhov, A., Kaiser, C., 2011. Space-time Dynamics of Topics in Streaming Text. *In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 01 – 01 November 2011, Chicago, IL, USA. New York, NY, USA: ACM 1–8

Sadilek, A., Kautz, H.A., Silenzio, V., 2012. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. *In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial*

Intelligence, 22-26 July 2012, Toronto, Ontario, Canada. Palo Alto, CA, USA: AAAI, 136-142

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *In: Proceedings of the 19th International Conference on World Wide Web*, 26-30 April 2010, Raleigh, NC, USA. New York, NY, USA: ACM, 851–860.

Shi, X., 2010. Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science, 24* (5), 643-660.

Shu, H., 2016. Big data analytics: six techniques. Geo-spatial Information Science, 19 (2), 119-128.

Signorini, A., Segre, A.M., Polgreen, P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One* 6, e19467.

Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. Boca Raton, FL, USA: CRC press.

Stefanidis, A., Crooks, A., Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78, 319–338.

Stoyan, D., Kendall, W.S., Mecke, J., Stochastic geometry and its applications. 1995. *Akademie-Verlag, Berlin*.

Takahashi, K., Kulldorff, M., Tango, T. and Yih, K., 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics*, 7 (1), 14.

Takahashi, K., Yokoyama, T., Tango, T., 2010. *FleXScan User Guide* [online]. Available from: https://sites.google.com/site/flexscansoftware/download_e [Accessed 30 May 2018].

Tang, Y., Zhang, Y.-Q., Chawla, N.V., Krasser, S., 2009. SVMs modeling for highly imbalanced classification. Tang, Y., Zhang, Y.Q., Chawla, N.V. and Krasser, S., 2009. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39 (1), 281-288.

Tango, T., Takahashi, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4 (1), 11.

Tao, R., Thill, J.C., 2016. Spatial cluster detection in spatial flow data. *Geographical Analysis*, 48 (4), 355-372.

Texas Department of State Health Services, 2013. *2012-2013 DSHS Flu Report Week 39* [online]. Available from: http://www.dshs.state.tx.us/idcu/disease/influenza/surveillance/2013/week39/ [Accessed March 03, 2015]

Texas Department of State Health Services, 2014. *Texas Influenza Surveillance Report 2013 – 2014 Season / 2014 MMWR Week 39 (Vol. 39)* [online]. Available from: http://www.dshs.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=8589992390 [Accessed March 03, 2015]

Thom, D., Bosch, H., Krueger, R., Ertl, T., 2014. Using Large Scale Aggregated Knowledge for Social Media Location Discovery. *In: 2014 47th Hawaii International Conference on System Sciences,* 6-9 January 2014, Waikoloa, HI, USA. Washington, D.C., USA: IEEE, 1464–1473.

Tobler, W.R., 1987. Experiments in Migration Mapping By Computer. *The American Cartographer,* 14 (2), 155–163.

Tobler, W.R., 1981. A Model of Geographical Movement. *Geographical Analysis,* 13 (1), 1–20.

Tsou, M.-H., Jung, C.-T., Allen, C., Yang, J.-A., Gawron, J.-M., Spitzberg, B.H., Han, S., 2015. Social Media Analytics and Research Test-bed (SMART Dashboard). *In: Proceedings of the 2015 International Conference on Social Media & Society*, 27 – 29 July 2015, Toronto, Ontario, Canada. New York, NY, USA: ACM, 2:1–2:7.

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

Tubergen, F. van, Maas, I., Flap, H., 2004. The Economic Incorporation of Immigrants in 18 Western Societies: Origin, Destination, and Community Effects. *American Sociological Review,* 69 (5), 704–727.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Icwsm, 10* (1), 178-185.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L., Clark, L.C., 1989. *Monitoring for clusters of disease; Application to leukemia incidence in upstate New York* [online]. Available from: https://ecommons.cornell.edu/bitstream/handle/1813/8724/TR000840.pdf?sequence=1 [Accessed 30 May 2018].

Wallenstein, S., Weinberg, C.R. and Gould, M., 1989. Testing for a pulse in seasonal event data. *Biometrics*, 45 (3), 817-830.

Wang, S., Cao, G., Zhang, Z., Zhao, Y., Padmanabhan, A., 2013. A CyberGIS Environment for Analysis of Location-Based Social Media Data. *In:* H.A. Karimi, ed. *Advanced Location-Based Technologies and Services*. Boca Raton, FL, USA: CRC, 187–205.

Wang, S., Armstrong, M.P., 2009. A theoretical approach to the use of cyberinfrastructure in geographical analysis. *International Journal of Geographical Information Science,* 23 (2), 169–193.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M.F., Liu, Y., Nyerges, T.L., 2013. CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science*, *27*(11), 2122-2145.

Wang, S., Hu, H., Lin, T., Liu, Y., Padmanabhan, A., Soltani, K., 2014. CyberGIS for Data-intensive Knowledge Discovery. *SIGSPATIAL Special,* 6 (2), 26–33.

Wang, S., 2010. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers, 100* (3), 535-557.

Wang, S., Liu, Y., Padmanabhan, A., 2016. Open cyberGIS software for geospatial research and education in the big data era. *SoftwareX*, 5, 1–5.

Wang, S., Goodchild, M. F., 2018. *CyberGIS for Geospatial Innovation and Discovery*. Dordrecht, Netherlands: Springer.

Washington State Department of Health, 2014. *Washington State Influenza Update Week 29: July 13 – July 19, 2014* [online]. Available from: http://www.doh.wa.gov/Portals/1/Documents/5100/420-100-FluUpdateSeason2014.pdf [Accessed March 03, 2015]

Wright, D.J. and Wang, S., 2011. The emergence of spatial cyberinfrastructure. *Proceedings of the National Academy of Sciences*, 108(14), 5488-5491.

Yin, J., Soliman, A., Yin, D. and Wang, S., 2017. Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, 31(7), 1293-1313.

Zhang, J.-D., Chow, C.-Y., 2013. iGSLR: Personalized Geo-social Location Recommendation: A Kernel Density Estimation Approach. *In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 05- 08 November 2013, Orlando, FL, USA. New York, NY, USA: ACM, 334–343.

Zhu, X., Guo, D., 2014. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Transactions in GIS,* 18 (3), 421–435.