The Pennsylvania State University

The Graduate School

Department of Geography

LOCAL INDICATORS OF TEMPORAL BURSTINESS FOR SPATIO-TEMPORAL EVENT ANALYSIS

A Dissertation in

Geography

by

Eun-Kyeong Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2018

The dissertation of Eun-Kyeong Kim was reviewed and approved* by the following:

Alan M. MacEachren
Professor of Geography
Dissertation Adviser
Chair of Committee

Donna J. Peuquet
Professor of Geography

Andrew M. Carleton
Professor of Geography

Clio Andris
Assistant Professor of Geography

Zhenhui Li
Associate Professor of Information Sciences and Technology

Cynthia A. Brewer
Professor of Geography
Head of the Department of Geography

*Signatures are on file in the Graduate School

ABSTRACT

Exploring space-time structures is essential to understanding geographic events that are localized in space and time. The need to understand a range of space-time events has prompted development of many exploratory spatio-temporal data analysis (ESTDA) methods that include spatiotemporal cluster detection (e.g. space-time scan statistics). The focus of this dissertation is on extending ESTDA methods and their application to include attention to temporal and spatio-temporal *burstiness* of dynamic, geographical scale phenomena. This focus on burstiness complements and extends existing research on spatial and spatio-temporal clusters. The concept of clusters can be defined in multiple ways; many statistical methods to detect clusters statistically define a clustered pattern as one that rejects a completely random process (CRP). CRP is based on the density of events within specific space and time. The dissertation introduces a new concept of temporal clusters, '*bursts*' that do not just reject CRP, but indicate a characteristic temporal pattern in which short term high activity periods labeled as 'bursts' alternate with extremely lengthy inactive periods. More specifically, this dissertation develops a new set of ESTDA statistics that characterize temporal burstiness, or frequency-invariant temporal regularity/irregularity, and applies the new statistics to analyzing geographic events including wildfires, geo-tweeting behaviors, and jet contrail outbreaks. In Chapter 1, I lay out the context for the overall work and describe the research goals and objectives. In Chapter 2, I review relevant literature on ESTDA statistics including point pattern analysis methods as well as the concepts and methods for burst analysis, and position the indicators of temporal burstiness presented here in the context of existing ESTDA methods. In Chapter 3, I develop a novel burstiness measure for event data with finite sample sizes, addressing the issue of small sample size effects of burstiness metrics as initially proposed in statistical physics. In Chapter 4, I propose a set of local indicators of temporal burstiness that integrates spatial containers filtering

input data points into the burstiness measure developed in Chapter 3; the resulting methods are applied to analysis of wildfire data as an illustration of their utility. In Chapter 5, I apply the proposed methodology in a more comprehensive application, for characterizing temporal regularity/irregularity patterns of contrail outbreaks in the conterminous United States. Geographically weighted regression (GWR) is adopted to explore relationships between the upper-troposphere (UT) meteorological factors, temporal burstiness, and frequency of contrail outbreaks. The concluding chapter, Chapter 6, highlights contributions, challenges, and future work. Overall, the components of this dissertation provide a methodology to explore the temporal burstiness of geographic events and help reveal new temporal aspects of geographic phenomena.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Chapters 1-2 and Chapter 6 of this dissertation were authored solely by Eun-Kyeong Kim.

Chapter 3 was authored by Eun-Kyeong Kim as the first author and Hang-Hyun Jo as the second author. In Chapter 3, Eun-Kyeong Kim initiated the research, uncovered the finite-size effect of the existing burstiness measure proposed by Goh & Barabási (2008), proposed the initial idea of the novel burstiness measure, and conducted empirical data analysis. Eun-Kyeong Kim and Hang-Hyun Jo devised the novel burstiness measure and were co-authors of a published paper (Kim and Jo, 2016); the text of this chapter is primarily derived from that paper. Hang-Hyun Jo proposed the localized model.

Chapter 4 presents contents of a co-authored paper to be submitted to a GIScience journal, subsequent to completion of the dissertation; Eun-Kyeong Kim is the first author and Alan MacEachren is the second author. In Chapter 4, Eun-Kyeong Kim devised and implemented the method proposed in the chapter, applied it to a case study, and was the primary author of the chapter text. Alan MacEachren provided input throughout the project, contributed to modifications of the method, and provided input on structure of the paper and revisions to portions of the chapter text.

Chapter 5 presents contents of a co-authored paper to be submitted to a climate science journal, subsequent to completion of the dissertation; this chapter was authored by Eun-Kyeong Kim as the first author and Andrew Carleton as the second author. In Chapter 5, Eun-Kyeong Kim designed the study and conducted data collection and processing for meteorological variables, data wrangling for existing contrail outbreak inventories, and statistical analyses, and was the primary author of the chapter text. Andrew Carleton provided contrail outbreak inventories of 2008-2009, suggested guidelines for data processing of the contrail inventories, and

contributed to the study design and revisions of portions of the chapter text. In addition to the authors of the paper, David Travis provided 2000-2002 contrail outbreak data.

ACKNOWLEDGEMENTS

This dissertation would not exist without the help and encouragement of Alan MacEachren. I thank him for being the best advisor that I could have. He taught me what qualities a good advisor should have, not by words but by his dedicated advising with scrutiny, fairness, consistency, transparency, and patience. I also thank Hang-Hyun Jo for intellectual discussions and mentoring over the years, which potentially inspired my research on bursts. He shared his knowledge and technical know-hows with me in collaboration for the paper edited in Chapter 3. I appreciate my committee members, Donna Peuquet, Andrew Carleton, Clio Andris, and Zhenhui Li for their devotion to offering thoughtful advice and support.

I thank the group of Marcel Salathé for sharing their geo-located Twitter data employed in Chapter 3 as well as Andrew Carleton and David Travis for sharing their U.S. contrail outbreak data used in Chapter 5. I would like to acknowledge Jennifer Balch's valuable inputs on the initial study on bursts of wildfires and Roger Downs' useful advice on my PhD proposal drafts at an early stage.

Many colleagues and academics have provided feedback on my manuscripts and presentations and I especially thank Mark Simpson, Yu-li Ko, Raechel White, and Jennifer Mason for their generosity in proofreading my writings many times.

I am grateful to Alan MacEachren, Thomas Lauvaux, and Jungwoo Ryoo for the opportunities to work as a graduate researcher on their research projects during my PhD program. It has been a great financial support to complete my dissertation as well as a chance to expand my practical skills and knowledge. Numerous other academics and professionals have influenced my journey toward earning a PhD and I particularly thank Chul Sue Hwang, Shi Hak Noh, Chang-Hyeon Joh, Frank Hardisty, and Alexander Klippel.

Words cannot express my gratitude and affection to my parents, siblings, significant other, friends, and Penny the cat for having made my life happier.

Chapter 1

Introduction

Time is critical to understanding human and natural phenomena. Sciences and society have striven to uncover temporal dynamics of human and environmental events. Have you ever observed how regularly you visit a local restaurant right next to your work? How regularly do you make a phone call to your best friends who live far away from you? How often do you post a photo with check-ins on Instagram? Have wildfires ever happened near where you live? How often have earthquakes with the magnitude of more than five occurred in California? What makes different temporal patterns of social or natural events? For any of these events, do they have any temporal regularity? Those questions may have been hard to answer in the past, due to the lack of data that capture dynamic characteristics of phenomena. Now, we live in the era of big data. Many data that include temporal components as well as geographic and social dimensions are being efficiently constructed, with advances in sensing and mobile technologies, crowdsourcing techniques, and social media. Open data movements have enhanced the accessibility to those data.

Data-driven research and the importance of spatiotemporal pattern analysis

Obtaining insightful information and knowledge from big data is becoming important in many disciplines. Disciplines including Geographic Information Science (GIScience) are reflecting this trend in research (Li *et al.*, 2016). CyberGIS has synthesized geographic information systems (GIS), spatial data analysis, and cyberinfrastructure (Wang, 2010). Spatial data science or geospatial data science has appeared as a form of broader data science

developments, specialized for processing and analyzing spatial or spatiotemporal data (Li *et al.*, 2016; Wang, 2016).

Following the CyberGIS revolution, data-driven research is a growing trend in GIScience and geography (Miller & Goodchild, 2015). This data-driven research trend is reshaping traditional scientific thought processes, with implications for the relative balance of inductive, deductive, abductive, and analogical reasoning. Deductive methods have been successful in science because they ensure logical consistency. Recently, radical arguments devaluing deductive methods in the petabyte age have appeared. Anderson (2008) argued 'the end of theory,' that data (i.e. numbers) speak for themselves because many dimensions and components are embedded in big data, and information is derived by using statistical methods and finding patterns. This perspective is aligned with John Stuart Mill's earlier argument that induction is the primary method of scientific discovery (Medawar, 1964; Wilson, 2014). However, relying solely on inductive approaches is not practical because observation of phenomena has limitations. In other words, some aspects of phenomena are still non-quantifiable or unobservable and a researcher's intervention is inevitable in observations and data analysis (Medawar, 1964; Boyd & Crawford, 2012).

Despite their limitations, data-driven approaches are, indeed, becoming an essential part of scientific thought processes. As an example, within GIScience, many spatiotemporal data mining methods have been developed to detect patterns having both spatial and temporal component and discover useful information and knowledge from data (e.g., Li, 2014). Predictive models combined with machine learning techniques learn patterns and rules from spatiotemporal data (e.g., Tran *et al.*, 2015). Even traditionally model-driven research is being combined with data. For example, the structure of recent agent-based models in geography is often designed referring to patterns extracted from massive data (e.g., Cenek & Franklin, 2018). In all those approaches, finding spatiotemporal patterns from data is critical to proceed with advanced

modeling and analysis. In this dissertation, I focus on developing an exploratory statistical method to detect spatiotemporal patterns of events.

Traditional point pattern analysis and the appearance of bursty pattern analysis

Extracting and characterizing spatiotemporal patterns from data are important steps in data-driven geographic research because examining patterns can potentially help infer the processes that may have caused observed patterns and be a way to test spatially and temporally related theories (Wiegand & Moloney, 2013). Among various exploratory data analysis methods, well-established exploratory spatio-temporal data analysis (ESTDA) statistics including point pattern analysis (PPA) methods (e.g., Ripley's K-function, STIK function) and local indicators of spatial association (LISA) (e.g., Local Moran's I) have enabled a first-hand examination of spatial or spatiotemporal patterns of the location, intensity, and relationships of geographic phenomena at both global and local scales (Ripley, 1976; Anselin, 1995; Gabriel & Diggle, 2009). Due to their simplicity, ESTDA methods have been widely used in many disciplines including ecology, spatial econometrics, and health sciences beyond geography (Nelson, 2012).

Traditional PPA methods are used to inspect whether spatial, temporal, or spatiotemporal distributions of positions of events in space and/or time (e.g., bike accidents, mass shootings, trees, wildfires, earthquakes) are regular, random, or clustered (O'Sullivan & Unwin, 2003). The results from PPA provide potentially useful information for establishing a hypothesis and designing further analysis. However, PPA has limitations in relation to characterizing various types of clustered patterns. Most PPA methods define clustered patterns in terms of the density of events, and distinguish clustered patterns from other patterns depending on how the observed pattern deviates from a completely random pattern generated from a null model of a Poisson random process (Wiegand & Moloney, 2004).

Recently, a special case of clustered patterns, *bursty patterns*, particularly in time, have been given attention outside of geography. Barabási (2005) initially proposed the concept of *bursts* to characterize temporal dynamics in human communications. A *bursty pattern* is defined as the pattern of time intervals (i.e., inter-event times) between succeeding events, in which short time periods of numerous events, analogically called '*bursts*,' alternate with very long inactivity periods with no events (Barabási, 2005; Vázquez *et al.*, 2006; Goh & Barabási, 2008; Kim & Jo, 2016).

Detecting bursty patterns is useful not only for inferring those hidden processes, but also for predicting collective behaviors of associated dynamic events. Extremely long inactivity periods potentially imply suppressive and inhibitive underlying mechanisms over those periods as well as longer influences of previous events on following events, referred to as a *long-range memory effect* (Goh & Barabási, 2008; Kim & Jo, 2016). One of the mechanisms yielding a bursty pattern is Barabási's (2005) preferential task selection based on a perceived priority; in email communications, people tend to reply quickly to important or urgent emails with a high priority, while they are likely to make delays in replying to or never responding to unimportant emails (Barabási, 2005). For a natural phenomenon, aftershock sequences of earthquakes are known to be bursty; the epidemic type aftershock sequence (ETAS) model is one of the most popular models that describe earthquake sequences by employing a self-similar process, in which every aftershock produces its own aftershocks (Utsu *et al.*, 1995; Bottiglieri *et al.*, 2009). In the social sphere, bursty patterns of human activities on a social network slow down or speed up epidemic spreading or diffusion (e.g., Vázquez *et al.*, 2007; Iribarren & Moro, 2011; Karsai *et al.*, 2011; Miritello *et al.*, 2011; Rocha *et al.*, 2011; Van Mieghem & Van de Bovenkamp, 2013; Perotti *et al.*, 2014; Delvenne *et al.*, 2015). In the application of epidemic spreading, detecting bursty patterns may help predict the speed of epidemic spreading based on temporal dynamics of interactions on social networks.

Burstiness measure and needs for improvements

One approach to characterizing bursty patterns is using a statistical indicator. A characteristic of the bursty pattern in time is that the variance of inter-event times is very large, while, in a completely random series, the variance of inter-event times is the same as the mean of inter-event times (Goh & Barabási, 2008; Kim & Jo, 2016). Using this trait, Goh & Barabási (2008) proposed a simple descriptive statistical indicator, the *burstiness parameter*, to measure burstiness. As defined by their parameter, *burstiness* indicates how bursty or regular events are over time (Goh & Barabási, 2008). The burstiness parameter is based only on the mean and variance of inter-event times and ranges from -1 to 0 to 1, respectively indicating regular, completely random, and completely bursty patterns (Goh & Barabási, 2008). Because it is simple and intuitive to interpret, the burstiness parameter has been widely used in many disciplines to analyze temporal patterns of events including earthquakes, heartbeats, cell phone communications, Wikipedia edits, geo-tagged tweets, and credit card trade (e.g., Goh & Barabási, 2008; Jo *et al.*, 2012a, 2015; Yasseri *et al.*, 2012; Kim & MacEachren, 2014; Zhao *et al.*, 2015; Gandica *et al.*, 2016).

Such a statistical measure can potentially serve as an exploratory data analysis tool to characterize temporal regularity patterns, or bursty patterns, of dynamic geographic phenomena and provide different perspectives from traditional PPA methods by focusing on the existence of long inactivity periods. However, research on leveraging a burstiness measure to analyze geographic events is still in its infancy. Specifically, while Goh & Barabási's (2008) burstiness parameter has been used for many spatiotemporal datasets, their measure assumes that the number of events to be analyzed is very large. There is need for a modification of the burstiness measure for small size datasets (Kim & Jo, 2016). Moreover, there is no guideline for adopting the burstiness measure for geographic events.

The goal and objectives of the research

Given the foregoing, the goal of this dissertation is to develop statistical indicators that characterize temporal regularity patterns of geographic events based on the variance of inter-event times. To achieve this, the objectives of the study are to:

- Review literature in both geography and complexity sciences closely relevant to the burstiness measure introduced above, and position the burstiness measure in the context of ESTDA methods (Chapter 2);
- Reveal effects of a small sample size in applying Goh & Barabási's (2008) burstiness parameter and propose a novel burstiness measure that addresses the small sample size effects while keeping the measure simple (Chapter 3);
- Provide guidelines for applying the novel burstiness measure to geographic events distributed over space and time, and propose a local burstiness measure that enables exploration of spatial distributions of temporal regularity patterns (Chapter 4);
- Apply the local burstiness measure to analyzing temporal regularity patterns of jet contrail outbreaks in the conterminous United States as a proof-of-concept case study, and discover spatial variations of relationships between upper-tropopause climate conditions, the local burstiness measure, and contrail outbreak frequency through geographically weighted regression analysis (Chapter 5);
- Summarize the dissertation and discuss the remaining challenges and opportunities to be explored in future work (Chapter 6).

The focus of each of the main chapters (Chapter 2, 3, 4 and 5) is detailed briefly below.

A New Concept of Clusters – 'Bursts' (Chapter 2)

As introduced above, bursts can be quantified by a simple measure. ESTDA methods and statistical indicators pertaining to temporal aspects of phenomena are reviewed in both geography and complexity sciences. This dissertation focuses on point-based events, so PPA methods among ESTDA methods are reviewed in more detail. PPA methods are categorized into a 'density-based' method or 'variance-based' one according to the properties that ESTDA methods use to characterize point patterns. Burstiness measures are positioned as a variance-based statistical indicator in the context of ESTDA approaches.

Burstiness Measure for Finite Event Sequences (Chapter 3)

Goh & Barabási's (2008) burstiness parameter has been widely used due to its simplicity, but the parameter in the form presented by Goh & Barabási (2008) is strongly affected by the finite number of events in the time series, called finite-size effects or small sample size effects. As the small sample size effects on the burstiness parameter have been largely ignored, I analytically investigate the finite-size effects of the burstiness parameter. Then I suggest a novel definition of burstiness that is free from small sample size effects and yet simple. Using the novel burstiness measure, the finite-size effects can be distinguished from the intrinsic bursty properties in the time series. I also demonstrate the advantages of the proposed burstiness measure by analyzing empirical datasets of Twitter data.

Local Indicators of Temporal Burstiness for Spatiotemporal Event Analysis (Chapter 4)

The novel burstiness measure devised in Chapter 3 enables one to characterize the burstiness of both large and small datasets. However, no local indicators for temporal burstiness

have been proposed. Hence, I extend the existing burstiness measure into a new set of ESTDA statistics, local indicators of temporal burstiness (LITB), to explore geographically local variation in temporal burstiness and develop algorithms to implement the temporal burstiness measures at both global and local spatial scales. Furthermore, this study provides guidelines on how to apply the burstiness measure for geographic events. Finally, I conduct a brief case study on wildfires to investigate temporal regularity patterns of wildfire events in California.

Spatio-Temporal Regularity Patterns of U.S. Contrail Outbreaks (Chapter 5)

Chapter 5 presents a more comprehensive proof-of-concept research application. This study applies the statistical indicators of local temporal burstiness proposed in Chapter 4 to spatial inventories of contrail outbreaks, in order to verify the usefulness of the methodology. This study first investigates spatial distributions of temporal regularity patterns of the contrail outbreaks in the conterminous United States in terms of the local temporal burstiness. The chapter explores spatial variations of relationships among local UT meteorological environments, the local temporal burstiness of contrail outbreaks, and the local temporal frequency of contrail outbreaks through geographically weighted regression (GWR) models, and examines spatial variations of the relationships.

Contributions of the research

This research has the potential to contribute to broader domains of study than those used as examples here. First, the methodology proposed in this study will benefit researchers in many disciplines involving spatiotemporal phenomena who are interested in quantifying the temporal regularity of spatiotemporal events and exploring spatial variations in temporal regularity. The

proposed methodology will enable researchers to inspect, model, and predict the temporal regularity of events, as existing ESTDA methods have contributed to many fields of study.

Second, this research contributes to methodological studies that extend statistical indicators originally designed with respect to only a temporal dimension for spatiotemporal analysis, as this study resolves the small sample size effects of an existing statistical measure and integrates spatial dimensions into the measure. Further, it can inspire geographic information scientists (GIScientists) and geographers to take and adapt a useful concept and analytical methods from other disciplines for geographic analysis.

Last but not least, this dissertation benefits the communities of complexity science and statistical physics by refreshing the importance of considering geometric spatial dimensions in research on temporal dynamics of complex systems. The spatial dimensions to be considered include the effect of spatial interactions among components in complex systems, the impact of spatial boundaries on dynamic phenomena, and spatial variations of statistical properties of complex systems.

Unique contributions of this study are introducing the concept of a bursty pattern as a different type of a clustered pattern from statistical physics to geography, and adapting a statistical methodology quantifying the temporal burstiness for spatiotemporal event analysis. It is expected that this research will bring attention to the long inactivity periods in geographic phenomena and their underlying suppressive processes. Further, this study raises attention to the importance of analytical methods to characterize extremely inhomogeneous distributions of events in time and/or space.

Chapter 2

A New Concept of Clusters – 'Bursts'

Exploratory Spatio-Temporal Data Analysis (ESTDA) for Spatiotemporal Events

Statistical and visual methods of ESTDA have been the key statistical techniques that enable researchers to summarize properties, detect patterns of spatiotemporal events, reason about spatiotemporal association and interactions, and formulate hypotheses (Tukey, 1977; Levine, 2004; She *et al.*, 2012). ESTDA techniques are useful when relationships among potentially associated factors are not intuitive enough to come up with a hypothesis or a formal model, particularly in cases where: 1) data are too large to do manual data inspection, 2) there are many known or unknown factors involved in spatiotemporal events, and 3) knowledge of domain experts on underlying processes of spatiotemporal events is not complete enough to rely upon. In the Big Data era, ESTDA is becoming a powerful tool to summarize data and find patterns at a collective level. In many ESTDA approaches, visual explorations (e.g., using a scatter plot) often allow us to find patterns intuitively. However, statistical methods are essential to find patterns when data are very large and high-dimensional. For instance, if we visualize movement trajectories of millions of individual passengers in a city on a map, one may only see visual clutters. In that situation, statistical methods provide a necessary complement to such visual approaches.

Many ESTDA statistics have been developed to describe and detect space-time associations (e.g., Knox Index, Mantel Index) and spatial, temporal, or spatiotemporal clusters in terms of point density (e.g., Ripley's K-function, STIK function, spatial or spatiotemporal scan statistics) or spatial, temporal, or spatiotemporal autocorrelation (e.g., Moran's I, Geary's c, Getis

and Ord's G) at both global and local levels (Moran, 1947, 1948; Geary, 1954; Knox & Bartlett, 1964; Mantel, 1967; Ripley, 1976, 1977; Getis & Ord, 1992; Kulldorff, 1997, 1999; Kulldorff *et al.*, 1998; Getis, 2007; Gabriel & Diggle, 2009). Unlike global indicators (e.g., Knox Index, Moran's I), local indicators (e.g., LISA; Anselin, 1995) describe a local tendency of spatial, temporal, or spatiotemporal distribution patterns and detect local-scale spatial/temporal clusters (e.g., local indicators of network-constrained clusters; Yamada & Thill, 2007).

Thanks to their simplicity, ESTDA methods have been widely used in various fields of study beyond geography, such as environmental science, economics, and medical science (Nelson, 2012). Many of those ESTDA statistics concern finding spatially, temporally, or spatiotemporally regular, random, or clustered patterns as those patterns provide an abstract representation of sets of individual observational data that enables understanding of phenomena that those data signify (Peuquet, 1994; Beard *et al*., 2008). Largely, there are two types of methods for detecting clustered patterns; one is based on the density or frequency of point patterns (e.g., K-function, scan statistics), and the other is based on spatial, temporal, or spatiotemporal autocorrelation in data values (e.g., Getis and Ord's G). Here, I focus on spatiotemporal or temporal point patterns of event-based phenomena. Many kinds of geographic events can be conceptualized as spatiotemporal point patterns: some phenomena occur momentarily over space and time (e.g., one geo-located tweet, the moment of a meteorite collision with the surface of the Earth); some phenomena are continuous over space and time, and they can be recorded as on or off regarding the occurrence of events within the given spatial, temporal, or spatiotemporal observational unit (e.g., a wildfire event in Sierra Nevada on the 26th of March in 2015).

Point Pattern Analysis: Density-Based vs. Variance-Based

Density-based PPA methods, including K-function and spatial or spatiotemporal scan statistics, have been effective in detecting local-scale spatial or spatiotemporal clusters and characterizing global spatial, temporal, or spatiotemporal regular, random, and clustered patterns. However, many density-based or frequency-based methods have limitations. First, the density-based methods usually detect point patterns by counting events within a range and rejecting a hypothesis that the density of events is likely to be obtained through a completely random process (CRP), so they do not provide information about distribution patterns of events inside of clusters and distribution patterns of clusters (Kulldorff, 1999; Zhang *et al.*, 2012). For instance, micro-scale patterns inside of detected clusters can be regular (e.g., Figure 2-1e) and global-scale patterns of detected clusters can be regular (e.g., Figure 2-1e and 2-1f). Second, even though a baseline point process is a homogeneous Poisson random process (e.g., Figure 2-1c and 2-1d), local-scale clusters can be detected if they have a higher density than the rest of a time period (e.g., Figure 2-1c). Third, those methods are not capable of characterizing patterns with extremely lengthy inactive periods, even though clusters can be detected in terms of the density (e.g., Figure 2-1g and 2-1h).

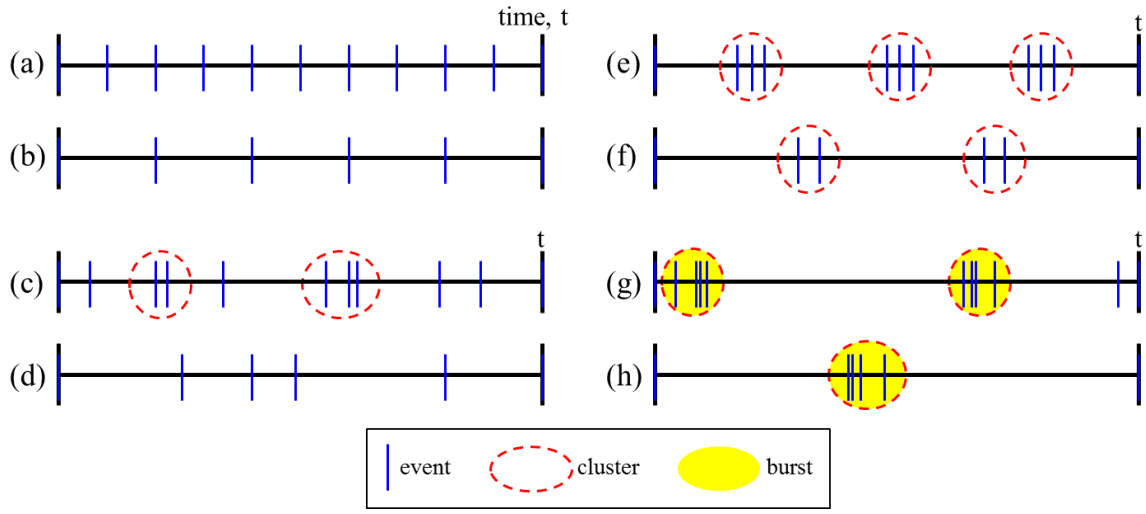


Figure 2-1. Differences between temporal clusters and temporal bursts in one dimensional space (time). (a) High-frequency regular pattern, (b) low-frequency regular pattern, (c) high-frequency random pattern, (d) low-frequency random pattern, (e) high-frequency clustered pattern, (f) low-frequency clustered pattern, (g) high-frequency bursty pattern, and (h) low-frequency bursty pattern. While temporal clusters can potentially be detected in (c), (e), (f), (g), and (h) in terms of the point density, apart from temporal patterns inside of those clusters, temporal bursts are only detected in (g) and (h) in which bursts are defined in relation to the presence of many lengthy inactive periods together with relatively high activity in narrow time ranges.

Although many PPA methods are focused on detecting clusters based on the density of events, spatial, temporal, or spatiotemporal interactions do not always appear as a form of spatial, temporal, or spatiotemporal clusters. Extremely lengthy inactive periods or extremely long distances between events can also provide substantial information to understand spatial/temporal interactions and processes in geographic events, particularly long inhibitive processes. As an example, while wildfires occur frequently in the dry season, they are often suppressed for a long time at some time other than the dry season. Also, the restoration time from previous severe wildfire disturbance results in fire return intervals of both short-term and long-term (Moritz *et al.*, 2011). Similarly, long-distance bike trips are less likely to be made when the weather is rainy and cold, and also during weekdays, owing to employment and social constraints, while middle-distance or short-distance bike trips can be made more frequently on a regular basis for

commuting. Patterns of bicycle use may also vary with different environments of places in terms of the propensity to rely on bicycles for commuting (e.g., a bike-friendly town like Boulder, Colorado versus the least safe transportation places like Pompano Beach, Florida[1]).

As described in the Introduction, temporal patterns with a large variance of inter-event times have been conceptualized as a 'bursty pattern' (Barabási, 2005; Clauset *et al.*, 2009) (Figure 2-1g and 2-1h). The rigid definition of a bursty pattern is based on heavy-tailed statistical distributions of inter-event time intervals including power-law or lognormal distributions that signify a generative process presuming long-term dependences between successive events, as opposed to Poisson random processes presuming independence among events (Barabási, 2005; Karsai *et al.*, 2012).

On the basis of the aforementioned statistical generative process of heavy-tailed distributions, previous studies of bursty patterns focused on modeling the underlying mechanisms of bursty patterns with respect to statistical parameters (e.g., power-law exponents) regarding human communication behaviors (e.g., Barabási, 2005; Jo *et al.*, 2012a; Jo *et al.*, 2012b), natural disasters (e.g., Karsai *et al.*, 2012), and queuing of jobs in supercomputers (e.g., Kleban & Clearwater, 2003). Traditionally, complexity scientists and statistical physicists have proposed mathematical or computational models yielding power-law distributions in terms of domino effect, positive feedback, and the influence of external factors' power-law behaviors. To validate those models, researchers have compared statistical parameters from patterns generated by a model with ones from empirical data. This validation method does not guarantee that only the proposed models generate such a temporal pattern; there can be multiple underlying mechanisms that generate similar bursty patterns (Shalizi, 2014). Nevertheless, those models are useful to predict patterns and simulate dynamic behaviors of phenomena depending on associated factors.

[1] http://www.care2.com/greenliving/americas-best-worst-bike-friendly-cities.html

However, before one models underlying mechanisms of a system that yields bursty patterns, whether a system exhibits bursty patterns should be determined. That is, to model a system that generates power-law behaviors, one needs to determine whether the system results in a power-law distribution. In this sense, exploratory approaches for characterizing bursty patterns are the first step to understand bursty phenomena. While there are no existing ESTDA methods that specifically characterize bursty patterns, as mentioned above, a burstiness measure proposed in the field of statistical physics that measures how bursty a temporal pattern is has been developed and widely used (e.g., Goh & Barabási, 2008). However, Goh & Barabási's (2008) burstiness measure does not consider small sample size effects (as mentioned in the previous chapter) nor does it specify how to consider the spatial context of geographic events. This dissertation focuses on developing spatially constrained local indicators for characterizing bursty patterns found in geographic events. The following section presents a more detailed description of the burstiness measure.

Temporal Burstiness: Variance-based Temporal Regularity/Irregularity

The most common method to detect bursty patterns has been to estimate statistical parameters of the heavy-tailed inter-event time distribution $P(\tau)$, where the inter-event time τ is defined as the time interval between two consecutive events, obtained from empirical data. However, there are several alternative statistical measures to complement the limitations of data fitting methods. These limitations include: 1) a difficulty in finding the best fit distribution for noisy data, and 2) a difficulty in comparing and interpreting statistical parameters (e.g., Goh & Barabási, 2008; Karsai *et al.*, 2012).

Alternatively, the burstiness measure proposed by Goh & Barabási (2008) is designed to quantify the burstiness within a range from -1 to 1:

$$B = \frac{\sigma - \mu}{\sigma + \mu} = \frac{r - 1}{r + 1}, \qquad \text{Eq. (1)}$$

where σ and μ denote the standard deviation and the mean of inter-event times, respectively, and $r = \sigma / \mu$ is the coefficient of variation. B has the value of -1 for regular time series as $\sigma = 0$, and 0 for Poissonian or random time series as $\sigma = \mu$. Finally, the value of B approaches 1 for extremely bursty time series as $\sigma \to \infty$ for finite μ.

Given that the number of events is very large, Goh & Barabási (2008)'s burstiness measure can be used to characterize temporal regularity patterns of geographic events that have different frequencies over different regions. Frequencies of wildfires vary by region; boreal regions have much lower fire frequencies than temperate regions, but their temporal regularity patterns could be similarly bursty due to seasonal climate. Conversely, the temporal regularity can vary by region; while two regions may have the same frequency of fire occurrences, one region can have temporally random patterns of wildfire events and the other region can have temporally bursty patterns of wildfire events. In this case, the burstiness measures can provide useful information on temporal regularity/irregularity of geographic events, in addition to the frequency of events.

The burstiness measure can be positioned within the context of ESTDA approaches as shown in Table 2-1. The burstiness measure is capable of characterizing bursty patterns that are not captured by existing ESTDA methods, so it generates another category of ESTDA. Statistical indicators of bursty patterns are only a few and no local indicators have been suggested for examining spatial variations of temporal burstiness (Table 2-1). I see this as an opportunity for unique contributions to research on ESTDA methods. First, a novel burstiness measure is proposed to amend the small sample size effects of Goh & Barabási (2008) in Chapter 3, and local indicators of temporal burstiness are developed in Chapter 4. The resulting methods are applied in a detailed case study of burstiness of jet contrail outbreaks in Chapter 5.

Table 2-1. Position of burstiness measure in ESTDA methods

	Clustered patterns		**Bursty patterns**
	Point Pattern	*Attribute*	***Point Pattern***
Global Indicator	K-function, L-function (density), G-function (nearest neighbors), STIK function	Space-time association: Knox index, Mantel index; Spatial/temporal autocorrelation: Moran's I, Getis and Ord's G	Statistical parameters of heavy-tailed distributions (inter-event times), Goh & Barabási (2008)'s burstiness measure **Kim & Jo (2016)'s burstiness measure (Chapter 3)***
Local Indicator	local (network-constrained) K-function (density), spatial/temporal/ space-time scan statistics (density)	Spatial association: LISA (e.g. local Moran's I)	**NONE** **Local indicators of temporal burstiness (Chapter 4)***

* ESTDA methods that are contributed by this dissertation

Chapter 3

Burstiness Measure for Finite Event Sequences[2]

Introduction

Goh & Barabási (2008)'s burstiness parameter[3], *B*, that characterizes bursty patterns has been widely used due to its simplicity, as mentioned in Chapter 2. We note that the behavior of the burstiness parameter in Eq. (1) may not be robust with respect to *finite-size effects* (i.e., *small-sample size effects*). The maximum value of *B* is strongly limited by the number of events. The behavior of *B* approaching 1 is expected only when the number of events in the time series is sufficiently large or approaching infinite.

However, the numbers of events in data sets are finite for real systems, and they are often very small. So, practically, the value of *B* never reaches 1, although the regularity pattern of events is conceptually at the maximum level of burstiness with the given number of events (i.e. all the events happen at once at the beginning or end of the time period of interest). Moreover, even in the case that the total number of events of a dataset is very large, the number of events or activity level per individual is typically highly skewed. For many cases in human dynamics (e.g., Radicchi, 2009; Jo *et al.*, 2012a; Kim & MacEachren, 2014), the majority of the population has a relatively small number of events. From a geographic perspective, the spatial distribution of events is often heterogeneous; some regions experience the majority of events over time, but

---

[2] This chapter contains an edited version of a paper that has been published as:
Kim, Eun-Kyeong, and Hang-Hyun Jo. "Measuring burstiness for finite event sequences." *Physical Review E* 94, no. 3 (2016): 032311. http://dx.doi.org/10.1103/PhysRevE.94.032311

[3] The term 'burstiness' of the original paper, Kim & Jo (2016), is interpreted as temporal burstiness in this chapter, without changing the term into 'temporal burstiness.'

other regions go through only a few events. Those individuals or regions with low activity have been arbitrarily ignored, or aggregated to form a group of low activity. The approach of excluding or aggregating individuals with low activity could be partly due to the absence of reliable measures of burstiness for finite event sequences.

Thus, the research question that arises is: How can one properly compare one bursty time series with another that shows a similar bursty pattern but has a different number of events? In other words, how can one isolate finite-size effects from the intrinsic temporal features in a time series?

To study finite-size effects on the burstiness parameter, we devise an analytically tractable model to calculate the coefficient of variation of inter-event times for finite event sequences. To simulate various event sequences, our model has two relevant factors: *a bursty period* (Δ) where events actually occur within the entire observation time window and *a lower bound of inter-event time* (τ_{min}) (i.e. the minimum time interval among input inter-event times). By tuning these two control parameters, we can obtain the analytic values of *B* for three reference cases; regular, random, and extremely bursty time series. Then we investigate the strong finite-size effects on *B* for reference cases, enabling us to suggest a novel definition of the burstiness measure that is free from finite-size effects and yet simple.

In this process, we take into account two specific types of *boundary condition* in time to calculate inter-event times from event sequences: a *periodic boundary condition* (PBC) and an *open boundary condition* (OBC). PBCs, also called the *Born–von Karman condition*, in time are chosen when an observed small part of a system can be used to approximate the large or infinite system where the observed tendency is periodically repeated over time so that the system is temporally stationary (Lebecki *et al.*, 2008). OBC does not assume such periodicity. Specifically, for phenomena whose temporal patterns are expected to continue or appear repeatedly over an extended period beyond the observation time window (e.g., a year-long period dataset of the

weather - four seasons), adopting PBC would be appropriate. If patterns for the extended period are uncertain and expected to be temporally non-stationary, choosing OBC would be proper. We consider the case of PBC in time first for simplicity in the following section and then the case of OBC in time later in the 'model with open boundary condition' section.

We demonstrate the advantages of our new burstiness measure by analyzing empirical datasets: geo-tagged Twitter datasets collected in the conterminous United States. Finally, we conclude our work in the last section.

Model with Periodic Boundary Condition

Uniform Case

We first consider an event sequence with n events, each taking place uniformly at random in the time interval [0,d). The events are ordered by their timings, and the timing of the *i*-th event is denoted by t_i for $i = 1,\ldots,n$. Inter-event times are defined as

$$\tau_{i,d} \equiv \begin{cases} d - t_n + t_1 & \text{if} \quad i = 1 \\ t_i - t_{i-1} & \text{if} \quad i \neq 1. \end{cases} \qquad \text{Eq. (2)}$$

By the order statistics (David & Nagaraja, 2003; Kivelä *et al.*, 2012), inter-event time distributions are written as follows:

$$P(\tau_{i,d}) = \begin{cases} \dfrac{(\tau_{1,d}/d)(1-\tau_{1,d}/d\,)^{n-2}}{B(2,n-1)d} & \text{if} \quad i = 1 \\ \dfrac{(1-\tau_{i,d}/d\,)^{n-1}}{B(1,n)d} & \text{if} \quad i \neq 1, \end{cases} \qquad \text{Eq. (3)}$$

where $B(n,m)$ denotes the beta function,

$$B(n,m) = \int_0^1 z^{n-1}(1-z)^{m-1}dz = \frac{(n-1)!\,(m-1)!}{(n+m-1)!}. \qquad \text{Eq. (4)}$$

Expectation values of $\tau_{i,d}$ and $\tau_{i,d}^2$ are obtained as

$$\langle \tau_{i,d} \rangle = \begin{cases} \dfrac{2d}{n+1} & \text{if} \quad i = 1 \\ \dfrac{d}{n+1} & \text{if} \quad i \neq 1 \end{cases} \qquad \text{Eq. (5)}$$

and

$$\langle \tau_{i,d}^2 \rangle = \begin{cases} \dfrac{6d^2}{(n+1)(n+2)} & \text{if} \quad i = 1 \\ \dfrac{2d^2}{(n+1)(n+2)} & \text{if} \quad i \neq 1. \end{cases} \qquad \text{Eq. (6)}$$

We have assumed that $\tau_{i,d}$s are independent of each other, although they are not independent but satisfy the condition $\sum_{i=1}^{n} \tau_{i,d} = d$. Instead we find on average that

$$\sum_{i=1}^{n} \langle \tau_{i,d} \rangle = \langle \tau_{1,d} \rangle + (n-1)\langle \tau_{i\neq 1,d} \rangle = d. \qquad \text{Eq. (7)}$$

This issue will be discussed later in the next subsection.

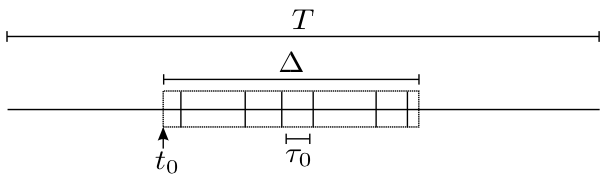


Figure 3-1. **Schematic diagram of the localized model.** *n* events are localized in the period Δ beginning at t_0 in [0,*T*), and they are separated from each other at least by τ_0.

Localized Model

We now consider the general case that all events are localized in the interval $[t_0, t_0 + \Delta)$ with $t_0 \geq 0$ and $t_0 + \Delta < T$, indicating that events do not take place in the intervals $[0, t_0)$ and $[t_0 + \Delta, T)$, as depicted in Figure 3-1. A similar model has been studied in a different context (Perotti et al., 2014). The localization parameter Δ is introduced to simulate the bursty limit for $\Delta \ll T$. Because we use periodic boundary condition, t_0 can be ignored. In addition, the lower bound of inter-event time, τ_0, is introduced, implying that events must be separated from each other at least by τ_0. Accordingly, it is assumed that

$$(n-1)\tau_0 \le \Delta \le T - \tau_0, \qquad \text{Eq. (8)}$$

leading to $\tau_0 \le \frac{T}{n}$. If $\tau_0 = \frac{T}{n}$, one gets the regular time series.

Then, we use definitions in Eq. (2) with *d* being replaced by $\delta \equiv \Delta - (n-1)\,\tau_0$ to define inter-event times as

$$\tau_i \equiv \begin{cases} \tau_{1,\delta} + T - \Delta & \text{if} \quad i = 1 \\ \tau_{i,\delta} & \text{if} \quad i \neq 1. \end{cases} \qquad \text{Eq. (9)}$$

Using Eq. (3), we get

$$\langle \tau_i \rangle = \begin{cases} T - \dfrac{n-1}{n+1}(\Delta + 2\tau_0) & \text{if} \quad i = 1 \\ \dfrac{\Delta + 2\tau_0}{n+1} & \text{if} \quad i \neq 1. \end{cases} \qquad \text{Eq. (10)}$$

and

$$\langle \tau_i^2 \rangle = \begin{cases} \dfrac{6\delta^2}{(n+1)(n+2)} + \dfrac{4(T-\Delta)\delta}{n+1} + (T-\Delta)^2 & \text{if} \quad i = 1 \\ \dfrac{2\delta^2}{(n+1)(n+2)} + \dfrac{2\tau_0\delta}{n+1} + \tau_0^2 & \text{if} \quad i \neq 1. \end{cases} \qquad \text{Eq. (11)}$$

Then we calculate the mean μ_n and the variance σ_n^2 of inter-event times to get the coefficient of variation $r_n = \frac{\sigma_n}{\mu_n}$:

$$\mu_n = \frac{1}{n}[\langle \tau_1 \rangle + (n-1)\langle \tau_{i\neq 1} \rangle], \qquad \text{Eq. (12)}$$

$$\sigma_n^2 = \frac{1}{n}\left[\langle \tau_1^2 \rangle + (n-1)\langle \tau_{i\neq 1}^2 \rangle\right] - \mu_n^2, \qquad \text{Eq. (13)}$$

$$r_n(x,y) = \sqrt{\frac{(n-1)[1 + n(1-x)^2 + n(n+1)y^2 - 2n(2-x)y]}{n+1}}, \qquad \text{Eq. (14)}$$

Here we have defined

$$x \equiv \frac{\Delta}{T}, \;\; y \equiv \frac{\tau_0}{T}, \qquad \text{Eq. (15)}$$

Satisfying the condition that

$$(n-1)y \le x \le 1 - y, \;\; y \le \frac{1}{n}. \qquad \text{Eq. (16)}$$

It is straightforward to show that $r_n(x, y)$ is a non-increasing function of *x* and *y*, respectively.

To study the strong finite-size effects in event sequences, we define the burstiness parameter using *x* and *y* as follows:

$$B_n(x, y) \equiv \frac{r_n(x, y) - 1}{r_n(x, y) + 1}. \quad \text{Eq. (17)}$$

We discuss three reference cases. Firstly, the regular time series means that all inter-event times are the same as μ_n, implying that $x = 1 - \frac{1}{n}$ and $y = \frac{1}{n}$. Since $r_n = 0$ independent of n, we get

$$B_n\left(1 - \frac{1}{n}, \frac{1}{n}\right) = -1. \quad \text{Eq. (18)}$$

Secondly, the Poissonian or random time series corresponds to the case with $\Delta = T$ and $\tau_0 = 0$, i.e., $x = 1$ and $y = 0$, leading to $r_n = \sqrt{\frac{n-1}{n+1}}$. We get

$$B_n(1,0) = \frac{\sqrt{n-1} - \sqrt{n+1}}{\sqrt{n-1} + \sqrt{n+1}}. \quad \text{Eq. (19)}$$

Note that $B_1(1,0) = -1$ and that $B_n(1,0)$ is always negative but approaches 0 as *n* increases, i.e., $B_n(1,0) \approx -\frac{1}{2n}$ for large *n*, as shown in Figure 3-2(a). Since this result is based on the assumption of independence of τ_is, we test our result by comparing it to numerical values of burstiness parameter. For this, we generate 10^5 event sequences for each *n* to obtain the burstiness parameter as depicted in Figure 3-2(a). We find that the deviation of our analytic results from the simulations is negligible.

Thirdly, the extremely bursty time series corresponds to the case that all events occur asymptotically at the same time, i.e., $x = y = 0$, leading to $r_n = \sqrt{n-1}$. Thus, one gets

$$B_n(0,0) = \frac{\sqrt{n-1} - 1}{\sqrt{n-1} + 1.} \quad \text{Eq. (20)}$$

Note that $B_1(0,0) = -1$ and $B_2(0,0) = 0$. $B_n(0,0)$ becomes positive for $n \geq 3$, and then approaches 1 as *n* increases, i.e., $B_n(0,0) \approx 1 - \frac{2}{\sqrt{n}}$, as shown in Figure 3-2(a). The finite-size

effect turns out to be the strongest for the bursty case. It is because $B = 1$ is realized only for infinitely many inter-event times. The strong dependence of B_n on the number of events n could introduce a serious finite-size effect in particular for the bursty case, i.e., for most empirical datasets.

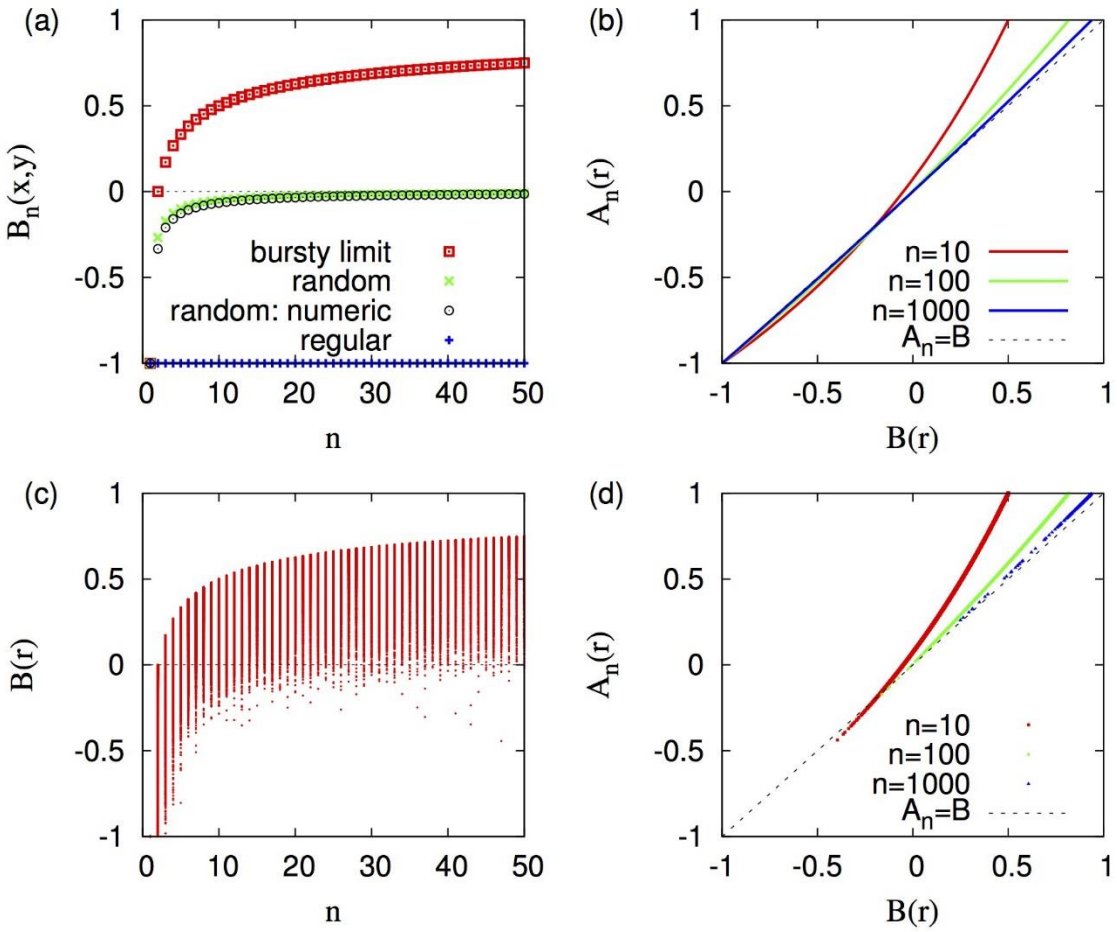


Figure 3-2. Analytic and empirical results of Goh & Barabási's (2008) burstiness parameter and our novel burstiness measure as a function of the number of events, *n*, for three reference cases of temporal patterns: regular, random, and extremely bursty time series. (a) Analytic results of $B_n(x, y)$ for three reference cases: Eq. (18) for regular time series, Eq. (19) for random time series, and Eq. (20) for the bursty limit. Numerical results for the random case are plotted for comparison to the analytic results. (b) Comparison of the novel burstiness measure $A_n(r)$ in Eq. (24) to the original burstiness parameter $B(r)$ in Eq. (1) for several values of n. (c) Scatter plot of $B(r)$ for individual Twitter users. (d) The same as (b) but using Twitter dataset.

General Formula of the Novel Burstiness Measure

To fix the finite-size effects in the original burstiness parameter while maintaining the simplicity of the original burstiness measure $B(r)$, we suggest a novel definition of the burstiness measure, denoted by $A_n(r)$. This new burstiness measure is assumed to be a function of $r = \frac{\sigma_n}{\mu_n}$, and *r* is a function of inter-event times and the number of events, *n* (i.e. the sample size of datasets). Because $B(r)$ was originally defined as $\frac{r-1}{r+1}$, we derive a general formula of the novel burstiness measure $A(r) = \frac{ar-b}{r+c}$ with coefficients *a, b, c* when the conditions for reference cases are given as follows:

$$\begin{aligned} A(r_-) &= -1, \\ A(r_0) &= 0, \\ A(r_+) &= 1. \end{aligned} \qquad \text{Eq. (21)}$$

Here, r_-, r_0, and r_+ denote the coefficients of variation for reference cases of regular, random, and extremely bursty time series, respectively. Using these conditions, one gets

$$A(r) = \frac{(r_+ - r_-)(r - r_0)}{(r_+ + r_- - 2r_0)r + (r_+ + r_-)r_0 - 2r_+r_-}. \qquad \text{Eq. (22)}$$

Novel Definition of Burstiness Measure

For a periodic boundary condition, a novel burstiness measure $A_n(r)$ must satisfy the following conditions:

$$\begin{aligned} A_n(0) &= -1, \\ A_n\left(\sqrt{\frac{n-1}{n+1}}\right) &= 0, \\ A_n\left(\sqrt{n-1}\right) &= 1, \end{aligned} \qquad \text{Eq. (23)}$$

which correspond to the cases of regular, random, and extremely bursty time series, respectively. According to the logic of the general formula stated above, we assume that $A_n(r) = \frac{a_n r - b_n}{r + c_n}$ with coefficients a_n, b_n, and c_n. Using a general formula of Eq. (22), we get

$$A_n(r) = \frac{\sqrt{n+1}r - \sqrt{n-1}}{\left(\sqrt{n+1} - 2\right)r + \sqrt{n-1}}, \qquad \text{Eq. (24)}$$

for $0 \le r \le \sqrt{n-1}$. Our novel burstiness measure A_n has no longer an upper bound due to the finite *n*, as depicted in Figure 3-2(b), where the curves for different *n*s are described by

$$A_n(r) = \frac{\sqrt{n+1} - \sqrt{n-1} + (\sqrt{n+1} + \sqrt{n-1})B(r)}{\sqrt{n+1} + \sqrt{n-1} - 2 + (\sqrt{n+1} - \sqrt{n-1} - 2)B(r)}. \qquad \text{Eq. (25)}$$

Then let us consider two event sequences with a different number of events, *n*s, but with the most bursty patterns of the given number of events, meaning that all the events occur at the beginning or end of the time window. The original burstiness parameter *B* has different values, while A_n gives the same value, 1. Thus, A_n can characterize bursty patterns more consistently by taking into account finite-size effects. So, our novel burstiness measure can be used to compare the burstiness of event sequences with different numbers of events (e.g., event sequences from region A and region B).

Empirical application: geo-tagged Twitter data

For practical applications, we show how A_n can be used to quantify the burstiness in the empirical dataset without finite-size effects. For a given event sequence of *n* events, one can calculate the coefficient of variation of inter-event times, denoted by $\tilde{r}$, to get the value of $A_n(\tilde{r})$ for the given event sequence. For the demonstration, we analyze a large-scale Twitter dataset collected for a year from October 1, 2012 to September 30, 2013 throughout the United States of America. The dataset was originally collected only for tweets with geographical

information, while we exploit only the temporal information of tweets in our work. After cleaning, the dataset contains approximately 698 million tweets posted by about 5.5 million user accounts. As shown in Figure 3-2(c), the original burstiness parameter for individual Twitter users in the dataset clearly has the upper bound for small values of *n*. Such upper bound is removed or corrected in A_n obtained from the same dataset in Figure 3-2(d).

Effect due to Minimum Inter-Event Times

In addition to the number of events in a given event sequence, one can also exploit more information from the sequence, such as the minimum inter-event time, $\tau_{min} = \min\{\tau_i\}$. The role of the minimum inter-event time has been discussed in various contexts. For example, it can be related to the refractory period of neurons, which may limit the response time of neuronal systems to external stimuli. The minimum inter-event time has been found to play a crucial role in spreading dynamics in complex systems (Jo *et al.*, 2014). Since different systems or different individuals in the same system may have different values of minimum inter-event time, its effect must be carefully investigated, particularly for comparing the temporal properties of different systems or different individuals in the same system.

We consider three reference cases for given τ_{min} or $\tilde{y} \equiv \frac{\tau_{min}}{T}$. The bursty limit case is achieved by maximizing $r_n(x, \tilde{y})$ with $x = (n-1)\,\tilde{y}$, see Eq. (16). We get the random case by setting $x = 1 - \tilde{y}$ for $r_n(x, \tilde{y})$. Finally, as for the regular case we consider a specific time series that one inter-event time is τ_{min}, while all other $n-1$ inter-event times are the same as $\frac{T-\tau_{min}}{n-1}$ for $n \geq 2$. We calculate the coefficient of variation of inter-event times as $r_n^*(\tilde{y}) = \frac{1-n\tilde{y}}{\sqrt{n-1}}$. Other regular cases can be considered. For example, *k* inter-event times are the same as τ_{min}, while other $n-k$ inter-event times are the same as $\frac{T-\tau_{min}}{n-k}$ for $n > k$. Then we get the coefficient of

variation as $\sqrt{\frac{k}{n-k}}(1-n\tilde{y})$, indicating that the minimal coefficient of variation is obtained when $k = 1$.

Note that τ_0 in the localized model is the possible lower bound of inter-event time, while τ_{min} is one of inter-event times in the given event sequence, hence it can be larger than τ_0. In sum, we get

$$r_n^*(\tilde{y}) = \frac{1-n\tilde{y}}{\sqrt{n-1}}, \quad \text{Eq. (26)}$$

$$r_n(1-\tilde{y},\tilde{y}) = \sqrt{\frac{n-1}{n+1}}(1-n\tilde{y}), \quad \text{Eq. (27)}$$

$$r_n((n-1)\tilde{y},\tilde{y}) = \sqrt{n-1}(1-n\tilde{y}), \quad \text{Eq. (28)}$$

for the cases of regular, random, and extremely bursty time series, respectively. Since the specific time series for the regular case does not fit within the localized model, we need to impose the condition that $r_n^*(\tilde{y}) < r_n(1-\tilde{y},\tilde{y})$, i.e., $n > 3$. One can calculate $B(r) = \frac{r-1}{r+1}$ with above results of rs for reference cases to find the strong effects due to the finite minimum inter-event time and the finite size of events. For example, the analytic curves for $n = 10$ are depicted in Figure 3-3(a), where we also plot the empirical values of B for individual Twitter users with $n = 10$ according to their own $\tilde{y}$.

Then, one can find the functional form of the burstiness measure $A_{n,\tilde{y}}(r) = \frac{a_{n,\tilde{y}}\, r - b_{n,\tilde{y}}}{r + c_{n,\tilde{y}}}$ with coefficients $a_{n,\tilde{y}}$, $b_{n,\tilde{y}}$, and $c_{n,\tilde{y}}$, satisfying

$$\begin{aligned} A_{n,\tilde{y}}[r_n^*(\tilde{y})] &= -1, \\ A_{n,\tilde{y}}[r_n(1-\tilde{y},\tilde{y})] &= 0, \\ A_{n,\tilde{y}}[r_n((n-1)\tilde{y},\tilde{y})] &= 1. \end{aligned} \quad \text{Eq. (29)}$$

Using a general formula of Eq. (22), we obtain the complete form as follows:

$$A_{n,\tilde{y}}(r) = \frac{(n-2)[\sqrt{n+1}\,r - \sqrt{n-1}(1-n\tilde{y})]}{[n\sqrt{n+1} - 2(n-1)]r + \sqrt{n-1}(n-2\sqrt{n+1})(1-n\tilde{y})} \quad \text{Eq. (30)}$$

for $\frac{1-n\tilde{y}}{\sqrt{n-1}} \leq r \leq \sqrt{n-1}(1-n\tilde{y})$. In Figure 3-3(b), we compare $A_{n,\tilde{y}}$ against $B(r)$ for the entire range of r, with empirical results for Twitter users with n = 10. We remark that for given event sequences, one can exploit even more information from those sequences, depending on which factors are to be controlled.

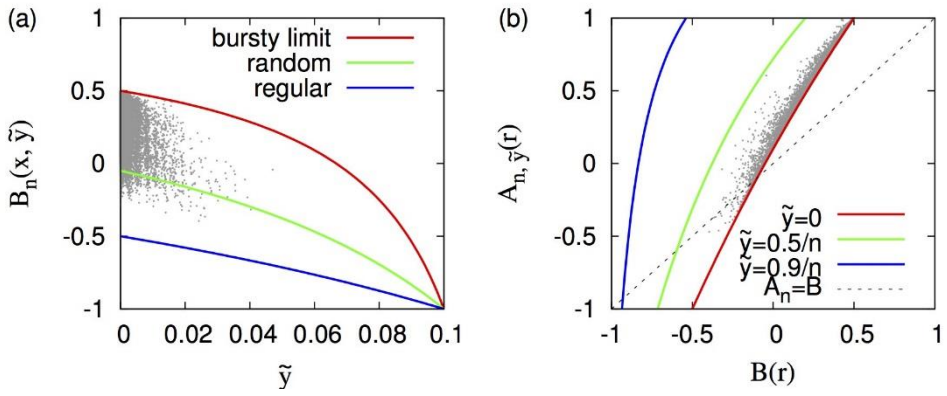


Figure 3-3. Analytic and empirical results of Goh & Barabási's (2008) burstiness parameter and our novel burstiness measure as a function of the ratio, $\tilde{y}$, of the minimum inter-event time, τ_{min}, to the entire time window, T, for three reference cases of temporal patterns: regular, random, and extremely bursty time series. (a) Analytic results of $B_n(x, \tilde{y})$ for bursty limit and random cases, and that of $B_n[r_n^*(\tilde{y})]$ for regular case, when n = 10. (b) Comparison of the novel burstiness measure $A_{n,\tilde{y}}(r)$ in Eq. (30) to the original burstiness parameter $B(r)$ in Eq. (1) for several values of $\tilde{y}$ for n = 10. Each gray dot in both panels corresponds to each individual Twitter user with n = 10.

Model with Open Boundary Condition

Uniform Case

Here we obtain the analytic results in the case of the open boundary condition as the periodic boundary condition may not apply to some empirical datasets.

We first consider an event sequence with *n* events, each taking place uniformly at random in the interval [0,*d*). The events are ordered by their timings, and the timing of the *i*-th event is denoted by t_i for $i = 1, \cdots, n$. Then, inter-event times are defined as

$$\tau_{i,d} = \begin{cases} t_1 & \text{if} \quad i = 1, \\ t_i - t_{i-1} & \text{if} \quad i = 2, \dots, n, \\ d - t_n & \text{if} \quad i = n+1 \end{cases} \qquad \text{Eq. (31)}$$

Here the time intervals from $t = 0$ to $t = t_1$ and from $t = t_n$ to $t = d$ have been taken as "inter-event times", although there is no event at $t = 0$ and $t = d$. The case when these time intervals are discarded is left for future work. By the order statistics (David & Nagaraja, 2003), inter-event time distributions read as follows:

$$P(\tau_{i,d}) = \frac{(1 - \tau_{i,d}/d)^{n-1}}{B(1,n)d} \qquad \text{Eq. (32)}$$

for all *i*. Expectation values of $\tau_{i,d}$ and $\tau_{i,d}^2$ are obtained as

$$\langle \tau_{i,d} \rangle = \frac{d}{n+1}, \qquad \text{Eq. (33)}$$

$$\langle \tau_{i,d}^2 \rangle = \frac{2d^2}{(n+1)(n+2)}. \qquad \text{Eq. (34)}$$

Here the inter-event times satisfy $\sum_{i=1}^{n+1} \langle \tau_{i,d} \rangle = d$. Thus, one gets

$$\mu_n = \frac{d}{n+1}, \qquad \text{Eq. (35)}$$

$$\sigma_n^2 = \frac{nd^2}{(n+1)^2(n+2)} \qquad \text{Eq. (36)}$$

Localized Model

We then consider the general case that all events are localized in the interval $[t_0, t_0 + \Delta)$ with $t_0 \geq 0$ and $t_0 + \Delta < T$, as depicted in Figure 3-1. In addition to the condition in Eq. (8), we have one more condition for t_0 that

$$\tau_0 \leq t_0 \leq T - \Delta - \tau_0. \qquad \text{Eq. (37)}$$

We use definitions in Eq. (31) with d being replaced by $\delta \equiv \Delta - (n-1)\,\tau_0$ to define inter-event times as

$$\tau_i = \begin{cases} \tau_{1,\delta} + t_0 & \text{if} \quad i = 1 \\ \tau_{i,\delta} + \tau_0 & \text{if} \quad i = 2, \dots, n \\ \tau_{n+1,\delta} + T - \Delta - t_0 & \text{if} \quad i = n+1. \end{cases} \qquad \text{Eq. (38)}$$

Using Eq. (32), one gets

$$\langle \tau_i \rangle = \begin{cases} \frac{\delta}{n+1} + t_0 & \text{if} \quad i = 1 \\ \frac{\delta}{n+1} + \tau_0 & \text{if} \quad i = 2, \dots, n \\ \frac{\delta}{n+1} + T - \Delta - t_0 & \text{if} \quad i = n+1 \end{cases} \qquad \text{Eq. (39)}$$

and

$$\langle \tau_i^2 \rangle = \begin{cases} \frac{2\delta^2}{(n+1)(n+2)} + \frac{2\,t_0\delta}{n+1} + t_0^2 & \text{if} \quad i = 1 \\ \frac{2\delta^2}{(n+1)(n+2)} + \frac{2\,\tau_0\delta}{n+1} + \tau_0^2 & \text{if} \quad i = 2, \dots, n \\ \frac{2\delta^2}{(n+1)(n+2)} + \frac{2(T-\Delta-t_0)\delta}{n+1} + (T - \Delta - t_0)^2 & \text{if} \quad i = n+1. \end{cases} \qquad \text{Eq. (40)}$$

The mean and the variance of inter-event times are obtained by

$$\mu_n = \frac{1}{n+1}\left[\langle \tau_1 \rangle + (n-1)\langle \tau_{i\neq 1,n+1} \rangle + \langle \tau_{n+1} \rangle\right], \qquad \text{Eq. (41)}$$

$$\sigma_n^2 = \frac{1}{n+1}\left[\langle\tau_1^2\rangle + (n-1)\langle\tau_{i\neq 1,n+1}^2\rangle + \langle\tau_{n+1}^2\rangle\right] - \mu_n^2. \quad \text{Eq. (42)}$$

Then we calculate the coefficient of variation of inter-event times as follows:

$$r_n(x,y,z) = \sqrt{\frac{2[x-(n-1)y][n+2-x+(n-1)y]}{n+2} + (n+1)[(1-x-z)^2+z^2+(n-1)y^2] - 1.} \quad \text{Eq. (43)}$$

Here we have defined

$$x \equiv \frac{\Delta}{T}, y \equiv \frac{\tau_0}{T}, z \equiv \frac{t_0}{T}, \quad \text{Eq. (44)}$$

satisfying the conditions that

$$(n-1)y \leq x \leq 1-y, \qquad y \leq z \leq 1-x-y, \qquad y \leq \frac{1}{n+1}. \quad \text{Eq. (45)}$$

The burstiness parameter for open boundary condition is obtained as $B_n(x,y,z) = \frac{r_n(x,y,z)-1}{r_n(x,y,z)+1}$.

We discuss the reference cases. The regular time series may correspond to the case of $r_n = 0$. It implies that all inter-event times, including τ_1 and τ_{n+1}, must be the same, i.e. $x = \frac{n-1}{n+1}$, and $y = z = \frac{1}{n+1}$. The random time series is obtained when $x = 1$ and $y = z = 0$, where $z = 0$ is needed to avoid any memory effects. Finally, the bursty limit for maximizing r_n implies the situation when all events occur at the same time, i.e., $x = y = 0$. In order to get the maximum value of r_n, we choose $z = 0$. In sum, one gets

$$\begin{aligned} B_n\left(\frac{n-1}{n+1}, \frac{1}{n+1}, \frac{1}{n+1}\right) &= -1, \\ B_n(1,0,0) &= \frac{\sqrt{n}-\sqrt{n+2}}{\sqrt{n}+\sqrt{n+2}}, \\ B_n(0,0,0) &= \frac{\sqrt{n}-1}{\sqrt{n}+1}. \end{aligned} \quad \text{Eq. (46)}$$

These results can be also obtained from those for the periodic boundary condition by replacing n by $n+1$, because we have one more inter-event time under the open boundary condition.

Novel Definition of Burstiness Measure

The novel definition of the burstiness measure for the open boundary condition reads

$$A_n(r) = \frac{\sqrt{n+2}r - \sqrt{n}}{\left(\sqrt{n+2} - 2\right)r + \sqrt{n}} \qquad \text{Eq. (47)}$$

for $0 \leq r \leq \sqrt{n}$.

Effect due to Minimum Inter-Event Times

We study the effect of minimum inter-event time, τ_{min} or $\tilde{y} \equiv \frac{\tau_{min}}{T}$, on the burstiness parameter. The bursty limit case is obtained when $x = (n-1)\tilde{y}$ and $z = \tilde{y}$. The random case is obtained when $x = 1 - \tilde{y}$ and $z = \tilde{y}$. For these two cases, we use Eq. (43). As for the regular case we consider a specific time series that one interevent time is τ_{min}, while all other n inter-event times are the same as $\frac{T-\tau_{min}}{T}$. Here z can be either $\tilde{y}$ or $\frac{1-\tilde{y}}{n}$, leading to the same result for the coefficient of variation of inter-event times as $r_n^*(\tilde{y}) = \frac{1-(n+1)\tilde{y}}{\sqrt{n}}$. Then, the calculation of novel burstiness measure $A_{n,\tilde{y}}(r)$ for the open boundary condition is straightforward using the following conditions:

$$\begin{aligned} A_{n,\tilde{y}}[r_n^*(\tilde{y})] &= -1, \\ A_{n,\tilde{y}}[r_n(1-\tilde{y},\tilde{y},\tilde{y})] &= 0, \\ A_{n,\tilde{y}}\left[r_n\left((n-1)\tilde{y},\tilde{y},\tilde{y}\right)\right] &= 1. \end{aligned} \qquad \text{Eq. (48)}$$

Conclusion

Bursts have been mostly characterized by heavy-tailed inter-event time distributions, or more simply by the burstiness parameter $B(r) = \frac{r-1}{r+1}$, where the coefficient of variation, r, denotes the ratio of standard deviation to mean of inter-event times. B has the value of -1 for regular time series as $r = 0$, and it is 0 for Poissonian or random time series as $r = 1$. Finally, $B(r)$ approaches 1 for extremely bursty time series as $r \to \infty$. Despite its successful applications, $B(r)$ turns out to be strongly affected by the finite size of event sequence, in particular when the event sequence is bursty.

To get the analytic limits of $B(r)$ for the given n in the reference cases, i.e., regular, random, and extremely bursty cases, we devise and study an analytically tractable model with n events. Then we suggest a novel definition of the burstiness measure that is free from finite-size effects and yet simple, denoted by $A_n(r)$; $A_n(r)$ has no upper bound due to the finite n. If two event sequences have the same r but different ns, the original burstiness parameter cannot distinguish which event sequence is burstier than the other, while our novel burstiness measure can do so. Thus, one can isolate the effect due to the finite size of event sequences. By analyzing a large-scale Twitter dataset, we show that $B(r)$ clearly has the upper bound due to the finite n for individual Twitter users, and that $A_n(r)$ no longer has such an upper bound.

This study advances a method of measuring burstiness, but there is still more work to address in the future. Above all, for our localized model, only one burst with period Δ is assumed to exist. This assumption is sufficient to simulate the above reference cases. It is because even when considering multiple bursts in the event sequence, the above reference cases will be the same as those in our localized model. In any case, the extension of our model to incorporate multiple bursts could be important but left for future work.

We can exploit more information other than *n* from the given event sequences, such as the minimum inter-event time. The minimum inter-event time is important to understand the intrinsic timescale of the system, e.g., the refractory period of neurons. For fixing the effects due to the finite minimum inter-event time and the finite size of event sequence, we suggest another burstiness measure, $A_{n,\tilde{y}}(r)$, with $\tilde{y}$ denoting the ratio of minimum inter-event time to the whole period. Using the $A_{n,\tilde{y}}(r)$, we can separate the intrinsic bursty dynamics from the effect due to the minimal timescale in the system. Even more information in the given event sequences can be exploited, which we leave for future work.

We have considered only the burstiness measure, while other quantities like memory coefficient (e.g., Goh & Barabási, 2008) and bursty train distribution (e.g., Karsai *et al.*, 2012) for higher order temporal correlations could be also investigated from the same perspective. In addition, our analytically tractable model can be extended to incorporate another kind of information called context, e.g., as studied in terms of contextual bursts (Jo *et al.*, 2013).

As our novel definition of burstiness $A_n(r)$ takes the sample size *n* into account, $A_n(r)$ is not a parameter but it could be interpreted as an unbiased estimator for $B(r)$ in the statistical sense. We however call $A_n(r)$ a measure instead of a parameter[4] mainly because it has been proposed as the descriptive statistic rather than as an optimal unbiased estimator of $B(r)$. Note that finding an unbiased estimator for the coefficient of variation *r* is not trivial, and it is beyond the scope of our current work (Sokal & Braumann, 1980; Breunig, 2001; Forkman, 2009; Albrecher *et al.*, 2010; Banik *et al.*, 2012; Jayakumar & Sulthan, 2015; Kivelä & Porter, 2015). Thus, statistical evaluation of the error of our novel definition $A_n(r)$ is left for future work.

[4] In statistics, the parameter is a value that describes an entire population.

Chapter 4

Local Indicators of Temporal Burstiness for Spatiotemporal Event Analysis

Introduction

A bursty pattern represents a temporal pattern of events. The burstiness measure as an ESTDA statistic is one of the simplest statistical indicators that explore such patterns. The novel burstiness measure proposed in the previous chapter addresses small sample size effects. The main characteristic of the burstiness measure is that it uses the variance of inter-event times normalized by the mean inter-event time. This characteristic means that no matter how many events occur over time, the temporal regularity/irregularity (i.e., burstiness) can be measured by the variability of inter-event times (see Table 4-1), which I call 'frequency-invariant temporal regularity/irregularity'.

The novel burstiness measure from the previous chapter can be useful to characterize temporal regularity patterns of geographic events that have different frequencies over different regions. Specifically, in Table 4-1, let us assume that sequences (a), (c), and (e) represent each geographic region. The patterns (a), (c), and (e) have the same frequency, but they all have different temporal regularity patterns. That is, while the frequency cannot be an indicator to distinguish those three patterns, the temporal burstiness measure provides additional information on the temporal regularity of events.

However, spatial constraints were not explicitly integrated into the novel temporal burstiness measure as introduced in the previous chapter. It is well-known that spatial scale and data aggregation choices are important factors in quantitative analyses of geographic phenomena (Openshaw, 1984; Rodrigues & Tenedório, 2016). From the perspective of these choices, there

are two high-level types of ESTDA statistics: global and local indicators. While global indicators yield a statistical summary for an entire region (e.g., a whole country), local indicators summarize the area of each sub-region (partition) and its corresponding neighborhood (e.g., states or counties as opposed to the whole country); this allows observation of spatial variations of statistical indicators (Anselin, 1995). In applying both global and local ESTDA methods to spatial data, the shape and scale of spatial containers (often referred to as aggregation units) can influence the result of spatial analysis, so it is critical to adopt an appropriate spatial container system for the phenomenon of interest (Openshaw, 1984). The impact of aggregation unit choice on analysis outcomes is known as the *modifiable areal unit problem* (MAUP) (Openshaw, 1984). This issue also affects the application of the temporal burstiness measure, as inter-event times—as an input variable of the temporal burstiness measure—are calculated for data points within the same spatial region. So, inter-event times change according to a spatial aggregation unit of input data points. The effect is explained in more detail in the following section. Hence, there is a need for guidelines for setting up the spatial aggregation unit as a spatial container of events.

The goal of this chapter is to extend the burstiness measure developed in Chapter 3 to local indicators as ESTDA statistics. Specific research objectives of this chapter are as follows:

(1) To introduce spatial containers, suggest possible options of spatial containers to be explicitly integrated with the temporal burstiness measure, and discuss appropriate spatial containers for phenomena of different types;

(2-1) To develop a set of local indicators of temporal burstiness (LITB) as ESTDA statistics to explore geographically local variations in temporal burstiness;

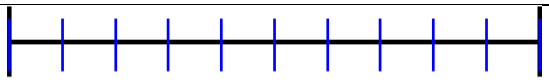
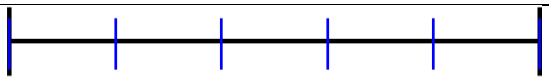
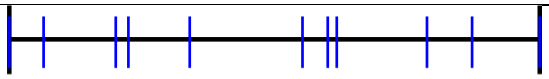
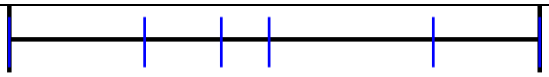
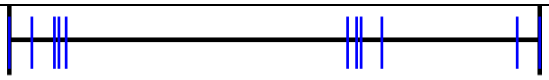
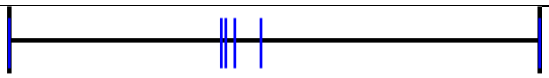
(2-2) To propose a statistical significance testing method to assess whether the burstiness measure produces statistically significant results;

(3) To design algorithms for calculating the temporal burstiness measures from event sequence data at both global and local spatial scales with Python scripts;

(4) To provide a proof-of-concept by applying LITB to the analysis of wildfire events in California, USA, compared to a simple exploratory descriptive statistic (i.e., frequency).

The four sections that follow are aligned with these objectives, in the order above.

Table 4-1. Temporal Burstiness: frequency-invariant temporal regularity/irregularity.

Temporal patterns of events	Frequency	Burstiness measure	Regularity pattern
(a) 	9	-1	completely regular
(b) 	4	-1	completely regular
(c) 	9	0	completely random
(d) 	4	0	completely random
(e) 	9	0.8	bursty
(f) 	4	0.8	bursty

Consideration of Spatial Containers for the Temporal Burstiness Measure

The temporal burstiness measure indicates the frequency-invariant temporal regularity of events at a global scale of time. The burstiness measure uses inter-event times as input data. The burstiness measure uses inter-event times as input data. From a spatial perspective, the temporal burstiness can be measured at both global and local scale of space; at a global scale, it is measured with inter-event times obtained from events over the entire study area; at a local scale, it is measured with those obtained from events in the local area. For some datasets (e.g., landing times of incoming airplanes at the University Park airport), spatial dimensions do not matter in calculation of the temporal burstiness.

For spatiotemporal events, spatial dimensions are often critical to measuring the temporal burstiness of events that occur throughout various locations. A fundamental reason for measuring the local temporal burstiness is that a spatial variation of the temporal burstiness potentially exists. That is, the temporal burstiness at a global scale is not the same as that at a local scale. Another reason is that spatiotemporal events far from each other have little interdependence in comparison to those close to one another, known as Tobler's (1970) first law in Geography. For instance, a wildfire in the Amazonian forest is barely associated with one in the Alaskan boreal forest, while wildfire events in southern California, USA are more likely to interact with each other. Thus, it is less likely useful to measure the temporal burstiness of events contained in the spatial extent where spatial distances of subsequent events are too far to interact with each other.

Along with the emphasis on the importance of considering local interactions among events, in complex systems where bursty behaviors of a phenomenon are often found, local interactions in space and time can lead to a global pattern, called *emergence* (Miller, 2004). Identifying the local interactions is essential to understanding and predicting a global behavior of the system. In this sense, measuring local temporal burstiness can enhance understanding of complex systems better. To ensure meaningful interpretations of temporal burstiness and allow exploration of a spatial variation of temporal burstiness, it is important to introduce the concept of a *spatial container* that narrows down a spatial extent of input data to a proper geographically local area.

Depending on the choice of spatial container in terms of its size and shape, the inter-event times can be very different, even with the same dataset of events (see Figure 4-1). In Figure 4-1a, inter-event times are obtained from events throughout the entire space; in Figure 4-1b, inter-event times are obtained from events belonging to three spatial containers (i.e., subsets of input data) depicted as green, blue, and yellow; inter-event times are shorter in Figure 4-1a than those in Figure 4-1b on average. Differences in calculation of the burstiness measure for the whole or

subset imply that considerations of spatial partitioning of events may lead to different values of the temporal burstiness measure. Events that are temporally bursty at a global spatial scale can be temporally random at a local scale.

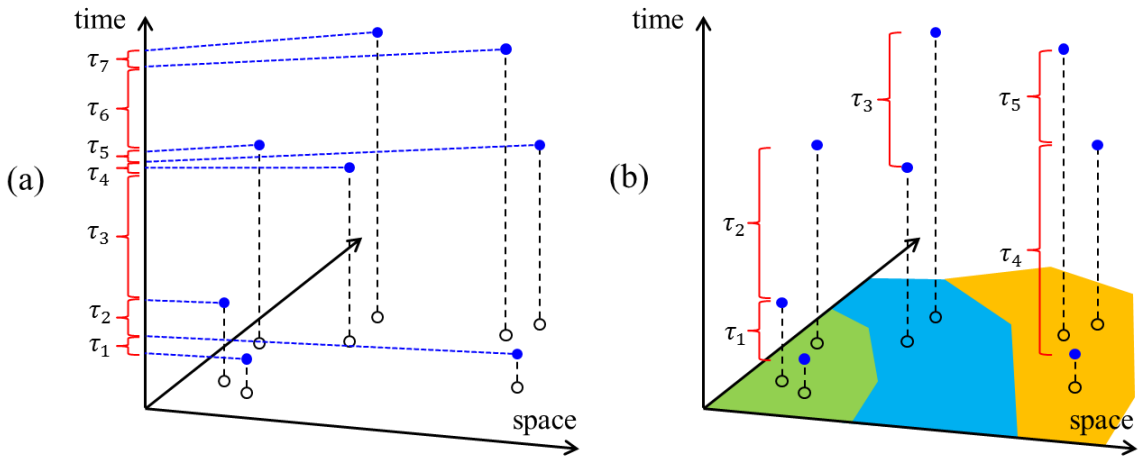
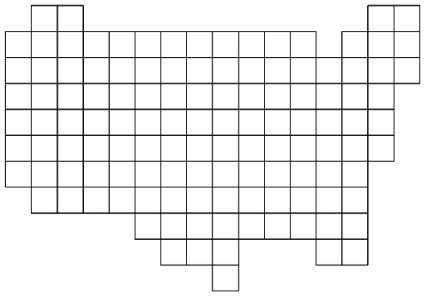
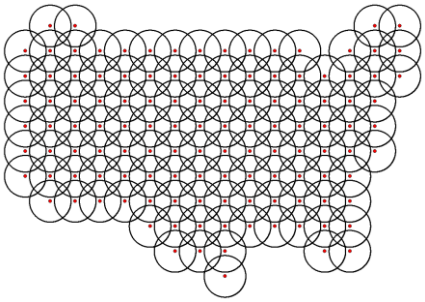
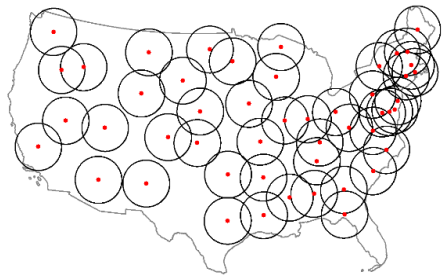


Figure 4-1. Inter-event times obtained by different methods. (a) without geographic context, and (b) with geographic context.

Then, how should one choose the size and form of spatial containers? There is no perfect solution for the selection of spatial aggregation units, but several studies on MAUP have compared available options for spatial aggregation schemas (e.g., Rodrigues & Tenedório, 2016; Nelson & Brewer, 2017).

In this study, I categorize different types of spatial containers: 1) using contiguous spatial aggregation units obtained from spatial partitioning, 2) using spatial buffers with a radius from arbitrary points (i.e., in a grid) or from data points, and 3) using spatial clusters (see Table 4-2). Each type of spatial container can be used for different types of event data sets and different purposes of spatiotemporal event data analysis. These spatial containers can be integrated into the burstiness measure, resulting in local indicators for each type of spatial container presented in the next section.

Table 4-2. Type of spatial containers.

Type	Prototype	
Using spatial partitioning	(a) Regular spatial partitions (e.g., grid/hexagon cells)	(b) Geographic boundaries (e.g., ecoregions, administrative boundaries)
		
Using spatial buffers with a radius from centers	(c) Regularly spaced center points	(d) All event locations as center points
		
(e) Using spatial clusters		

Due to the potential effect of MAUP, decisions on the size and shape of spatial containers should be carefully made (Openshaw, 1984; Briant *et al.*, 2010; Rodrigues & Tenedório, 2016; Nelson & Brewer, 2017). One approach to determining the form and size of spatial containers is to analyze the sensitivity or stationarity of statistics across varying forms of spatial containers. Previously, Briant *et al.* (2010) analyzed the impact of the size and shape of spatial aggregation units on the results of spatial analysis and concluded that the size is more important than the

shape. Rodrigues & Tenedório (2016) conducted a sensitivity analysis of global spatial autocorrelation statistics on two forms of aggregation units: administrative regions and hexagonal grids, and concluded that stronger patterns of spatial autocorrelation appear when spatial aggregation units ensure functional homogeneity (e.g. natural regions, urban area). Nelson & Brewer (2017) examined the non-stationarity in variables across varying scales of administrative units by measuring and visualizing local indicators of spatial association (LISA) statistics and their statistical significance. Such existing analyses on the effect of MAUP can provide some insights into how to select spatial containers; however, as those analyses were conducted on using specific data and particular domains, their conclusions cannot be taken at face value to other data and analyses; this is a point emphasized by Nelson & Brewer (2017) about their findings.

Hence, to decide the type of spatial containers, it is useful to try multiple sizes and forms to calculate statistics and find an option that has the higher goodness-of-fit index values (e.g., r-squared, R^2) (e.g., Chen *et al.*, 2008; Kumari *et al.*, 2017). Simultaneously, it can be helpful if researchers have prior knowledge about characteristics of events and their environments as well as the traits of each type of spatial containers. First, spatial partitioning methods divide a space into multiple disjoint subsets (i.e., contiguous spatial aggregation units); see the first column of Table 4-1. An event belongs to only one spatial aggregation unit. Spatial partitioning methods are most appropriate when there is reason to assume some functional relationship between the nature of events and their spatial aggregation units. For instance, to analyze the temporal burstiness of wildfire occurrences, an *ecoregion*—an area exhibiting relative homogeneity of ecosystems (Loveland & Merchant, 2004)—can be used as a spatial container because the underlying mechanisms of wildfire occurrences may be homogeneous within the ecoregion. For social phenomena affected by state government policies (e.g., the illegal sale of marijuana), states can be spatial containers of events.

Second, for spatial buffers, the locality is defined by spatial proximity (i.e., spatial distance or bandwidth) and a spatial aggregation unit is defined as a spatial buffer from each fixed point (see the second column of Table 4-1). The fixed center point can be the location of each individual event (e.g., each vehicle accident) or an arbitrary or systematic location (e.g., grid). Then, the temporal burstiness is measured for events within a spatial buffer. With the spatial buffering method, an event can belong to multiple spatial aggregation units (i.e., spatial buffers). This method would be more appropriate if environmental factors that may trigger following events are rather continuously represented over space and the spatial proximity between successive events is critical. Using the location of each event as the center point for spatial buffering can be viable when it is known that spatial distributions of events are inhomogeneous, and events in some regions are extremely scarce (see Table 4-1d).

Third, with a similar motivation to use the location of each event, spatial clusters of events can be used as spatial containers for the temporal burstiness measure when events are spatially clustered, and those spatial clusters are far apart from each other (see Table 4-1e). Spatial clusters are detected in various ways. For most, the dissimilarity of properties and/or location of events is minimized within the same cluster and maximized between clusters (e.g., k-means clustering) or the density of events within a moving window is higher than the designated threshold (e.g., DBSCAN) (Hartigan & Wong, 1979; Ester *et al.*, 1996). Using spatial clusters as spatial containers ensures the homogeneity in locations or attributes of events in a spatial container. When neither administrative units (e.g., census tracts) nor regular grids (e.g., hexagonal grids) guarantee the functional homogeneity, spatial clusters can be an alternative. Earlier, Andrienko *et al.* (2010) suggested using spatial clustering of point-based data to generate space compartments as spatial aggregation units. Their approach is implemented as a clustering-based spatial partitioning, or regionalization, in which one data point belongs to only one compartment. In my approach, one event can belong to multiple spatial clusters according to

clustering methods. In the following section, I propose multiple LITBs integrated with these three types of spatial containers.

Local Indicators of Temporal Burstiness as ESTDA statistics

A Global Indicator of Temporal Burstiness (GITB)

In Chapter 3, the novel burstiness measure was proposed respectively for a periodic boundary condition (PBC) and an open boundary condition (OBC) (see Eq. (24) and Eq. (47) in Chapter 3). I define a global indicator of temporal burstiness (GITB), equivalent to the burstiness measure (i.e., Eq. (24) and Eq. (47)) applied to all events in the entire region without any spatial constraints. As opposed to GITB, the burstiness measure can be applied to events within the local region that is confined by a spatial container, which is introduced in the next section.

Local Indicators of Temporal Burstiness (LITB)

Local indicators of temporal burstiness (LITB) are designed to calculate the temporal burstiness for a geographically local area and, like other spatially local statistical methods, enable spatial variations of temporal burstiness to be explored. The simplest way to define local indicators is counting events that belong to a local area, as opposed to counting events over the entire region (e.g., Yamada & Thill's (2007) local network K-function). The local area that contains a subset of events serves as a spatial container that filters out events occurring outside of the local area.

Let us define S as a set of events that occur in the entire region, and S_i as a set of events that occur in a local area i where $S_i \subsetneq S$. Then, a general LITB is defined as:

$$Local\ A_i(n) \equiv \begin{cases} \dfrac{\sqrt{n+2}r - \sqrt{n}}{\left(\sqrt{n+2}-2\right)r + \sqrt{n}}, & \text{for } 0 \le r \le \sqrt{n} \quad (OBC) \\ \dfrac{\sqrt{n+1}r - \sqrt{n-1}}{\left(\sqrt{n+1}-2\right)r + \sqrt{n-1}}, & \text{for } 0 \ \le \ r \ \le \ \sqrt{n-1} \quad (PBC), \end{cases}$$

where $\boldsymbol{\tau}$ is inter-event times, $e_a, e_b,$ and $\forall\ e_a, e_b \in\ S_i$, and $n(S_i) > 1$. Two options for the boundary condition are available, as indicated above: OBC and PBC. In the following sections, this general definition of LITB is slightly amended in terms of each of three types of spatial container: contiguous spatial aggregation units, spatial buffers, and spatial clusters. These variations on the definition of LITB are necessary because each type has different conditions in restricting input data points.

LITB with Contiguous Spatial Aggregation Units

Among $\boldsymbol{k}$ contiguous spatial aggregation units, LITB for $\boldsymbol{i}$th spatial aggregation unit (SAU) is defined as:

$$Local\ A_i(n)\text{for SAU} \equiv \begin{cases} \dfrac{\sqrt{n+2}r - \sqrt{n}}{\left(\sqrt{n+2}-2\right)r + \sqrt{n}}, & \text{for } 0 \le r \le \sqrt{n} \quad (OBC) \\ \dfrac{\sqrt{n+1}r - \sqrt{n-1}}{\left(\sqrt{n+1}-2\right)r + \sqrt{n-1}}, & \text{for } 0 \ \le \ r \ \le \ \sqrt{n-1} \quad (PBC), \end{cases}$$

where $\boldsymbol{\tau}$ is inter-event times, $e_a, e_b,$ and $\forall\ e_a, e_b \in\ S_i$, S_i is a set of events in $\boldsymbol{i}$th spatial aggregation unit, $\sum_{i=1}^{n} n(S_i) = n(S)$, and $n(S_i) > 1$. According to this definition, the number of events in the entire region is the sum of the number of events in each spatial aggregation unit.

LITB with Spatial Buffers

LITB for a location $\boldsymbol{i}$, with a spatial buffer (SB) with a radius, $\boldsymbol{d}$ is defined as:

$$Local\ A_i(n)\text{for SB} \equiv \begin{cases} \dfrac{\sqrt{n+2}r-\sqrt{n}}{\left(\sqrt{n+2}-2\right)r+\sqrt{n}}, & \text{for } 0 \le r \le \sqrt{n} \quad (OBC) \\ \dfrac{\sqrt{n+1}r-\sqrt{n-1}}{\left(\sqrt{n+1}-2\right)r+\sqrt{n-1}}, & \text{for } 0 \le r \le \sqrt{n-1} \quad (PBC), \end{cases}$$

where $\boldsymbol{\tau}$ is inter-event times, $e_a, e_b,$ and $\forall\ e_a, e_b \in S_i$, S_i is a set of events within a spatial buffer zone generated with a radius $\boldsymbol{d}$ from a reference location $\boldsymbol{i}$, and $S_i \subseteq S$, and $n(S_i) > 1$.

LITB with Spatial Clusters

LITB for a spatial cluster (SC) $\boldsymbol{i}$ is defined as:

$$Local\ A_i(n)\text{for SC} \equiv \begin{cases} \dfrac{\sqrt{n+2}r-\sqrt{n}}{\left(\sqrt{n+2}-2\right)r+\sqrt{n}}, & \text{for } 0 \le r \le \sqrt{n} \quad (OBC) \\ \dfrac{\sqrt{n+1}r-\sqrt{n-1}}{\left(\sqrt{n+1}-2\right)r+\sqrt{n-1}}, & \text{for } 0 \le r \le \sqrt{n-1} \quad (PBC), \end{cases}$$

where $\boldsymbol{\tau}$ is inter-event times, $e_a, e_b, and\ \forall\ e_a, e_b \in S_i$, S_i is a set of events within a SC $\boldsymbol{i}$, and $S_i \subseteq S$, and $n(S_i) > 1$.

A Method for Statistical Significance Test for GITB and LITB

One important component of ESTDA statistics is a statistical significance test for the statistics. While a theoretical distribution of statistics based on a random assumption that each event happens independently is known or relatively easy to deduce, that of statistics for interdependent events (e.g., clustered and bursty patterns) is often unknown or complicated (Adèr *et al*., 2008; Kim & Jo, 2016).

For this reason, many existing ESTDA statistics adopt a bootstrapping method or Monte Carlo simulation approaches. As examples, the Knox index and Mantel index are tested through a

bootstrapping method because those indexes assume that observations are interdependent in space and time (Knox & Bartlett, 1964; Mantel, 1967; Levine, 2004). In contrast, Yamada and Thill (2007) employed a Monte Carlo simulation approach for their local network k-function to test if an observed pattern rejects a random hypothesis.

For the burstiness measure, Monte Carlo simulation approaches do not work. Randomly sampled points in the given period can consist of a new set of hypothetical events and inter-event times can be derived from those new events; in this situation, the Monte Carlo method tests if temporal patterns of events are a result of a completely random process (CRP), but rejecting the CRP does not mean that the temporal pattern is bursty.

In contrast, the bootstrapping method is viable because it does not sample hypothetical random points, but resamples inter-event times derived from original event points, and then calculates the burstiness measure for resampled inter-event times. Thus, it is possible to avoid a problem of rejecting the CRP in spite of a non-bursty pattern, something that can happen with the Monte Carlo approach.

In this paper, I will adopt a bootstrapping method for three reasons. First, a theoretical distribution of a burstiness measure is unknown and difficult to deduce as observations are not assumed to be independent. Second, a bootstrapping method resamples observations to generate an empirical distribution that is used to test if a statistic is a result of a sampling error (Adèr *et al*., 2008). Third, the bootstrapping method is a resampling method with replacement (Efron, 1992); although only a few time intervals are obtained from a local area, the resampling method enables repeated sampling of time intervals from the local area.

Implementation of Bootstrapping for GITB and LITB

Procedures of the bootstrapping method are as follows.

(1) Randomly sample *k* time intervals among *k* observed time intervals with replacement; this means that each time interval can be sampled multiple times up to *k* times. For example, if there are five time intervals, *A, B, C, D, E*, it is possible to sample a time interval, *A*, for five times.

(2) Calculate the burstiness measure, A_i with *k* sampled time intervals.

(3) Repeat (1) and (2) *n* times (e.g. 10,000 times).

(4) Order the simulated burstiness measure values $(A_1, A_2, \dots, A_{10000})$.

(5) At the 95% confidence level, the bootstrap confidence interval is between the 250th smallest simulated burstiness measure value and the 250th largest simulated burstiness measure value. This means that if the burstiness measure value from *k* observed time intervals falls into the bootstrap confidence interval, it can be assumed to be statistically significant.

Implementation of the Proposed Methodology

This section implements the proposed methodology. Algorithms were devised to compute 1) local indicators of temporal burstiness and conduct 2) a statistical significance test on those indicators.

Algorithms

General functions: burstiness measure and bootstrapping

Three basic functions for implementing a set of LITBs are 1) obtaining inter-event times, 2) calculating the burstiness measure, and 3) bootstrapping for a statistical significance test. The three functions are defined in Table 4-3.

Table 4-3. Functions of obtaining inter-event times, the burstiness measure, and bootstrapping.

DEFINE $\boldsymbol{\tau}$ AS an inter-event time between two subsequent events;
DEFINE $\boldsymbol{n}$ AS the number of events;
DEFINE array **arr_$\boldsymbol{\tau}$** AS a set of inter-event times;
DEFINE $\boldsymbol{A}$ AS a burstiness measure value for events;

FUNCTION **get inter-event times** (events):

    DEFINE time1 AS occurrence time for the current event;

    DEFINE time2 AS occurrence time for the next event;

    FOR all events:
        $\tau \leftarrow$ time2 – time1 // inter-event times
        ADD τ TO arr_τ
    RETURN arr_τ;

FUNCTION **burstiness measure** (n, arr_τ, type):
    DEFINE m_τ AS the mean of elements of arr_τ;
    DEFINE σ_τ AS the variance of elements of arr_τ;
    IF type is OBC:
        IF $\left(\left(\sqrt{n+2}-2\right)r+\sqrt{n}\right)$ is not zero AND $0 \le r \le \sqrt{n}$
        THEN $A \leftarrow \left(\sqrt{n+2}r-\sqrt{n}\right)/\left\{\left(\sqrt{n+2}-2\right)r+\sqrt{n}\right\}$
        RETURN A;

    ELSE IF type is PBC:
        IF $\left(\sqrt{\text{n}+1}\text{-}2\right)\text{r}+\sqrt{\text{n-}1}$ is not zero AND $0 \le r \le \sqrt{\text{n-}1}$
        THEN $A \leftarrow (\sqrt{\text{n}+1}\text{r-}\sqrt{\text{n-}1})/\{\left(\sqrt{\text{n}+1}\text{-}2\right)\text{r}+\sqrt{\text{n-}1}\}$
        RETURN A;

FUNCTION **bootstrapping** (arr_τ, test measure value, iteration number):
    DEFINE array arr_A AS a set of burstiness measure values computed with resamples;
    FOR iteration number:

        DEFINE arr_τ'' AS random.sample(arr_τ, size(arr_τ)); // resample inter-event times
        ADD **burstiness measure** (arr_τ'', type) TO arr_A; // compute the burstiness measure with resamples

    ORDER arr_A BY values;
    DEFINE upper significance envelope AS arr_A[int(iteration number * 5%)]
    DEFINE lower significance envelope AS arr_A[int(numpy.ceil(iteration number * 95%))]
    IF test measure value is between lower and upper significance envelopes THEN

        PRINT "Test measure value is statistically significant with 90% confidence level."

```
        RETURN array[valid significance level, upper significance envelop, lower
        significance envelop];
```

An algorithm for LITBs

Table 4-4. An algorithm for LITBs.

```
FUNCTION LITB (events, a shapefile of spatial containers, iteration number):
    CASE spatial containers OF
        spatial aggregation unit:
            DEFINE event_i AS events within a spatial aggregation unit i;
        spatial buffer:
            DEFINE event_i AS events within a spatial buffer i;
        spatial cluster:
            DEFINE event_i AS events within a spatial cluster i;

    FOR each unit i:
        DEFINE dictionary dict_output AS a set of outputs including burstiness measure
            values and results of a statistical significance test for each unit i;
        arr_τ ← get inter-event times (event_i);
        A ← burstiness measure (arr_τ, type)
        ADD {i: [B]} TO dict_output;
        ADD bootstrapping (arr_τ, A, iteration number) TO dict_output[i];
    JOIN dict_output INTO a shapefile for spatial containers;
    VISUALIZE dict_output as independent maps;
```

Case Study: Spatial Distributions of Temporal Burstiness of Wildfire Events in California, USA

A case study of wildfire events in California, USA was conducted as a proof-of-concept for the methods proposed above. To verify the utility of LITB proposed in the previous section, this case study examines spatial distributions of temporal burstiness of wildfire occurrences in California and compares the temporal burstiness with wildfire frequency.

Background

Wildfires cause tremendous social and ecological costs. For effective forest fire management and ecosystem conservation, it is important to grasp the characteristics of fire regimes by understanding the interplay between the traits of a wildfire event and its drivers. Among fire return intervals, seasonality, fire severity, and spatial extent, fire return intervals are the most important characteristics of fire regimes because they reflect temporal aspects of interactions between fire and the ecosystem, such as fuel accumulation and seasonality of ignitions (Moritz *et al.*, 2011). Wildfires occur if 1) enough fuel is accumulated, 2) fuel is dry enough to burn in terms of seasonal climate, and 3) lightning or humans ignite fires (Moritz *et al.*, 2011), which influence the temporal patterns of wildfires as well as other environmental factors including topography, vegetation type, and soil type.

It is important to characterize the temporal regularity/irregularity of wildfire events over different geographic regions because it allows analysts to consider the association of long suppressions of wildfires (i.e., bursty patterns) with environmental factors. Also, temporal patterns of landscape disturbances (e.g., wildfires) can be used to predict other ecological variables including carbon stores in ecosystems (Smithwick *et al.*, 2007).

The key advantage of the burstiness measure is that it is capable of characterizing power-law behaviors of wildfires in fire return intervals. Measuring the temporal burstiness precedes modeling underlying mechanisms that generate power-law behaviors of wildfires in fire return intervals to better understand the dynamics of wildfires. However, the temporal regularity patterns of wildfire occurrences have not been explored with the burstiness measure as an exploratory data analysis approach.

In contrast to the lack of exploratory research on fire return intervals, there have been studies of power-law relationships between fire size and the corresponding number of

occurrences from a spatial perspective (e.g., Cui & Perera, 2008; Moritz *et al.*, 2011). One mechanism that describes a fire size distribution is *self-organized criticality (SOC)*, which is a simple model concerning initiation and propagation of forest fires that is relevant to the 'domino effect' (Malamud *et al.*, 1998); in wildfire events, the domino effect is the situation in which a fire in one location spreads to a nearby region, and from there to another region perhaps multiple times (e.g., Keeley *et al.*, 2009). Another mechanism is *highly-optimized tolerance (HOT)*, which emphasizes allocation of resources and distributions of costs. HOT considers not only physical conditions such as energy, matter, and information but also engineering factors for systems (Carlson & Doyle, 1999). Moritz *et al.* (2005) compared SOC and HOT mechanisms with empirical data and presented evidence that HOT is more plausible than SOC.

Those models are useful to understand the power-law behaviors of wildfires, once power-law patterns are detected from data. LITBs enable one to detect a power-law behavior, or highly bursty patterns of events as well as regular patterns. Thus, measuring LITBs can prompt interesting hypotheses on dynamics of wildfires.

In this case study, I measure LITBs of wildfire events in California, USA for different types of spatial containers, and then compare geographical distributions of the temporal burstiness with those of the frequency to verify the utility of LITBs.

Data and Method

The study area is California, where wildfires frequently occur. I employed the Federal wildland fire occurrence data provided by the United States Geological Survey for the period from 1980 to 2013 (The USGS, 2014). Fire occurrence has been recorded as a point data set based on the latitude and longitude coordinates of the fire occurrence reports. Here, three types of spatial containers are applied to measuring LITBs and the frequency of wildfires in California

including: regular spatial aggregation units (rectangular grid cells) (Figure 4-2), irregular (natural) spatial aggregation units (ecoregions[5]) (Figure 4-3), and spatial buffers with regularly spaced centers (Figure 4-4). The burstiness measures and their statistical significance levels were calculated for each of these types of spatial container. The results regarding spatial buffers were interpolated with inverse distance weighting (IDW) (Bartier & Keller, 1996). The results contain many data points, so it is not efficient to use point symbols to visualize the spatial variations of temporal burstiness and frequency. An interpolation method with IDW that estimates a value in a spatial location is useful to present the major outcomes of the spatial analysis while minimizing noise in the display.

Results

The main results comprise LITB statistics and the frequencies of wildfire events for each type of spatial container: rectangular grid cells, ecoregions, and spatial buffers (see Figures 4-2, 4-3, and 4-4). The results were visualized on a map with a percentile classification method. With the percentile classification, the number of events assigned to the same color is similar across all the classes, so this method makes it easy to visualize relatively high or low values of events in the data (Slocum, 1999). The bootstrapping-based significance test, as introduced above, has been applied and the results are depicted as grid meshes in the figures.

In general, spatial distribution patterns of the temporal burstiness are different from those of the frequency. First, in the southern California coast area, the frequency is relatively the highest, but the temporal burstiness is close to random (see Figures 4-2c and 4-2d). The contrast between the burstiness and frequency appears stronger in Figure 4-4c and 4-4d. Second, from

[5] http://worldwildlife.org/biomes

analysis with ecoregions, the Sierra Nevada ecoregion exhibits the highest frequency in comparison to other ecoregions, but its temporal burstiness is not the highest. Third, in the ecoregion of the East Cascades - Modoc Plateau, the frequency is relatively low but the temporal burstiness is the highest in California (see Figure 4-3c and 4-3d). Furthermore, for different types of spatial containers there are clear differences of spatial distribution between the temporal burstiness and the frequency of wildfire events. Acknowledging that the dynamics of wildfires involve many diverse factors and their relationships are complex, spatial patterns of LITBs and the frequencies may help establish potential hypotheses on the underlying processes of wildfires.

Discussion

This study made three innovations: 1) the development of LITBs, 2) the introduction of a significance test for GITB and LITB, and 3) the implementation of the proposed methodology. The case study on wildfires showed how such proposed methods can be applied to exploring temporal regularity/irregularity patterns of geographic events. Simultaneously, some important issues were discussed including how to apply this proposed method to analyzing geographic events: how to choose a spatial container.

Notwithstanding, there are topics to be investigated and they are left for future work: how to define an event and how to quantify the effect of the size and shape of spatial containers on temporal burstiness. First, while this research defines an event as a point over space and time, real-world events occur with time duration, spatial extent (e.g., area burned by fire), and different levels of strength (e.g., fire severity). Moreover, it will be beneficial to conduct a sensitivity analysis to examine the effect of MAUP on LITBs. Overall, this research and further work outlined above will contribute to ESTDA research on spatiotemporal events from a geographic

perspective as well as burst research in computational social science and complexity science, bridging the gaps between various fields of study.

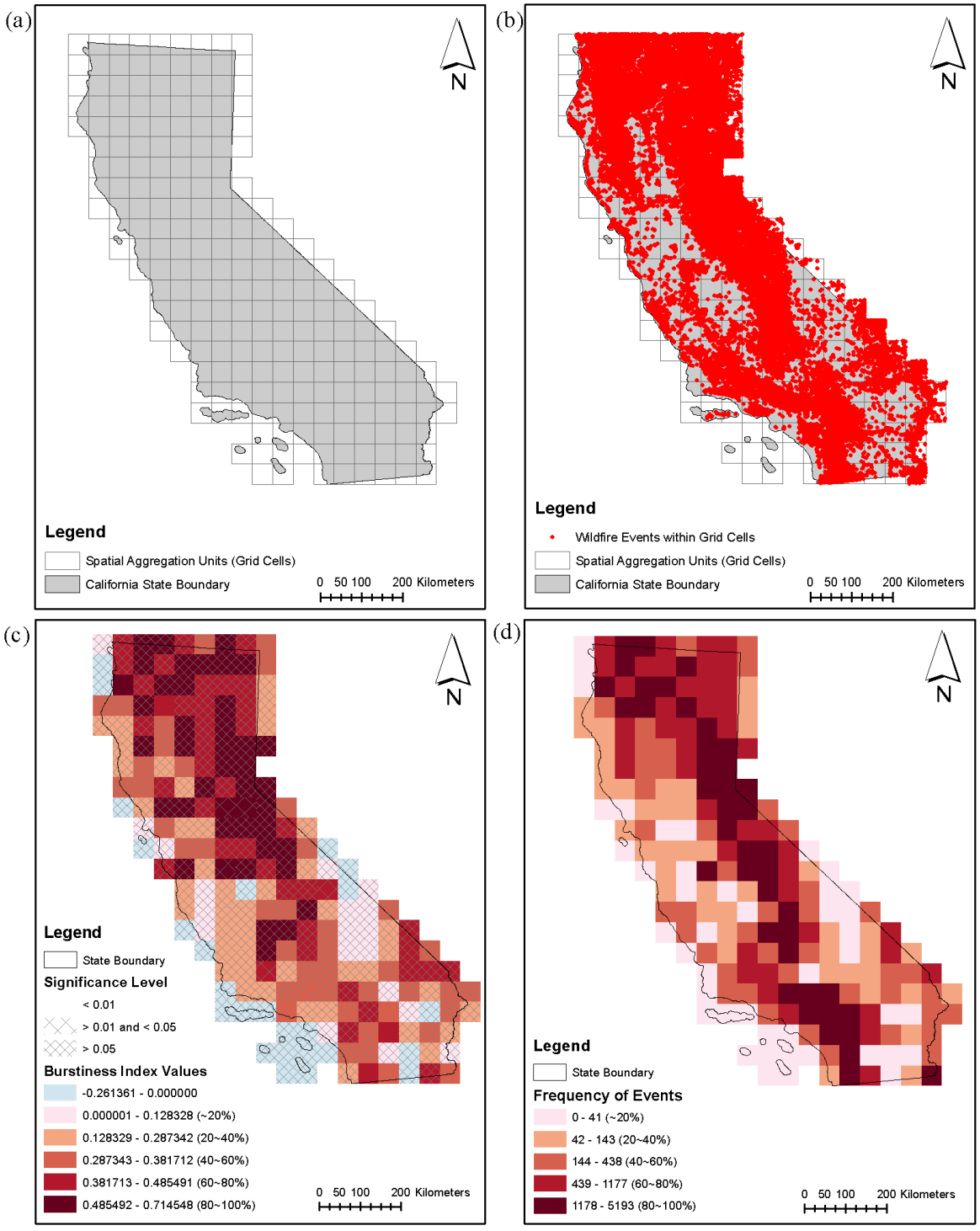


Figure 4-2. LITB of wildfires in California with spatial aggregation units (rectangular grid cells). (a) Grid cells with the length of side of 50km, (b) targeted wildfire events, (c) burstiness measure values with significance levels, (d) frequencies of wildfires.

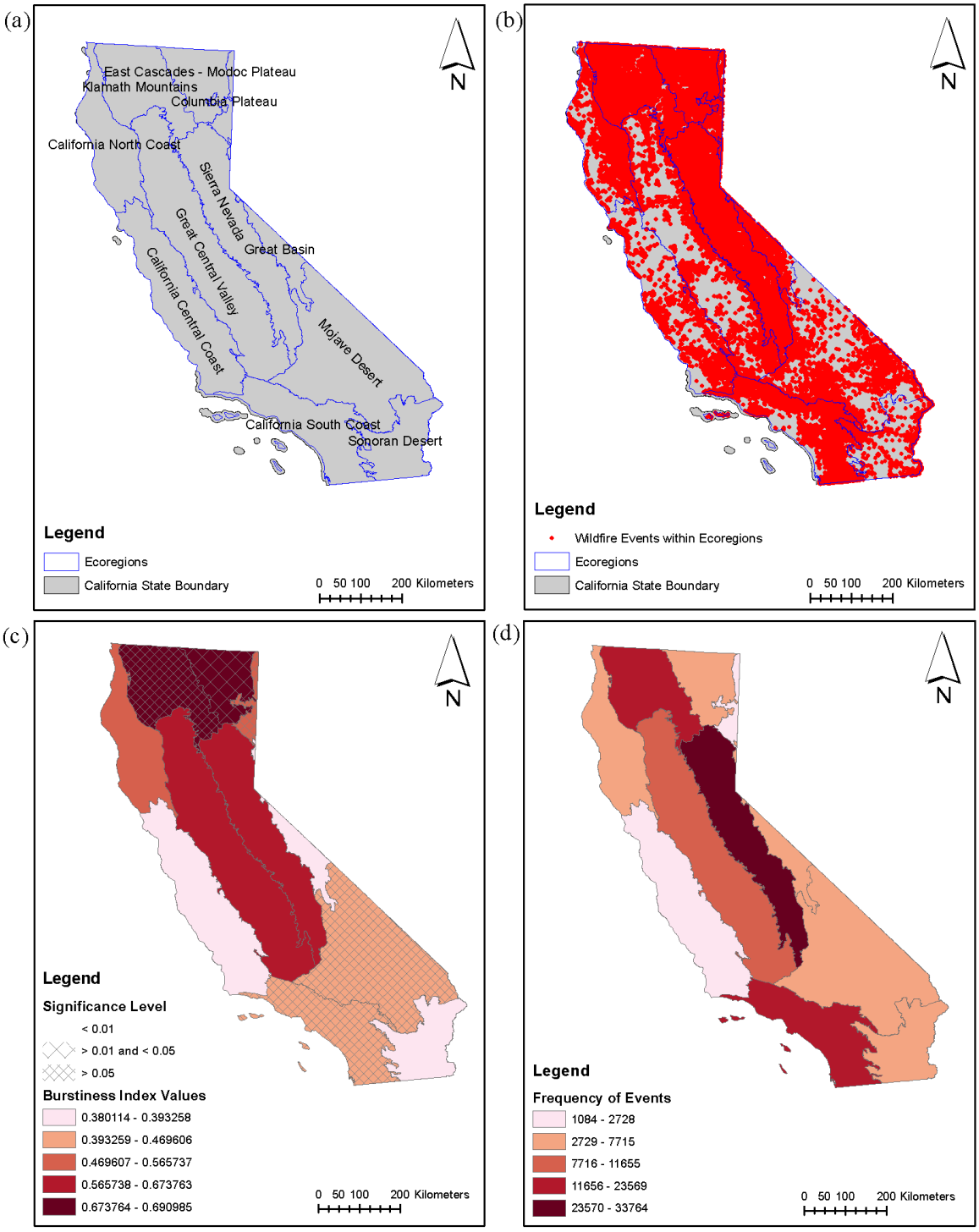


Figure 4-3. LITB of wildfires in California with spatial aggregation units (ecoregions). (a) Ecoregions in California, (b) targeted wildfire events, (c) burstiness measure values with significance levels, (d) frequencies of wildfires.

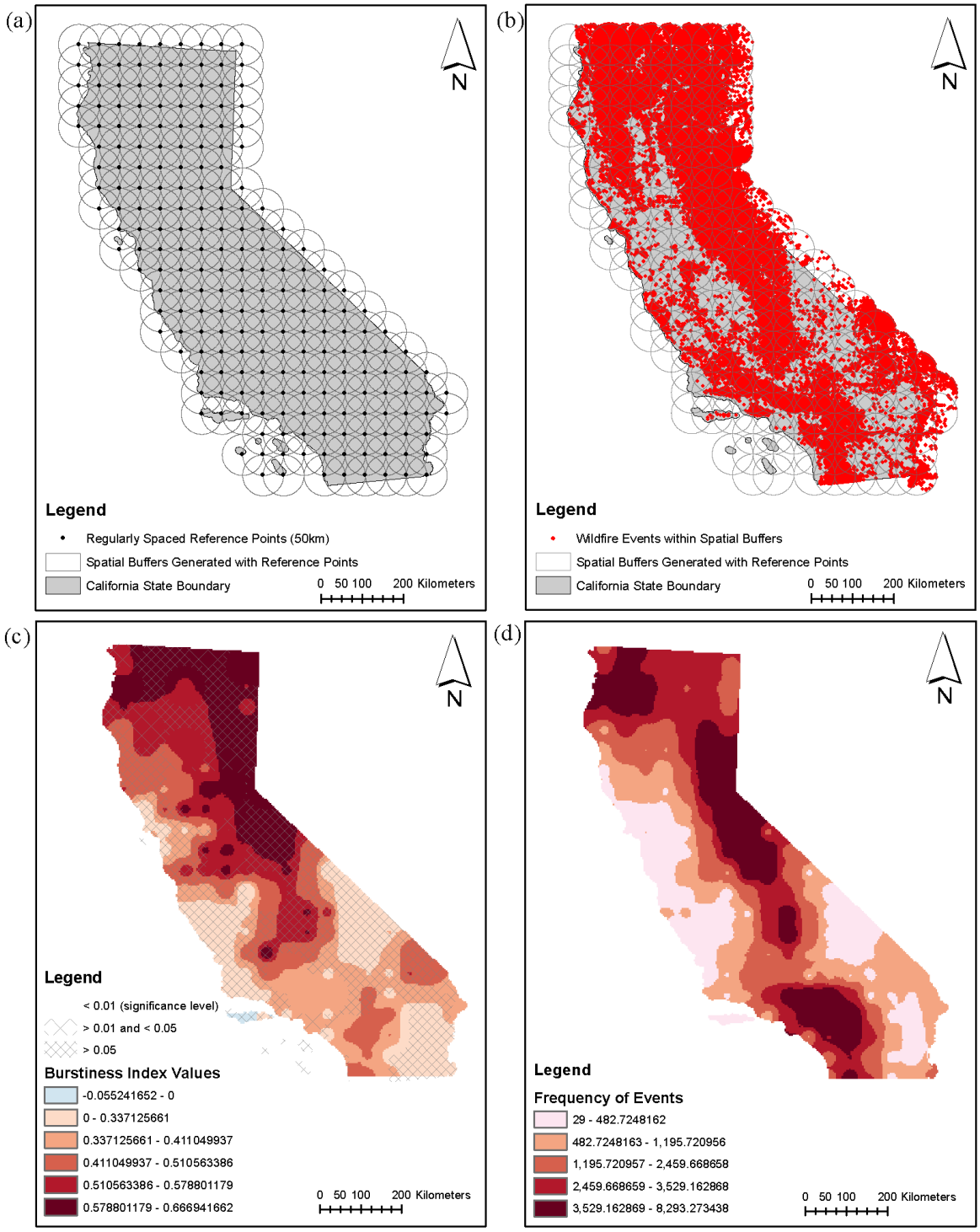


Figure 4-4. LITB of wildfires in California with spatial buffers. (a) Spatial buffers with a radius of 50km regarding regularly spaced reference points (space: 50km), (b) targeted wildfire events, (c) interpolated burstiness measure values with significance levels, (d) interpolated frequencies of wildfires.

Chapter 5

Spatio-Temporal Regularity Patterns of Continental U.S. Contrail Outbreaks[6]

Introduction

The continuous increase in commercial jet traffic has prompted speculation on its impact on regional-scale climate changes via increased cirrus-level *condensation trails (contrails)* (Changnon, 1981; Sassen, 1997; Minnis *et al.*, 1999; Travis *et al.*, 2007). Such contrails potentially influence the Earth's energy balance by altering the *radiative forcing (RF),* contributing to climate changes particularly in the mid-latitudes (Forster *et al.*, 1999, their fig. 6-9; Stuber *et al.*, 2006; Carleton *et al.*, 2013). Positive RF appears when contrails absorb, transmit, and reradiate outgoing longwave radiation from the Earth (positive = warming); negative RF occurs when contrails reflect incoming solar radiation (negative = cooling) (Stuber *et al.*, 2006). Multiple contrails persisting over a long period of time (~1 to 6h) in a region are often developed into *contrail outbreaks* as shown in Figure 5-1, potentially enhancing fluctuations in weather and regional climate (Burkhardt & Kärcher, 2011; Carleton *et al.*, 2013). As an example of the mechanism by which this happens, occurrences of contrail outbreaks alter the surface diurnal temperature range in the conterminous United States (CONUS) (Travis *et al.*, 2002, 2004; Bernhardt & Carleton, 2015).

Occurrence patterns of contrail outbreaks vary regionally and temporally, due to the spatial distribution of jet aircraft flight activity and spatiotemporal dynamics of contributing

[6] The research reported in this chapter is a case study proof of concept that the methods proposed in previous chapters are generally applicable. This chapter is presented as a journal submission written for the climate science community.

climatic factors, such as atmospheric moisture, temperature, and winds (Carleton *et al.*, 2013). Jet aircraft traffic differs over different regions, as cities and their population size are spatially heterogeneous (Palikonda *et al*., 2005; Moninger *et al.*, 2010). It has been confirmed that the western CONUS has lower flight activity than the eastern CONUS including regions of the Midwest, North-east, Mid-Atlantic, South, and Southeast (Palikonda *et al.*, 2005; Moninger *et al.*, 2010; Carleton *et al.*, 2013). In addition, climate conditions in the upper-troposphere (UT) leading to forming and persisting contrails over regions vary according to scales of hour, day, month, season, and year. Previous studies have consistently affirmed that the regional and seasonal variations of the frequency of contrail outbreaks are associated with atmospheric conditions including air temperature (T), relative humidity (RH), vertical motion of air (i.e., ascending versus descending air), horizontal motion of wind, and geopotential height (Travis *et al.*, 2007; Carleton *et al.*, 2008, 2013, 2015). Understanding such relationships between spatiotemporal patterns of contrail outbreaks and contributing UT meteorological factors is the first step to assessing anthropogenic influences of jet aircraft flight activity on climate changes at regional scales and developing appropriate mitigation options for them (Carleton *et al.*, 2013; Carleton & Travis, 2013). Based on such knowledge, individual flight paths and airplane's altitude could be altered in contrail-prone regions and time periods to alleviate formation of contrail outbreaks (Lee *et al*., 2009; Carleton *et al.*, 2013).

The occurrence frequency and sky coverage of contrails and contrail outbreaks are often estimated through satellite-based observations (e.g., DeGrand *et al.*, 2000; Travis *et al.*, 2007). Observations using satellite images allow sensing of atmospheric status for almost the entire region of the CONUS across a wide range of time periods including nighttime (Travis *et al.*, 2007). Several spatial inventories of contrails or contrail outbreaks have been manually built from satellite-based observations for the northern mid-latitude regions including the CONUS and Europe where jet flight paths are densest (see Table 5-1) (Carleton *et al.*, 2015).

Travis *et al.* (2007) and Carleton *et al.* (2013) established spatial inventories of contrail outbreaks in the CONUS for two multiyear periods of 2000-2002 and 2008-2009 for midseason months of January, April, July, and October. Those inventories have been used to analyze: (1) regional and/or seasonal variations in contrail frequencies or coverage (e.g., Travis *et al.*, 2007; Carleton *et al.*, 2008, 2013, 2015), (2) their associations with meteorological factors that form contrail outbreaks (e.g., Carleton *et al.*, 2013, 2015), (3) their associations with jet airplane flight activity (e.g., Travis *et al.*, 2007), and (4) their potential influences on (sub-)regional climates (e.g., Bernhardt & Carleton, 2015). For those analyses, various descriptive and/or confirmatory approaches were adopted; more concretely, statistical composite variables were created for the outbreak frequency from multiyear contrail inventories and visualized on a map (e.g., Travis *et al.*, 2007; Carleton *et al.*, 2013); inferential statistical methods were applied including analysis of variance (ANOVA) (e.g., Travis *et al.*, 2004), a correlation analysis (e.g., Travis *et al.*, 2007), and binary logistic regression to determine which climate variables are significant predictors for contrail outbreak incidences (e.g., Travis *et al*., 1997; Carleton *et al*., 2015). These studies improved knowledge on the regional climatology of contrail outbreaks.

Notwithstanding, some aspects of contrail outbreaks have not been explored, so this study addresses the following three issues. First, in existing research, the local occurrence frequency of contrail outbreaks provided essential information on the likelihood of contrail outbreak incidence in the given time and space. However, the frequency is unable to capture another important aspect of temporal dynamics of phenomena: the *local temporal burstiness* – how regular or bursty the events are over time in a local area. Here, temporal burstiness is defined as the extent to which time intervals of events vary, no matter what the temporal scale of the phenomenon; the temporal burstiness is high if extremely large time intervals appear together with many small time intervals in an event sequence (Kim & Jo, 2016). Although the frequencies of contrail outbreaks might be equally high in two cities (e.g., Detroit, MI and Atlanta, GA), the temporal regularity patterns of

their contrail outbreaks could be totally different. It could be temporally highly regular (e.g., an outbreak occurs on every other day) or temporally very bursty (e.g., all the outbreaks occur on one day for a month). This temporal regularity aspect can be characterized by a *local indicator of temporal burstiness* (LITB) proposed in Chapter 4. Second, the temporal burstiness of outbreaks may vary throughout different regions, potentially depending on temporal dynamics of UT synoptic climates and other atmospheric conditions, which has not been investigated. Third, previously applied inferential statistical models (e.g., correlation analysis, multiple logistic regression analysis) assume the spatial homogeneity of relationships between variables, but actually relationships can change over different regions. This means that, although one explanatory variable (e.g., relative humidity) is the best predictor of contrail occurrences in one region, another variable (e.g., temperature) could be the best predictor in other regions. Geographically weighted regression (GWR) models are known to reconcile such spatially heterogeneous statistical relationships with a general regression model (Brusdon *et al.*, 1996).

Hence, the present study adopted statistical methods of LITB and GWR and spatial inventories of contrail outbreaks used in Travis *et al.* (2007) and Carleton *et al.* (2013) to tackle the following research questions for each analysis:

- Local temporal burstiness of contrail outbreaks:
    - whether contrail outbreaks occur regularly or are bursty in time in each region in the CONUS;
    - whether the temporal regularity patterns of contrail outbreak occurrences are spatially heterogeneous across the CONUS;
- GWR models for inferring relationships among local UT atmospheric conditions, the local temporal burstiness, and the local frequency:

- whether UT meteorological factors including T, RH, vertical motion of air, horizontal motion of wind, and geopotential height have a significant explanatory power for the temporal burstiness or the frequency of contrail outbreaks;
- whether the local temporal burstiness is a significant predictor for the local frequency of contrail outbreaks;
- whether those relationships are spatially heterogeneous.

The rest of the paper proceeds as follows. The 'background' section introduces closely relevant concepts, methods, and preceding studies. The 'data and method' section describes the data source, data processing procedures, and analytical methods used in this study. The 'results' section reports the outcomes from analyses and interpretations of results. The last section summarizes the research and discusses the implications of this study.

Background

Contrails, contrail outbreaks, and contrail cirrus

Contrails (condensation trails) are visible linear ice-crystal clouds, the condensate of aircraft-emitted water vapor and sublimated ambient moisture, formed around the upper troposphere (Penner *et al.*, 1999; Schumann, 2000). Most visible contrails tend to dissipate after only a few minutes, but some contrails last longer even for many hours and some evolve into *contrail cirrus*, irregularly shaped cirrus cloud cover spread from single or multiple contrails (Jensen *et al.*, 1999; Burkhardt & Kärcher, 2011; Minnis *et al.*, 2013). In northern middle latitudes where commercial flight traffic is heavy, and in contrail-prone atmospheric environments, multiple contrails occur as clusters, called *contrail outbreaks* (see Figure 5-1) (Carleton & Lamb, 1986; Carleton *et al.*, 2008, 2015). The spatial scale of contrail outbreaks is at

least 1×10^3 km^2, extending over approximately 10^4~10^5 km^2, and they endure over ~1 to 6 h (Minnis *et al.*, 1998; Carleton *et al.*, 2013, 2015).

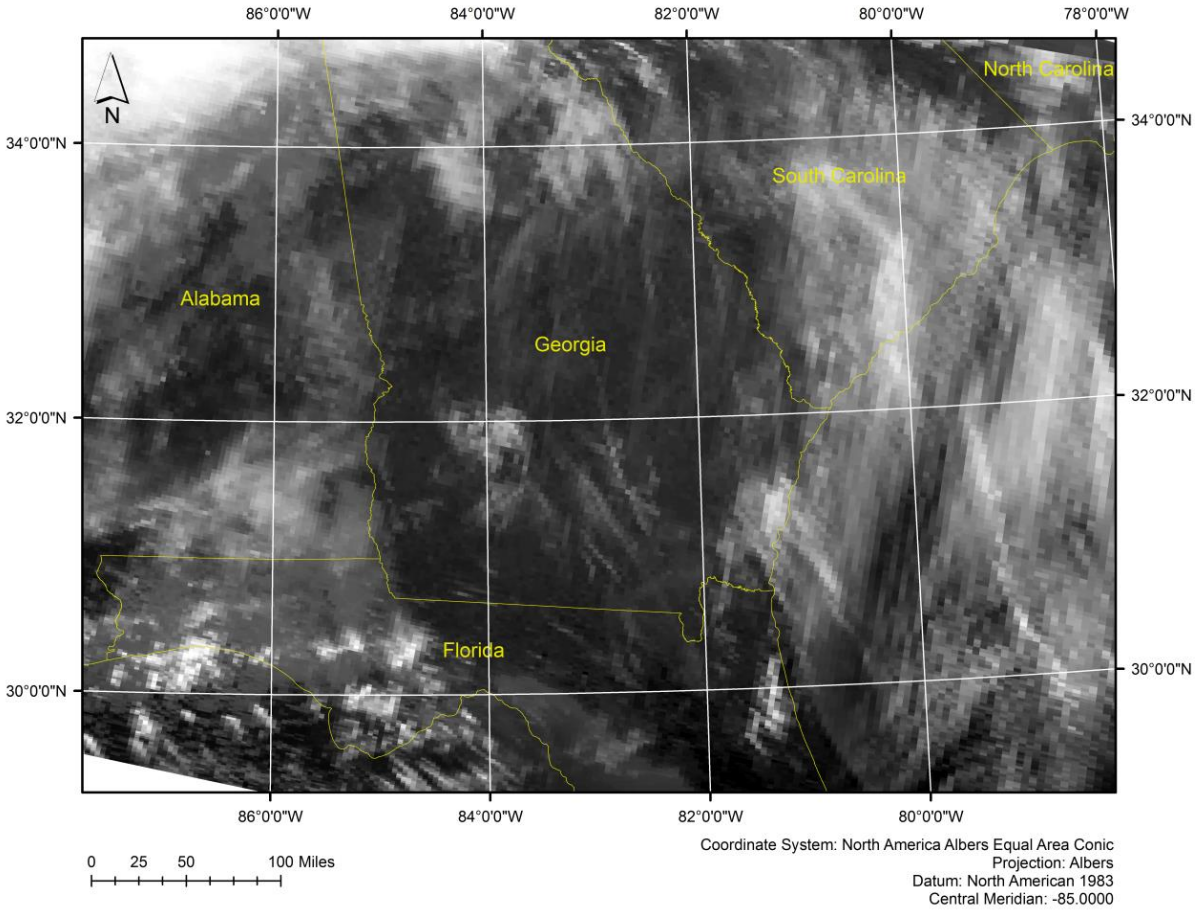


Figure 5-1. AVHRR thermal IR image (channel 4: 10.3 µm - 11.3 µm) of contrail outbreaks occurred in Georgia on January 29th, 2008 between 3:44 pm ~ 5:22 pm.

Formation and persistence of contrails and contrail outbreaks

Due to their potential climate change impacts, the formation conditions of long-lasting contrails and contrail outbreaks have been studied by many researchers. The formation of ice crystals in the atmosphere, a main composition of contrails, requires significant supersaturation with respect to ice (i.e., $RH_{ice} > 100\%$), partly because only about one of a million particles serves as an ice nucleus and, for some strongly acidic particles (e.g., H_2SO_4), the solution needs to be

highly diluted for freezing (Jensen *et al.*, 1998; Gierens *et al.*, 2012). Aircraft-emitted water vapor and soot can contribute to the formation and persistence of contrails in the very cold upper troposphere, when the ambient temperature is low and the ambient relative humidity is substantially high, leading to *ice supersaturations* (Burkhardt & Kärcher, 2011; Gierens *et al*., 2012). The existence of an *ice-supersaturated region* (ISSR) in the upper troposphere where contrails last long and cirrus clouds can form has been generally accepted and integrated into the weather forecast models (e.g., ECMWF) (Detwiler & Pratt, 1984; Gierens *et al.*, 2012). The geographical distributions of ice supersaturation occurrence frequencies vary with the pressure layer in the atmosphere and with season (Lamquin *et al.*, 2012, their fig. 9 and 10), implying that spatial distributions of contrail outbreak frequencies may change over different pressure layers and seasons. While ISSR partially explains the spatiotemporal variability of contrail outbreaks, several studies have focused on how UT meteorological factors including T, RH, vertical motion of air, horizontal motion of wind, and geopotential height are associated with the occurrence of contrail outbreaks at a regional scale (Bakan *et al.*, 1994; Jensen *et al.*, 1998; DeGrand *et al.*, 2000; Travis *et al.*, 2007; Carleton *et al.*, 2008, 2013, 2015). Needless to say, jet aircraft activity is heterogeneous across regions in the CONUS, resulting in spatiotemporal variations of contrail outbreaks (e.g., Machta & Carpenter, 1971; Liou *et al.*, 1990; Changnon, 1981; Travis *et al.*, 2007).

Observations and inventories of contrails and contrail outbreaks

The occurrence and the spatial extent of contrails or contrail outbreaks have been detected via satellite-based observations (e.g., Minnis *et al.*, 1998; DeGrand *et al.*, 2000; Travis *et al.*, 2007), ground-based observations (i.e., surface observations) (e.g., Changnon, 1981; Liou *et al.*,1990; Minnis *et al.*, 2003), or citizen self-reports (e.g., Duda & Minnis, 2009). Surface

observations or self-reporting based observations can be interfered by intervening clouds and changing sky-view perspective and they are difficult to make at night (Travis *et al.*, 2007). Thus, thermal infrared (IR) satellite-based observations have become commonly used in contrail studies, due to the ability of this technology to discern almost the entire region of the CONUS over a long time period, at all hours of the day or night (Travis *et al.*, 2007; Bernhardt & Carleton, 2015). Imagery datasets obtained from various Earth observation satellites have been adopted to detect contrails and their outbreaks; at least three satellite imagery datasets have been used: (a) the Geostationary Operational Environmental Satellite (GOES) data (1 km pixel resolution) (e.g., Minnis *et al.*, 1998, 2013), (b) the IR channel hard-copy Defense Meteorological Satellite Program (DMSP) imagery (0.6 km pixel resolution) (e.g., DeGrand *et al.*, 2000), and (c) Advanced Very High Resolution Radiometer (AVHRR) satellite thermal IR images (1.1 km pixel resolution) (e.g., Travis *et al.*, 1997, 2007) (see Table 5-1).

Contrails can be detected from image data either automatically (e.g., Zhang *et al.*, 2012) or manually (e.g., Carleton *et al.*, 2015). Although it is labor-intensive, the manual detection can yield more accurate results than currently available automated detection techniques that sometimes fail at distinguishing contrails from geomorphological features (e.g., rivers, mountain ranges) (Travis *et al.*, 2007). Following DeGrand *et al.*'s (2000) contrail inventories of the CONUS for years of 1977-79, Travis *et al.* (2007) and Carleton *et al.* (2013) manually constructed spatial inventories of contrails in the CONUS for years of 2000-2002 and 2008-2009 respectively, based on a large number of AVHRR satellite IR images.

Table 5-1. Previous studies on analysis of contrail outbreaks at regional scales using thermal infrared satellite imagery data.

Study	Imagery data	Spatial extent	Temporal extent	Variables	Analytical method
Bakan *et al.* (1994)	AVHRR[7]	western Europe; eastern North Atlantic	September 1979 - December 1981; September 1989 - August 1992	•the averaged contrail cover; •contrail frequencies; •atmospheric conditions	•comparison of two different periods •comparison between the averaged contrail cover and atmospheric conditions
Travis *et al.* (1997)	AVHRR[7]; DMSP[8]	CONUS	January and April, 1987	•contrail coverage of 33 contrail outbreaks; •atmospheric conditions: water vapor[9], temperature[10]	•**multivariate binary logistic regression** for predicting widespread contrail occurrences
DeGrand *et al.* (2000)	DMSP[8]	CONUS	January, April, July, October for 1977-1979	•contrail frequencies on 1°×1° grid cells; •density of flight routes; •synoptic-scale atmospheric circulation: geopotential height, westward wind, air temperature	•visual UT-map technique for composite normalized frequencies of contrails; •outbreak climate diagnostics with synoptic-scale atmospheric circulation factors; •categorization of circulation type and cloud type; •contrail case event studies
Travis *et al.* (2004)	AVHRR[7]	CONUS	January, April, July, October for 2000-2002	•contrail frequencies on 1°×1° grid cells; •atmospheric factors[11]: temperature, humidity, vertical wind shear, vertical motion	•**a Pearson correlation coefficient** between contrail frequency and diurnal temperature range (DTR) •**statistical analysis of contrail outbreak retro-prediction** with statistical composites

[7] Digital imagery data of the advanced very high-resolution radiometer (AVHRR) - thermal infrared (IR) channel (1.1 km pixel resolution)

[8] Hard-copy imagery data of defense meteorological satellite program (DMSP) - thermal infrared (IR) channel (0.6 km pixel resolution)

[9] the Geostationary Operational Environment Satellite (GOES) - water vapor absorption band (6.7 μm)

[10] National Weather Service (NWS) cloud cover data

[11] Data source: NCEP-NCAR reanalysis data, http://www.cdc.noaa.ogv/

Travis *et al.* (2007)	AVHRR[7]	CONUS	January, April, July, October for 1977-1979 and 2000-2002	•contrail frequencies on 1°×1° grid cells; •atmospheric factors[11]: air temperature, pressure-altitude; •jet aircraft flight activity[12]: daily total scheduled number of flights, total scheduled flight distance, daily average length of flights	•visual UT-map technique for the mean contrail outbreak frequency and the mean tropopause temperature change; •**ANOVA** for testing inter-seasonal variations in the mean contrail frequency and jet airplane flight activity by year and month; •**correlation test** for 1) contrail frequency change × mean tropopause temperature change (C°) and 2) contrail frequency change × mean tropopause pressure (mb)
Carleton *et al.* (2008)	AVHRR[7]	CONUS	January, April, July, October for 2000-2002	•contrail frequencies on 1°×1° grid cells; •atmospheric factors[11]: air temperature, humidity, vertical motion, east-west wind component, and geopotential height;	•regionalization of contrail outbreaks •annual midseason month totals of contrail outbreaks •analysis of composite (i.e., multicase average) "synoptic climatology"
Duda & Minnis (2009)	AVHRR[13]; GOES-12[14]	CONUS	April 2004 – 27 June 2005 (15 months)	•persistent contrail occurrence •atmospheric factors[15]: temperature, humidity, horizontal wind speed and direction, and vertical velocity	•**binary logistic regression models** based on numerical weather analyses for predicting contrail formation
Carleton *et al.* (2013)	AVHRR[7]	CONUS	January, April, July, October for 2000-2002 and 2008-2009	•contrail frequencies on 1°×1° grid cells; •atmospheric factors[11]: air temperature, humidity, vertical motion, east-west wind component, and geopotential height;	•visual UT-map technique for the mean contrail outbreak frequency (i.e. contrail outbreak overlap normalized frequencies; •outbreak climate diagnostics for sub-regions of higher-frequency outbreak overlaps determined semi-objectively by GIS

[12] Data source: BACK Aviation Solutions (2015); US domestic flights

[13] Digital imagery data of AVHRR - infrared (10.8 μm) minus split window (12.0 μm) brightness temperature difference (BTD) data

[14] Digital imagery data of the Geostationary Operational Environmental Satellite-12 (GOES-12) - infrared (10.8 μm) and water vapor (6.5 μm) channel

[15] Data from hourly meteorological analyses from the Advanced Regional Prediction System (ARPS) (27-km resolution) and the Rapid Update Cycle (RUC) (20-km resolution)

Carleton *et al.* (2015)	AVHRR[7]	CONUS	January, April, July, October for 2000-2002 and 2008-2009	• contrail frequencies on 1°×1° grid cells; • atmospheric factors[11]: air temperature, relative humidity, vertical motion, and east-west wind component	• visual UT-map technique for contrail outbreak frequencies • comparison of daily maps with the composite fields for outbreak days (CON) versus nonoutbreak days (NON); evaluation with **standard skill measures** • **binary logistic regression** with **sensitivity tests** for determining which UT variables are significant predictors, individually and in combination
Bernhardt & Carleton (2015)	AVHRR[7]	Two sub-regions of CONUS	South in January & Midwest in April, 2008-2009	• contrail frequencies on 1°×1° grid cells; • DTR	• **nonparametric Mann–Whitney U-test** to determine statistical significance of the composite differences in DTR between outbreak and non-outbreak stations pairs

Contrail research using spatial inventories of contrails in the CONUS

Spatial inventories built by Travis *et al.* (2007) and Carleton *et al.* (2013) were employed for multiple descriptive and explanatory analyses on the regional climatology of contrail outbreaks in the CONUS (see Table 5-1). Travis *et al.* (2007) created composite frequencies of contrails from multiyear contrail inventories and visualized them on a map of 1°×1° grid cells. Carleton *et al.* (2013) regionalized high-frequency areas of contrail outbreaks in the CONUS in terms of contrail frequencies via a GIS-based regionalization method, and compared composite meteorological variables to the composite contrail frequencies in each region. Travis *et al.* (2007) adopted an analysis of variance (ANOVA) test to verify inter-seasonal variations in the mean contrail frequency and jet airplane flight activity by year and month, and did correlation analysis of changes in the contrail frequency and those in each of the mean tropopause temperature and pressure. Carleton *et al.* (2015) determined significant predictors of climate variables for the incidence of contrail outbreaks by applying binary logistic regression. Bernhardt & Carleton (2015) confirmed, through a nonparametric Mann–Whitney U-test, that clear-sky contrail outbreaks significantly reduce the diurnal temperature range (DTR) at a regional scale. These studies improved knowledge on (1) where and in which season contrail outbreaks in the CONUS occur more frequently, (2) which UT climate factors contribute to forming and maintaining contrail outbreaks in the CONUS, and (3) whether contrail outbreaks make significant impacts on regional climates.

Although there were attempts to identify seasonal variations in the contrail frequency throughout the CONUS (e.g., Travis *et al.*, 2007; Carleton *et al.*, 2013) or the lifetime of contrails (e.g., Minnis *et al.*, 1998), little research has been conducted on the temporal regularity of occurrences of contrail outbreaks that can be measured by the LITB. The temporal regularity information has the potential to both enhance comprehension of dynamic mechanisms of contrail

formation and persistence, and support development of strategies to mitigate contrail outbreaks by regulating jet airplane flight traffic. Therefore, it is worthwhile to explore the spatial distributions of the LITB of contrail outbreaks and their relationships with UT climate variables.

Data and Methodology

This section introduces the data and methodology used in the study, which are visually depicted in Figure 5-3. The discussion below is linked to relevant parts of the figure through call-outs to the lettered sections, A-F.

Satellite-derived spatial inventories of contrail outbreaks

This study adopted two sets of satellite-based daily spatial inventories of contrail outbreaks in the CONUS constructed for the midseason months of January, April, July, and October of 2000-2002 and 2008-2009, respectively, by Travis *et al.* (2007) and Carleton et al. (2013) (see Figure 5-3a). Both inventories were built upon AVHRR thermal IR digital images (channel 4: 10.3 μm - 11.3 μm, 1.1 km^2 nadir pixel resolution). These images are available online at the website[16] of the U.S. National Oceanic and Atmospheric Administration (NOAA) Comprehensive Large Array-data Stewardship System (CLASS). First, Travis *et al.* (2007) selected 2126 images for 2000-2002 in pairs that cover the CONUS for three local time periods per location: nighttime (00-09 UTC), morning (09-18 UTC), and afternoon (18-00 UTC) for both the western and eastern halves of the CONUS by using remote sensing and Geographic Information Systems (GIS) software. The images were selected to have about 4.5 h spacing

[16] https://www.class.ngdc.noaa.gov/saa/products/search?datatype_family=AVHRR

between each image pair to avoid double-counting contrail outbreaks, which resulted in a median of six images per day (Travis *et al.*, 2007). Second, Carleton *et al.* (2013) built the inventory of 2008-2009 in a similar way to that done by Travis *et al.* (2007); observations were possible six to eight times per day (Bernhardt & Carleton, 2015).

Contrail outbreaks in both sets of inventories were manually detected according to all the following criteria pertinent to the definition of contrail outbreaks and their climate impacts: (1) three or more contrails occur in the same region over at least three adjacent 1˚ grid cells in any direction; and (2) the contrail outbreaks occur in an area of 50% or less natural cloud coverage. The occurrence time and spatial extent of each outbreak were recorded in spreadsheets. Each outbreak was identified as a bounding box (BBOX) with geographic coordinates (longitude, latitude) of upper-left and the bottom-right corners (see Figure 5-2); these data were tabulated in a spreadsheet by midseason month.

Adjustment of spatial inventories of contrail outbreaks

Additional adjustment of contrail data was needed for further analysis. Figure 5-3a illustrates the processing. BBOXs of contrail outbreaks from two spatial inventories were rearranged to improve consistency in the size of contrail outbreaks between months, because the initial criteria used to construct raw inventories do not specify the adjacency in space and time between contrail outbreaks. Two or more BBOXs of contrail outbreaks were merged together if they met the following rules: first, (1) BBOXs are smaller than 3 degrees in either width or length; (2) one BBOX contains or overlays other BBOX(s) or BBOXs are adjacent to one another within 1.5 degrees in both latitude and longitude; (3) in either case of (1) and (2), the duration time between BBOXs must be less than 6 hours for purposes of comparisons with the atmospheric data. In addition to the merging, if a BBOX is smaller than a 1˚ × 1˚ grid cell and

does not have adjacent BBOXs in space and time, the BBOX is eliminated. Once a group of BBOXs meet these rules, the BBOXs are replaced by one with the upper-left corner (the minimum longitude, the minimum latitude) and the bottom-right corner (the maximum longitude, the minimum latitude), and the occurrence time of the merged contrail outbreak is determined by averaging occurrence times of the input contrail outbreaks.

As a result, some small BBOXs are eliminated and the variations of the width and length of BBOXs as well as the number of BBOXs between months decreased after the adjustment. As shown in the last row of Table 5-2, the standard deviation (STDEV) for the number of BBOXs in each month decreased from 70.04 to 44.53. STDEVs for the mean length and the mean width of BBOXs of each month also decreased from 0.74 to 0.66 and from 1.60 to 1.48, respectively, signifying an improvement of the consistency of observations by excluding too small or possibly double-counted contrail outbreaks. According to the ‘difference’ column group (right side of Table 5-2), in most of the months when BBOXs that were close in space and time were merged, the mean length and mean width of BBOXs increased after the alteration, particularly for months having a high number of contrail outbreaks (e.g., July and October of 2009). For some cases, the mean length of BBOXs decreased for July of 2008; this was because multiple BBOXs of large size were merged together due to their spatiotemporal proximity. After the rearrangement of BBOXs, fluctuations in BBOX size over seasons and years still exist; revealing what produces such differences is to be a focus of future analysis.

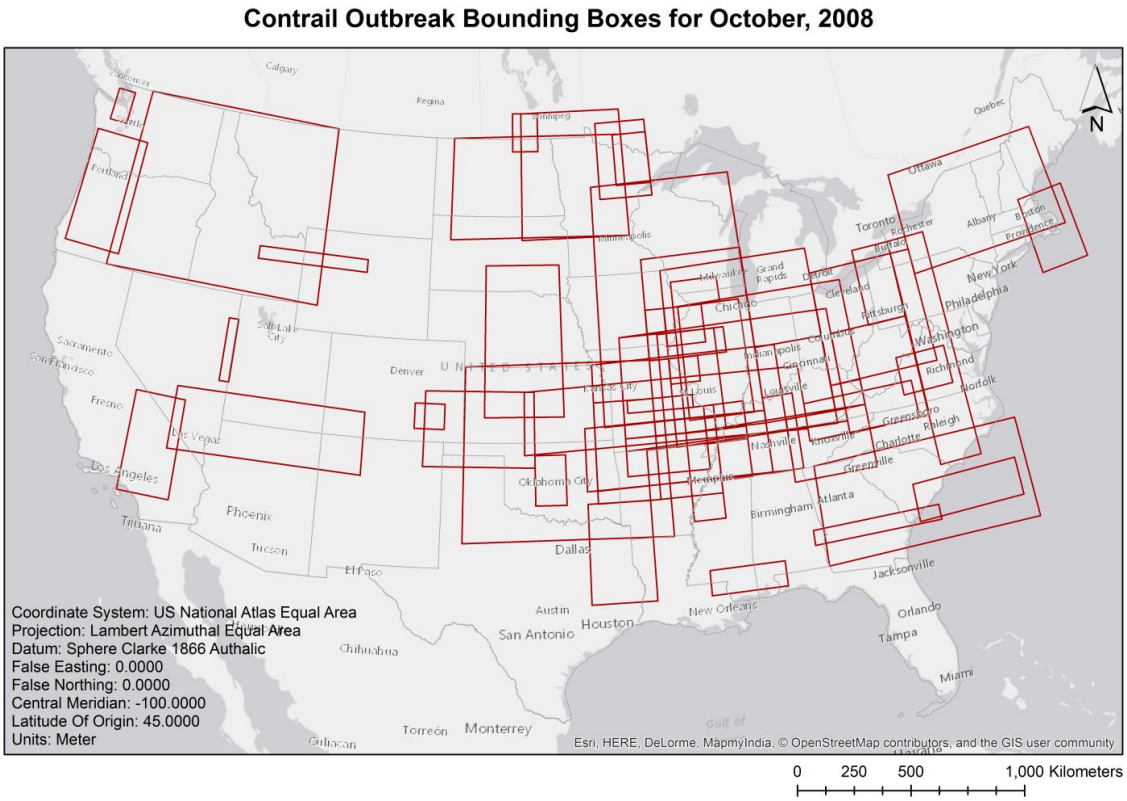


Figure 5-2. Contrail outbreak bounding boxes for October 2008 after adjustment.

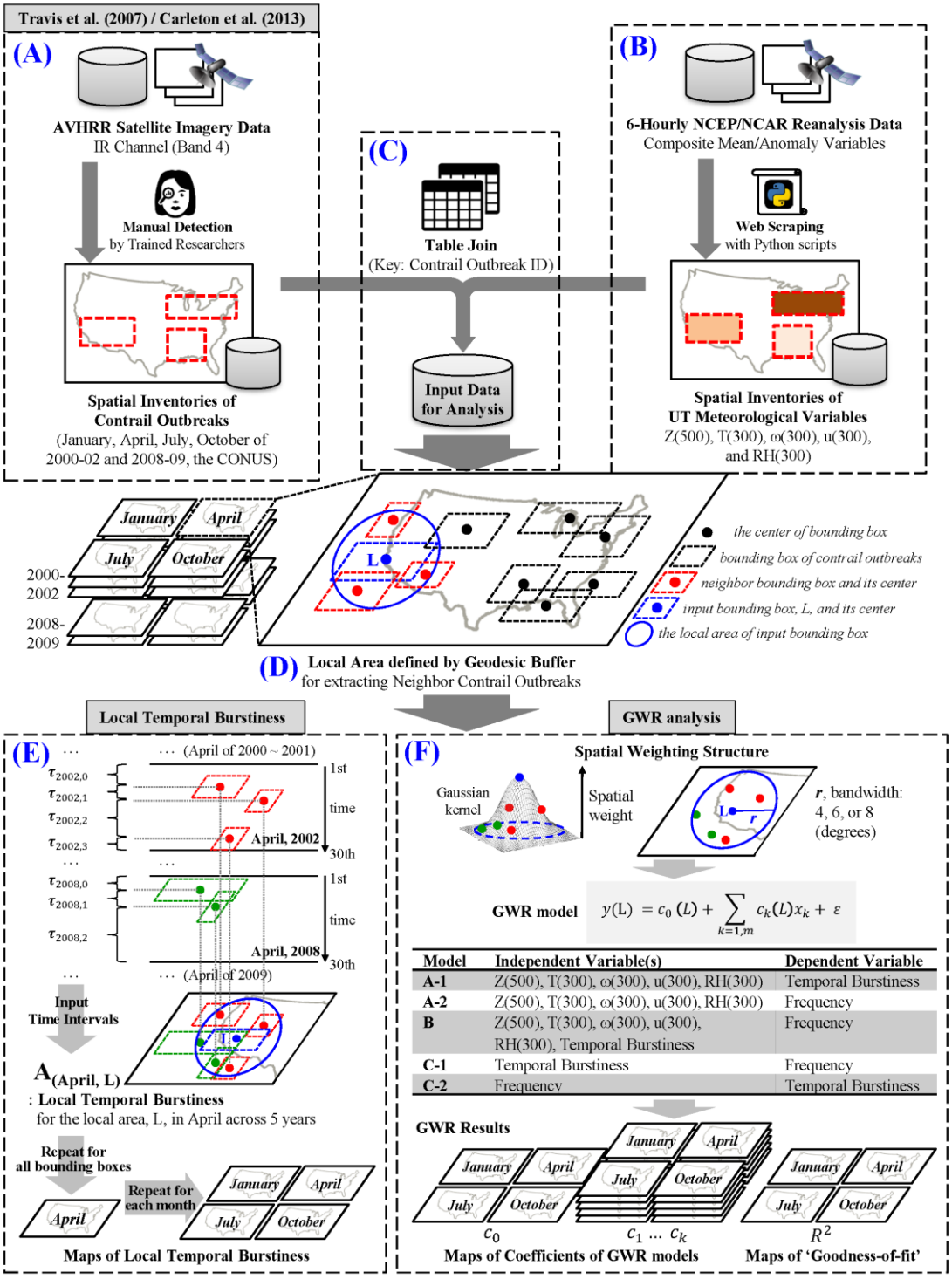


Figure 5-3. Data and methodology of the current study. The process is organized into six components: A - construction of spatial inventories of contrail outbreaks in U.S.; B - construction of UT meteorological variables; C – Integration of BBOXs of contrail outbreaks and UT climate variables; D – the definition of a local area by geodesic buffering; E – burst analysis; F – GWR analysis.

Table 5-2. Changes in the size of bounding boxes of contrail outbreaks before and after adjustments. The color scheme for the size of bounding boxes was applied in pairs of the original and adjusted statistics for each of the mean length, the mean width, the standard deviation of length, and standard deviation of width.

Year	Month	Number of Original Contrail Outbreaks	Number of Adjusted Contrail Outbreaks	Number of Removed Contrail Outbreaks	Number of Merged Groups of Outbreaks	The Size of Bounding Boxes of Original Contrail Outbreaks*				The Size of Bounding Boxes of Adjusted Contrail Outbreaks*				Difference (Adjusted – Original)			
						Mean		STDEV		Mean		STDEV		Mean		STDEV	
						Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
2000	January	11	10	1	0	2.70	7.80	0.78	2.27	2.70	7.80	0.78	2.27	0.00	0.00	0.00	0.00
	April	34	34	0	0	4.91	7.41	2.16	3.27	4.91	7.41	2.16	3.27	0.00	0.00	0.00	0.00
	July	10	10	0	0	2.70	7.80	0.78	2.27	2.70	7.80	0.78	2.27	0.00	0.00	0.00	0.00
	October	22	21	1	0	4.38	8.43	2.01	3.39	4.38	8.43	2.01	3.39	0.00	0.00	0.00	0.00
2001	January	30	27	3	0	2.93	6.70	1.94	3.74	2.93	6.70	1.94	3.74	0.00	0.00	0.00	0.00
	April	22	22	0	0	3.14	4.77	1.69	2.17	3.14	4.77	1.69	2.17	0.00	0.00	0.00	0.00
	July	14	14	0	0	3.57	5.50	2.61	4.14	3.57	5.50	2.61	4.14	0.00	0.00	0.00	0.00
	October	37	32	0	3	2.49	5.00	1.13	2.94	2.78	5.53	0.99	2.87	0.29	0.53	-0.14	-0.07
2002	January	20	19	0	1	3.60	6.30	2.18	4.16	3.58	6.58	2.23	4.08	-0.02	0.28	0.05	-0.08
	April	30	29	0	1	3.47	5.60	2.26	3.66	3.59	5.83	2.24	3.62	0.12	0.23	-0.03	-0.04
	July	28	28	0	0	4.07	6.89	1.60	3.49	4.07	6.89	1.60	3.49	0.00	0.00	0.00	0.00
	October	13	13	0	0	2.77	4.38	0.80	2.10	2.77	4.38	0.80	2.10	0.00	0.00	0.00	0.00
2008	January	42	35	0	7	3.85	7.96	2.71	4.67	3.97	7.97	2.93	4.93	0.11	0.01	0.22	0.25
	April	53	47	2	3	2.61	4.37	1.69	2.69	2.78	4.44	1.72	2.71	0.17	0.07	0.03	0.02
	July	71	58	1	9	2.73	3.76	1.84	2.43	2.71	3.89	1.97	2.65	-0.01	0.12	0.13	0.22
	October	66	56	1	8	3.19	4.51	1.92	3.05	3.25	4.66	2.10	3.25	0.05	0.15	0.18	0.19
2009	January	94	71	0	17	2.32	6.02	1.74	3.45	2.67	6.72	1.90	3.57	0.35	0.70	0.16	0.13
	April	93	75	0	15	2.37	3.63	1.44	2.69	2.70	4.19	1.50	2.84	0.33	0.56	0.06	0.15
	July	220	159	2	47	2.52	3.70	1.79	2.08	2.83	4.23	1.95	2.38	0.31	0.54	0.16	0.29
	October	277	169	1	71	2.27	3.89	1.49	2.28	2.86	4.66	1.75	2.76	0.59	0.77	0.26	0.48
STDEV		**70.04**	**44.53**			**0.74**	**1.60**	**0.55**	**0.78**	**0.66**	**1.48**	**0.59**	**0.76**				

Legend	Percentile of Length, Width	**0%**	**50%**	**100%**	Range (Difference)	**-1**	**0**	**1**
	Color scheme (continuous)				Color scheme (continuous)			

* Removed contrail outbreaks were excluded in calculating the statistics; STDEV stands for the standard deviation.

Construction of UT meteorological data

To infer associations between UT meteorological conditions and the local temporal burstiness and frequency of contrail outbreaks, I used 6-hourly NCEP-NCAR reanalysis (NNR) data that are available online at the website[17] of The NOAA Earth System Research Laboratory (ESRL) Physical Sciences Division (PSD), Boulder, Colorado, USA (Kalnay *et al.*, 1996; Kistler *et al.*, 2001) (see Figure 5-3b). Composite mean and anomaly variables of meteorological factors are available for the entire globe with 2.5° × 2.5° grid resolution at 0:00, 6:00, 12:00, and 18:00 (UTC). Composite anomaly variables are calculated by subtracting the 6-hour mean value of each meteorological variable by the 30-year mean value of the corresponding variable in a grid cell, emphasizing local deviations from normal; here, long term means are based on a 30-year period of 1981-2010. A benefit of using composite anomaly variables for statistical analysis is cancellation of seasonal or monthly variations from regular meteorological variables. Thus, one can compare meteorological conditions between different months of January, April, July, and October without additional data processing removing the seasonality from the data.

I specifically allied anomaly variables of geopotential height at the pressure level of 500 mb, and air temperature, zonal wind (i.e., west-east component of the total wind), omega (i.e., vertical motion of winds), and relative humidity at 300 mb. Each variable is referred to as Z(500), T(300), ω(300), *u*(300), and RH(300) for statistical models. I retrieved the 6-hourly NNR UT variables only for the 2.5° × 2.5° grid cells intersected with each BBOX for the closest time at which the corresponding contrail outbreak occurred, by devising Python scripts that selectively scrape an online text file of NNR data as illustrated in Figure 5-3b. Then, the median of each variable was calculated from the fetched grid cell values and attached to the BBOX as additional

[17] https://www.esrl.noaa.gov/psd/data/composites/hour/

attributes, as depicted in Figure 5-3c. This new approach has two major advantages, in comparison to previous research that analyzed composite NNR UT climate variables fetched based on higher-frequency subregions of contrail outbreaks (e.g., Carleton *et al.*, 2013, 2015): (1) an increase in precision to estimate statistics of UT atmospheric conditions of the area directly affected by contrail outbreaks, and (2) improved flexibility to generate a local summary of each UT climate variable by calculating the composite median of each variable of multiple adjacent BBOXs. The composite median of each UT climate variable is used as an independent variable in GWR models of this study.

Local indicator of temporal burstiness (LITB)

To address the first research question of whether the temporal regularity patterns of contrail outbreaks in the CONUS are bursty or regular over different regions, I applied the LITB (explained in more detail below) to spatial inventories of the outbreaks for January, April, July, and October of 2000-2002 and 2008-2009 (see Figure 5-3e). In the context of contrail outbreaks, a bursty pattern of contrail outbreaks indicates the pattern of time intervals (i.e., inter-event times) between subsequent contrail outbreaks, in which a short time period of many contrail outbreaks, analogically called ‘bursts,’ alternate with a long time period with no contrail outbreaks so that the variance of the inter-event times is very large; in contrast, a regular pattern of contrail outbreaks means that the inter-event times of contrail outbreaks are more or less identical so that their variance is small or close to zero (Barabási, 2005; Goh & Barabási, 2008; Kim & Jo, 2016). As many variables are involved in the dynamics of contrail outbreak incidences, temporal patterns of contrail outbreaks are hard to identify without quantitative analysis. Thus, examining the temporal burstiness patterns of contrail outbreaks can be the beginning of inferring the thermodynamic mechanisms of the formation and persistence of contrail outbreaks.

The LITB is an exploratory spatio-temporal data analysis (ESTDA) statistic designed to identify spatio-temporal patterns of events. It measures the local temporal burstiness of events within a given spatially local area, based on inter-event times of events. Given that σ_τ is the variance of inter-event times and m_τ is the mean of inter-event times, the LITB (*Local A*$_i$) is defined as follows, as proposed in Chapter 4:

$$Local\ A_i(n) \equiv \frac{\sqrt{n+2}r - \sqrt{n}}{\left(\sqrt{n+2}-2\right)r + \sqrt{n}},$$
$$r = \sigma_\tau / m_\tau\,, \qquad 0 \le r \le \sqrt{n}\,,$$

where S is a set of events that occur in the entire region, S_i is a set of events that occur in a local area i where $S_i \subsetneq S$, and $n(S_i) > 1$; $\boldsymbol{\tau}$ is time intervals between two successive events, $e_a, e_b,$ and $\forall\ e_a, e_b \in S_i$. The LITB ranges from -1 to 0 to 1, indicating temporal patterns that are completely regular, completely random, and completely bursty, respectively (see Chapter 4).

Because contrail outbreaks do not occur many times in an individual month, we calculated the local temporal burstiness for each midseason month across five years of 2000-2002 and 2008-2009, allowing observations of seasonal variations. The local area is operationally defined as a geodesic circular buffer generated with the center of a BBOX by the selected bandwidth (i.e., buffer radius), so all the generated buffers have the same area as illustrated in Figure 5-3d. Three bandwidths of 4, 6, and 8 degrees were adopted for both temporal burstiness analysis and GWR analysis to ensure the local area to be larger than most of the BBOXs and cover the regional scale. These bandwidths were applied to GWR models described in the following section. With each bandwidth, *Local A*$_i$ was calculated for each BBOX (i.e., each contrail outbreak); the inter-event times were derived from the BBOXs of the same midseason month whose center is contained in the local area of the contrail outbreak.

GWR models

Because the local temporal burstiness was obtained, it is possible to explore how other environmental factors and the local frequency of contrail outbreaks interplay with the temporal regularity of contrail outbreaks. Multiple GWR models were designed to infer relationships between UT atmospheric conditions and each of the temporal burstiness and frequency of contrail outbreaks at a regional scale and explore the spatial and seasonal variations of those relationships (Brunsdon *et al.*, 1996). As with the temporal burstiness analysis, each of multiple GWR models is applied to each midseason month across five years (see Figure 5-3f).

The structure of GWR model

A GWR model incorporates spatial dependency among locations into a simple linear regression model to resolve the issues of spatial non-stationarity (Brunsdon *et al.*, 1996). GWR modifies the structure of a simple linear regression, or the ordinary least squares (OLS) regression, by accommodating spatial heterogeneity of the coefficients (Brunsdon *et al.*, 1996). The basic GWR equation applied in this study is

$$y(i) = c_0(i) + \sum_{k=1,m} c_k(i)x_k + \varepsilon$$

where $c_k(i)$ is the value of the *k*-th parameter at location *i*, the center of the *i*-th BBOX. The parameters are estimated with the spatial weighting structure for all neighbor BBOXs within a local area of the *i*-th BBOX. The neighbor BBOXs are obtained using a geodesic circular buffer in the same manner used for the local temporal burstiness (Figure 5-3d). Each neighbor BBOX is weighted by distance from the center of the *i*-th BBOX as depicted in Figure 5-3f. To obtain the weight as a function of the distance, I employed a Gaussian kernel, meaning that the further from

the center that the neighbor BBOX is, the lower the weight. Three fixed bandwidths of 4, 6, and 8 degrees are adopted to ensure consistency of the coverage of a local area. These bandwidths are identical to those for the temporal burstiness analysis. This estimation process is repeated for all BBOXs in the same midseason month.

GWR models

No GWR analysis has yet been undertaken in existing research on contrail outbreaks, but it has been done in other meteorological studies (e.g., for rainfall, by Kumari *et al.*, 2017). Five GWR models were designed to examine the following relationships at a local level in the CONUS (see Table 5-3; Figure 5-3f):

A-1. whether each UT meteorological variable of Z(500), T(300), ω(300), *u*(300), and RH(300) has a significant explanatory power for the temporal burstiness of contrail outbreaks (LITB-CO);

A-2. whether each UT meteorological variable of Z(500), T(300), ω(300), *u*(300), and RH(300) has a significant explanatory power for the frequency of contrail outbreaks (Freq-CO);

B. whether including LITB-CO as an independent variable on top of UT meteorological variables of Z(500), T(300), ω(300), *u*(300), and RH(300) improves the prediction performance for Freq-CO;

C-1. whether temporal burstiness is a significant predictor for the frequency of contrail outbreaks;

C-2. whether the frequency of contrail outbreaks is a significant predictor for the temporal burstiness;

The dependent and independent variables for each model are listed in Table 5-3.

Table 5-3. Dependent and independent variables of GWR models

Model	Independent variables	Dependent variables
A-1	Z(500), T(300), ω(300), u(300), RH(300)	LITB-CO
A-2	Z(500), T(300), ω(300), u(300), RH(300)	Freq-CO
B	Z(500), T(300), ω(300), u(300), RH(300), LITB-CO	Freq-CO
C-1	LITB-CO	Freq-CO
C-2	Freq-CO	LITB-CO

Performance evaluation

Models should be evaluated to determine which model best explains data in terms of the structure of the model (e.g., bandwidth, variables) (Kumari *et al.*, 2017). The model performance can be quantified by statistical indicators including the corrected Akaike information criterion (AICc, a statistical model quality metric) and the coefficient of determination (r-squared, R^2) (Akaike, 1973; Cavanaugh, 1997; Nakagawa & Schielzeth, 2013). I used AICc because it adjusts the negative bias of AIC that appears in small-sample applications (Cavanaugh, 1997). Contrail inventories are small size data and the number of data points in a local area is even smaller. The r-squared is the most commonly used 'goodness-of-fit' measure for a regression model (Kumari *et al.*, 2017). The lower AICc implies that the goodness-of-fit of a model is better than a higher one (Cavanaugh, 1997). The r-squared ranges from 0 to 1, and a higher r-squared indicates a better performance of the model (Nakagawa & Schielzeth, 2013).

Results and discussion

Local temporal burstiness of contrail outbreaks in the CONUS

The first research question is whether contrail outbreaks occur temporally regularly or are bursty in each region in CONUS. To answer this question, local temporal burstiness and the local

frequency of contrail outbreaks were calculated for each contrail outbreak. For the calculation, the local area was defined with three bandwidths of 4, 6, and 8 degrees. However, the summary presented in this section is based on results from the 6-degree bandwidth option, because the summary statistics are similar among three options and the 6-degree bandwidth represents the intermediate case. The summary of those statistics is presented in Figure 5-4. The graph shows that the mean local frequency increases from January to October, but the mean local burstiness fluctuates over those months, making a trough in April and a peak in July (Figure 5-4). In general, the temporal patterns of contrail outbreaks are not bursty; it is close to random in April on average. However, in July, the mean local burstiness across the CONUS is about 0.5, indicating that contrail outbreak incidences in July are more irregular in time than in other midseason months.

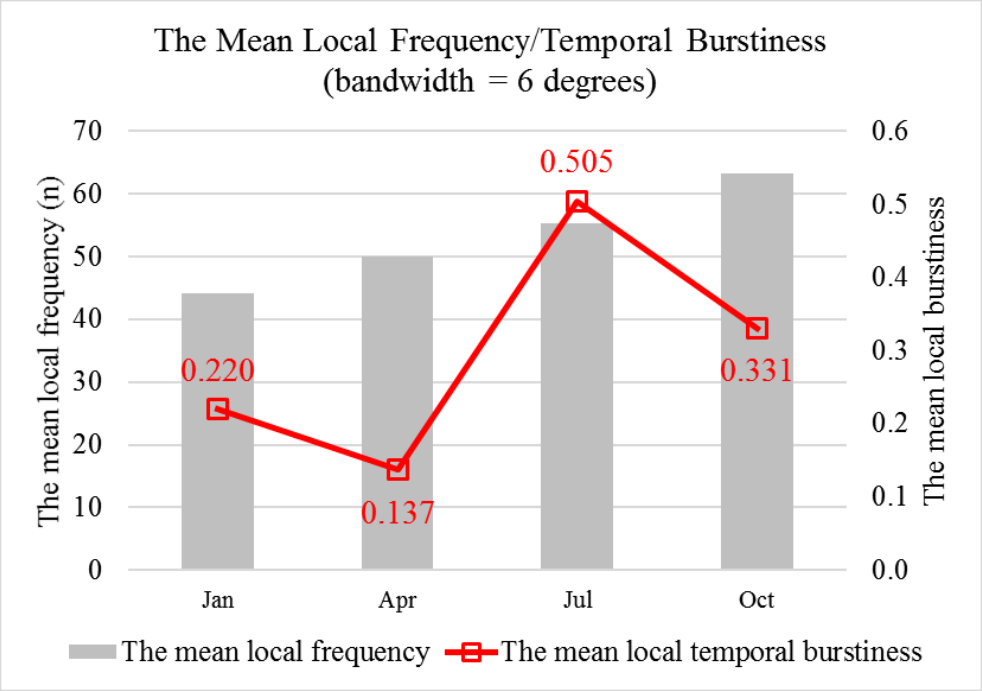
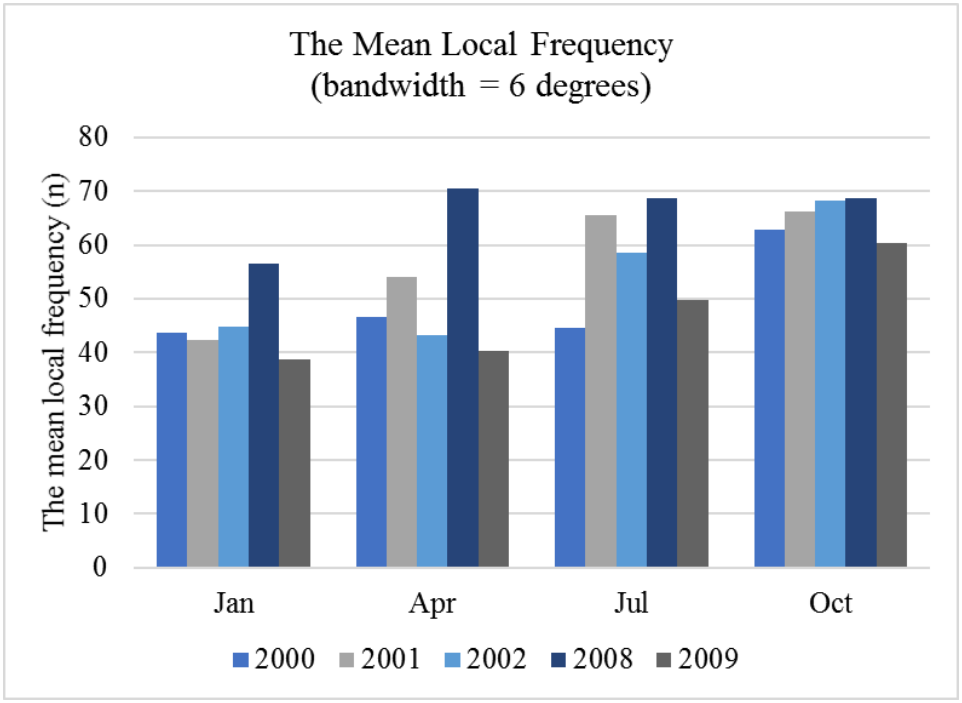


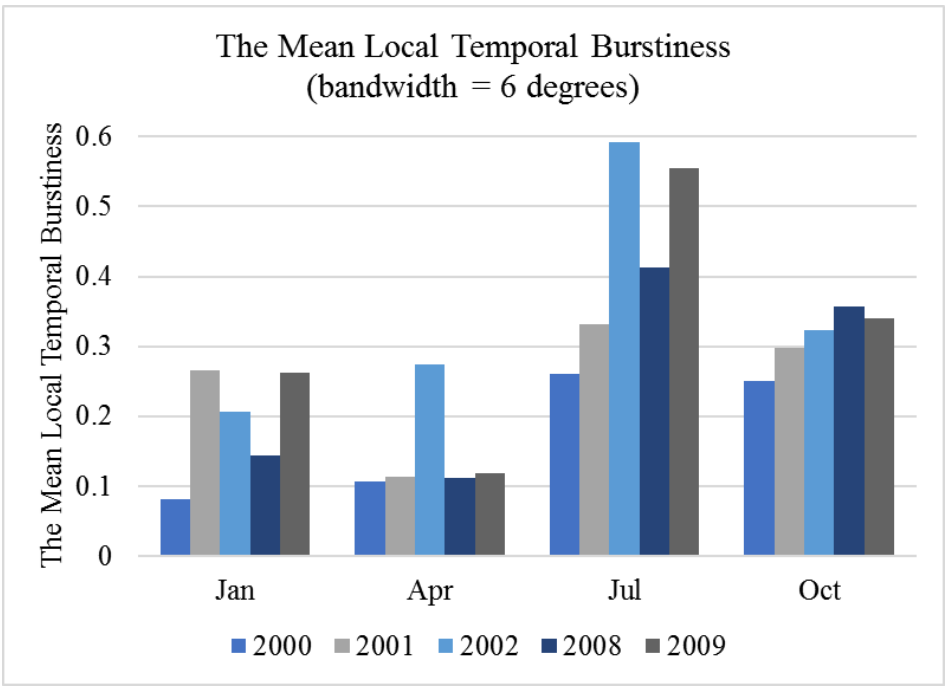
Figure 5-4. The mean local frequency and the mean local temporal burstiness for each midseason month (January, April, July, and October) across five years (2000-2002 and 2008-2009).

To identify the yearly variations of the local temporal burstiness and the local frequency for each season in detail, charts shown in Figure 5-5 were drawn based on the mean and standard deviation (STDEV) of the local temporal burstiness and the local frequency. Although yearly fluctuations in those two statistics exist for each month, the increasing trend in the local frequency from January to October is consistent in each year. Also, the pattern marking a peak of local temporal burstiness in July, in contrast to other midseason months shown in Figure 5-4, coincides with the patterns in Figure 5-5a and 5-5b. In contrast, STDEVs of local temporal burstiness vary across years for the same season (month).

(a)



(b)



(c)

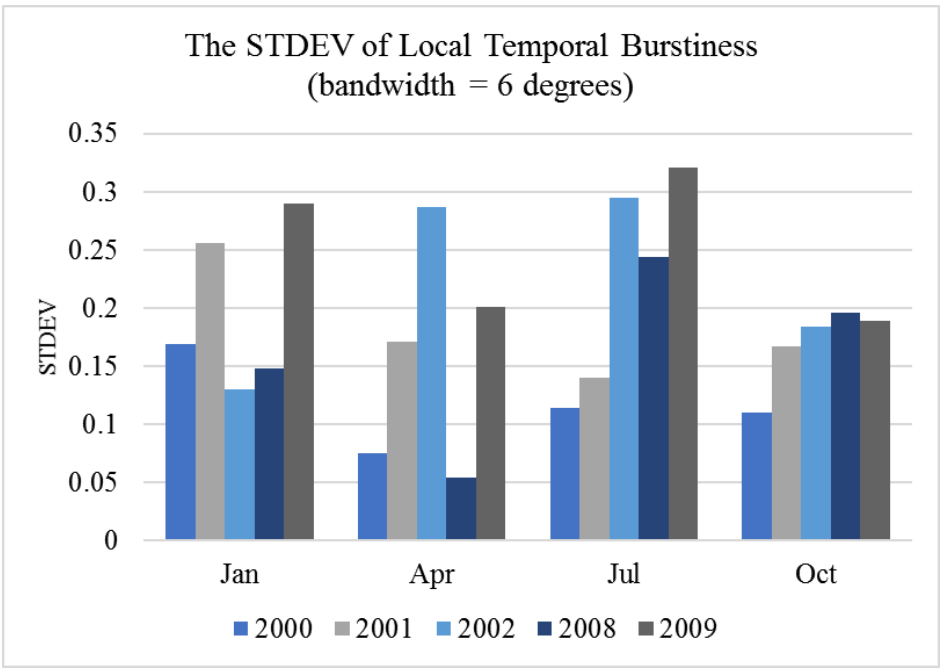


Figure 5-5. Yearly variations in the mean and standard deviation of local temporal burstiness and local frequency over 2000-2002 and 2008-2009 for each month of January, April, July, and October.

The second research question is whether the temporal regularity patterns of contrail outbreak occurrences are spatially heterogeneous across the CONUS. As the local temporal burstiness is obtained for each contrail outbreak annotated with their occurrence location, it is possible to visualize a spatial pattern of the local temporal burstiness, to determine if the pattern is spatially heterogeneous.

Figure 5-6 and Figure 5-7 map the local temporal burstiness for each option of bandwidths of 4, 6, and 8 degrees; the two figures use the same data but in a different fashion according to the adoption of spatial interpolation. The color of each dot in Figure 5-6 is assigned according to the value of the local temporal burstiness (blue for the regions with more temporally regular contrail outbreak occurrences, dark red for the regions with burstier ones). Among different bandwidths, the pattern of the local temporal burstiness calculated with the bandwidth of four degrees exhibits the larger range of values in the local temporal burstiness. Local areas with more regular patterns turn into ones with close to random patterns in the larger bandwidth options.

As illustrated in Figure 5-6a, the East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin) and Southeast CONUS regions broadly exhibit close to random patterns. Then, very bursty patterns appear surrounding those regions, particularly on the north side; more regular patterns appear on the south side.

Similar patterns to those appear in Figure 5-7, but this figure shows more clearly the tendency of the Northwest regions to exhibit local temporal burstiness; contrail outbreak occurrences in this region are much burstier than other regions in the CONUS. California and nearby regions exhibit more or less temporally random patterns of contrail outbreaks.

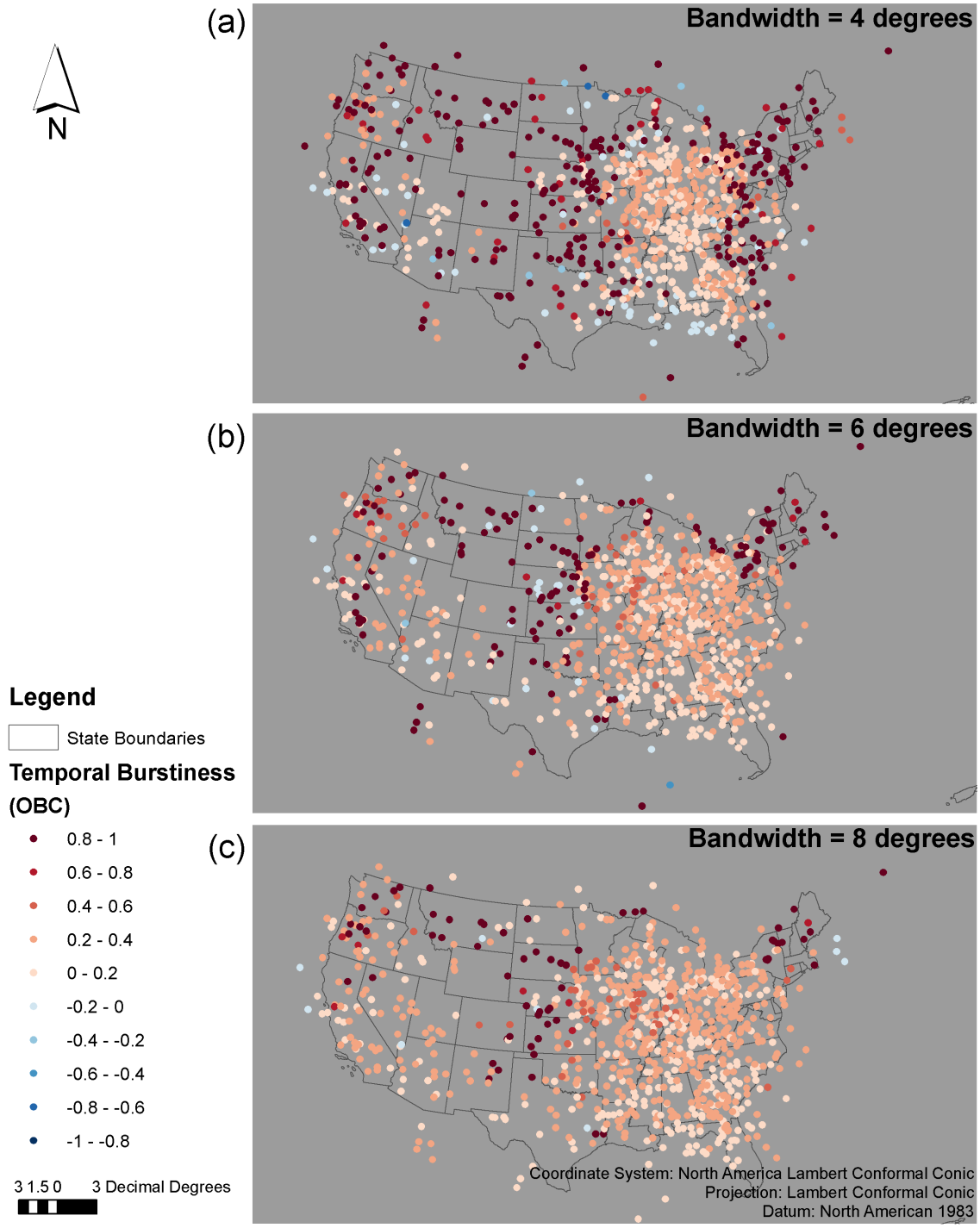


Figure 5-6. Spatial variations in local temporal burstiness in the CONUS (dot = the center of BBOX of contrail outbreaks)

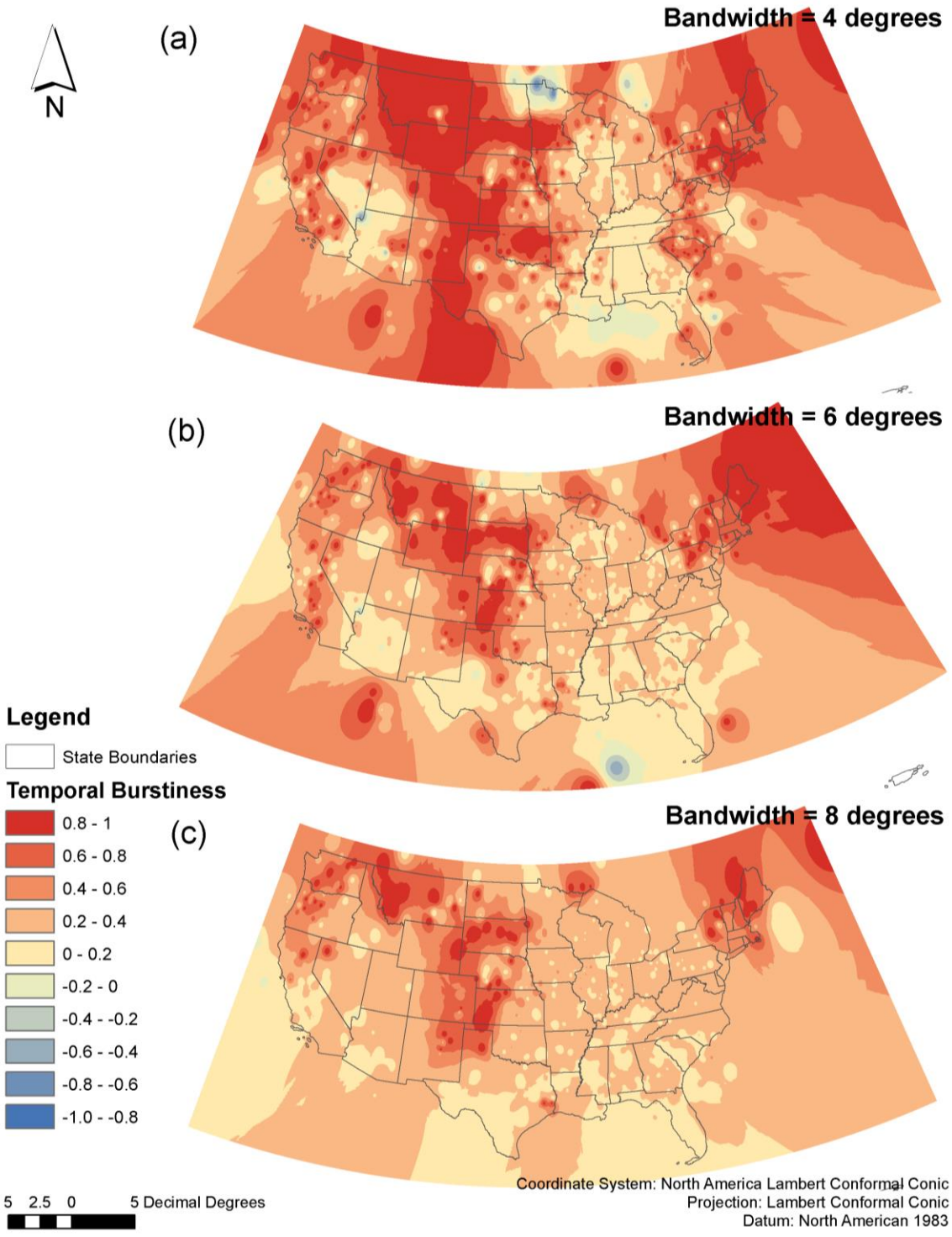


Figure 5-7. Spatially interpolated local temporal burstiness in the CONUS; the spatial interpolation technique adopted for this visualization is IDW.

As noted above, there are seasonal variations in the local temporal burstiness and the local frequency. Accordingly, I also visualized spatial distributions of the local temporal frequency and local temporal burstiness separately for each month of January, April, July, and October (see Figure 5-8 and Figure 5-9, respectively). For the local frequency shown in Figure 5-8, the patterns are similar across different seasons, while there are differences in the strength of the frequency for the region of most frequent contrail outbreaks and the location of the areal center. As seen in Table 5-5a, the area of most frequent contrail outbreaks (red-colored regions) increases from January to October. The most frequent outbreaks are centered in the Southeast region in January and in the East North Central region in other midseason months.

Spatial distributions of the local temporal burstiness (Figure 5-9) are more diverse than those of the local frequency. Regions with random or regular patterns change over different months. In January, those regions are the Southwest and the Southeast; in April, California, the Midwest, the East North Central, the Southeast, the Northeast, and the Atlantic regions; in July, the Southeast, the east side of the Southwest, and Arizona; in October, the Southeast, the upper East North Central near Great Lakes, and Southern California. In April and October, there is a band of higher temporal burstiness over the West and the Southwest. The patterns in July exhibit strong spatial variation; contrail outbreaks then are (very) bursty for most regions except the Southeast.

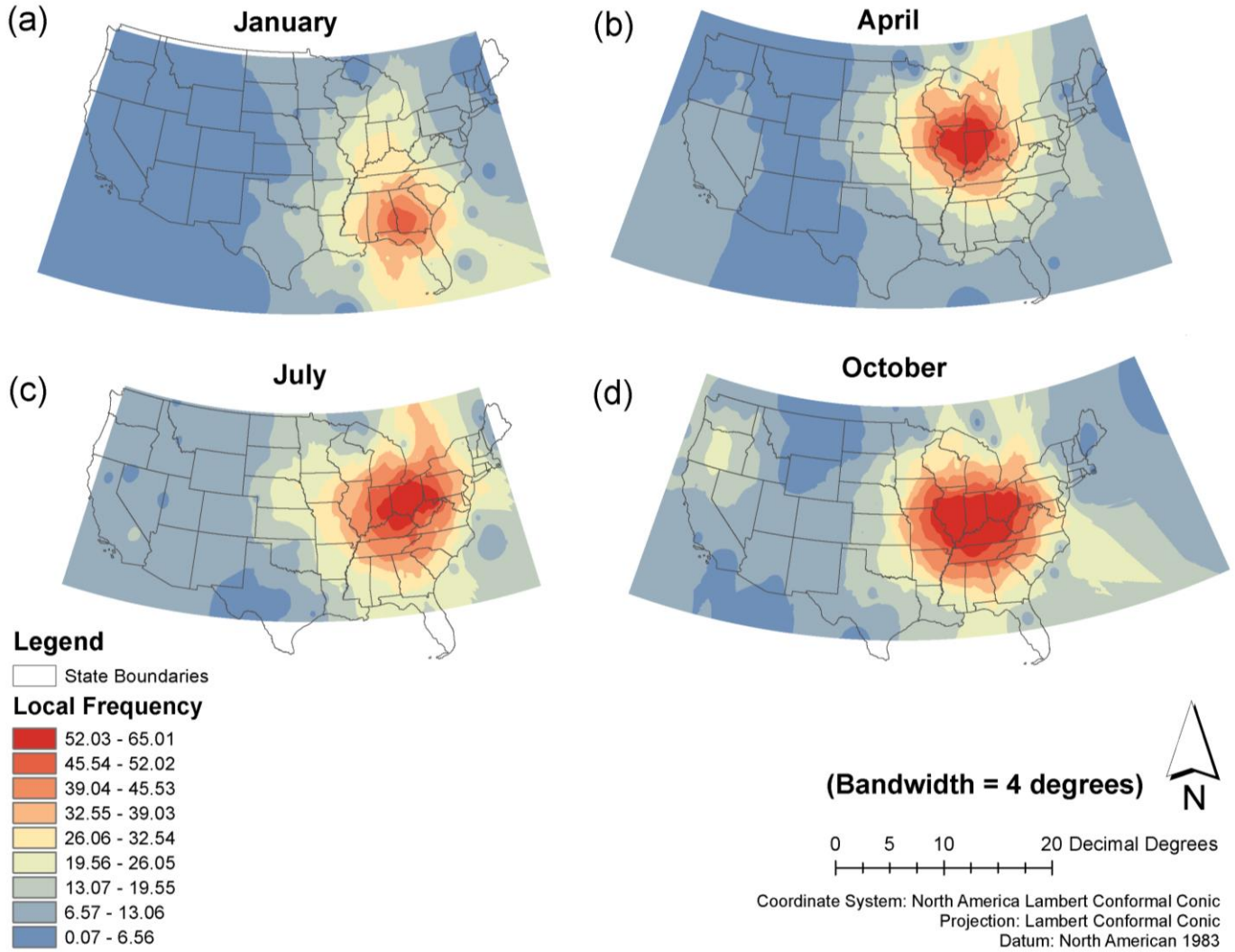


Figure 5-8. Spatial distributions of the local frequency for each month of January, April, July, and October (bandwidth = 4 degrees) (classification methods = quantile)

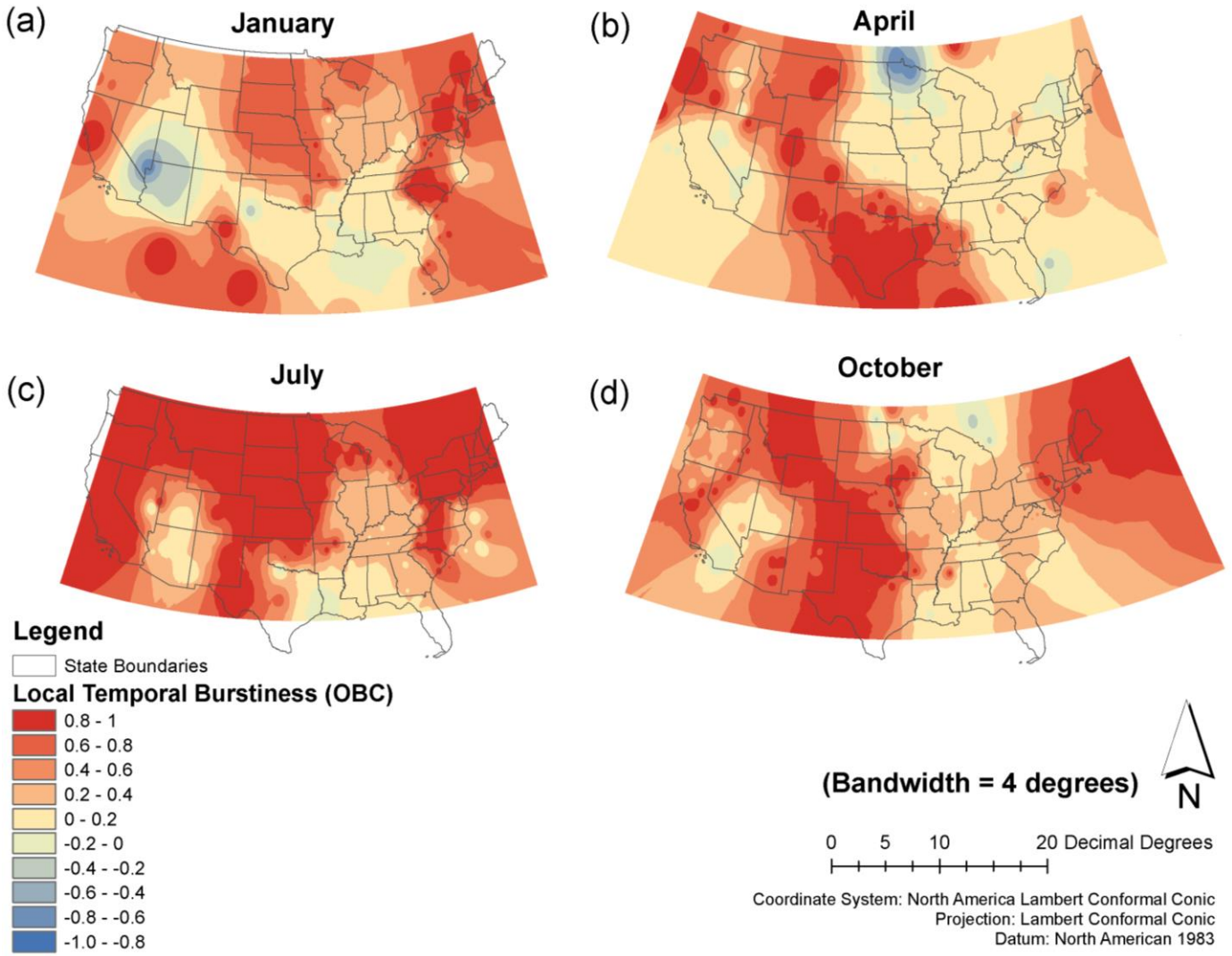


Figure 5-9. Spatial distributions of the local temporal burstiness for each month of January, April, July, and October (bandwidth = 4 degrees).

Results from GWR analysis

For comparison purposes, five GWR models were analyzed and the model performance evaluation of those models is presented in Table 5-4. Thus, now I can answer the research questions listed in the introduction. First, do UT meteorological factors including T, RH, vertical motion of air, horizontal motion of wind, and geopotential height have significant explanatory power for the frequency or the temporal burstiness of contrail outbreaks? According to the model evaluation results (Table 5-4), the answer is both yes and no. Model A-1 described in Table 5-3 reflects the relationships between UT climate variables and the temporal burstiness; Model A-2 is for one between UT variables and frequency. The two models are different in dependent variables that UT climate variables explain. Model A-1 is focused on how UT climate variables affect the temporal burstiness characterizing the temporal regularity of contrail outbreaks; Model A-2 is focused on the impact of UT climate variables on the contrail outbreak frequency. The goodness-of-fit values (AICc, R^2, and R^2 Adjusted) imply that Model A-1 has weak explanatory power for temporal burstiness and Model A-2 has moderate explanatory power for frequency; thus, Model A-2 explains the contrail outbreak frequency at a local scale in terms of UT climate variables.

Second, is the local frequency a significant predictor for the local temporal burstiness of contrail outbreaks? The answer is no, but interestingly, the local temporal burstiness has some explanatory power for local frequency. Thus, the relationship between the local temporal burstiness and the local frequency is asymmetric. The model evaluation for Model B also supports this assertion. In Model B, the local temporal burstiness is added as an explanatory variable on top of other UT meteorological variables to explain a dependent variable, the local frequency. The model performance of Model B is slightly higher than that of Model A-2. This result means that the local temporal burstiness has a small amount of explanatory power for the local frequency of contrail outbreaks.

Table 5-4. Results of GWR model evaluation of contrail outbreaks for the CONUS

Bandwidth	Model	AICc	R^2	R^2 Adjusted
4 degrees	**A-1**	445.350	0.465	0.353
	A-2	6965.879	0.721	0.663
	B	6920.958	0.746	0.685
	C-1	6822.895	0.711	0.691
	C-2	346.102	0.416	0.378
6 degrees	**A-1**	-79.580	0.378	0.317
	A-2	8136.975	0.663	0.630
	B	8119.276	0.673	0.635
	C-1	8099.372	0.647	0.634
	C-2	-42.312	0.289	0.267
8 degrees	**A-1**	-549.453	0.285	0.242
	A-2	8849.346	0.613	0.589
	B	8771.864	0.648	0.624
	C-1	8780.172	0.619	0.612
	C-2	-558.514	0.249	0.236

Third, are relationships represented in each GWR model spatially heterogeneous? The mapped results of GWR analysis indicate yes to this question (see Figure 5-10). Figure 5-10 represents the results of Model B. Positive coefficients are visualized with red color schemes, negative ones are mapped with blue color schemes, and zero coefficients are depicted in light yellow. Positive coefficients denote that an explanatory variable and a dependent variable fluctuate in the same sense.

Figure 5-10 shows that the impact of UT climate variables and the local temporal burstiness on the local frequency is spatially heterogeneous. For instance, the vertical motion of winds at 300 mb, *u*(300) influences the frequency positively in the upper West and negatively in the Midwest (Figure 5-10g). Meanwhile, UT temperature at 300 mb, T(300) shows the opposite pattern (Figure 5-10c); it has positive impacts on the frequency in the Midwest, but negative ones in the east side of the West and the Southwest. Another interesting result is the spatial distribution

of coefficients of the local burstiness. The local burstiness has positive impacts on the frequency in the Northeast and the Atlantic regions and negative impacts in the Midwest and the Southeast regions. This result implies that temporally regular or random patterns in the Midwest and the Southeast may relate to the high frequency of contrail outbreaks in those regions, while bursty patterns in the Northeast and the Atlantic areas may be associated with a lower frequency of outbreaks there. With those spatial variations of each coefficient, the goodness-of-fit of GWR model, B also varies across the different regions (see Figure 5-10a). The model *B* performs better for the West and Midwest regions than for the East North Central and the Southeast regions. Accordingly, it is expected that those regions with the lower model performance potentially have other factors that influence the local frequency of contrail outbreaks in addition to the given explanatory variables. As an example, the spatial density of jet flight traffic can be a potential explanatory variable for the local contrail outbreak frequency (Machta & Carpenter, 1971; Liou *et al.*, 1990; Changnon, 1981; Travis *et al.*, 2007).

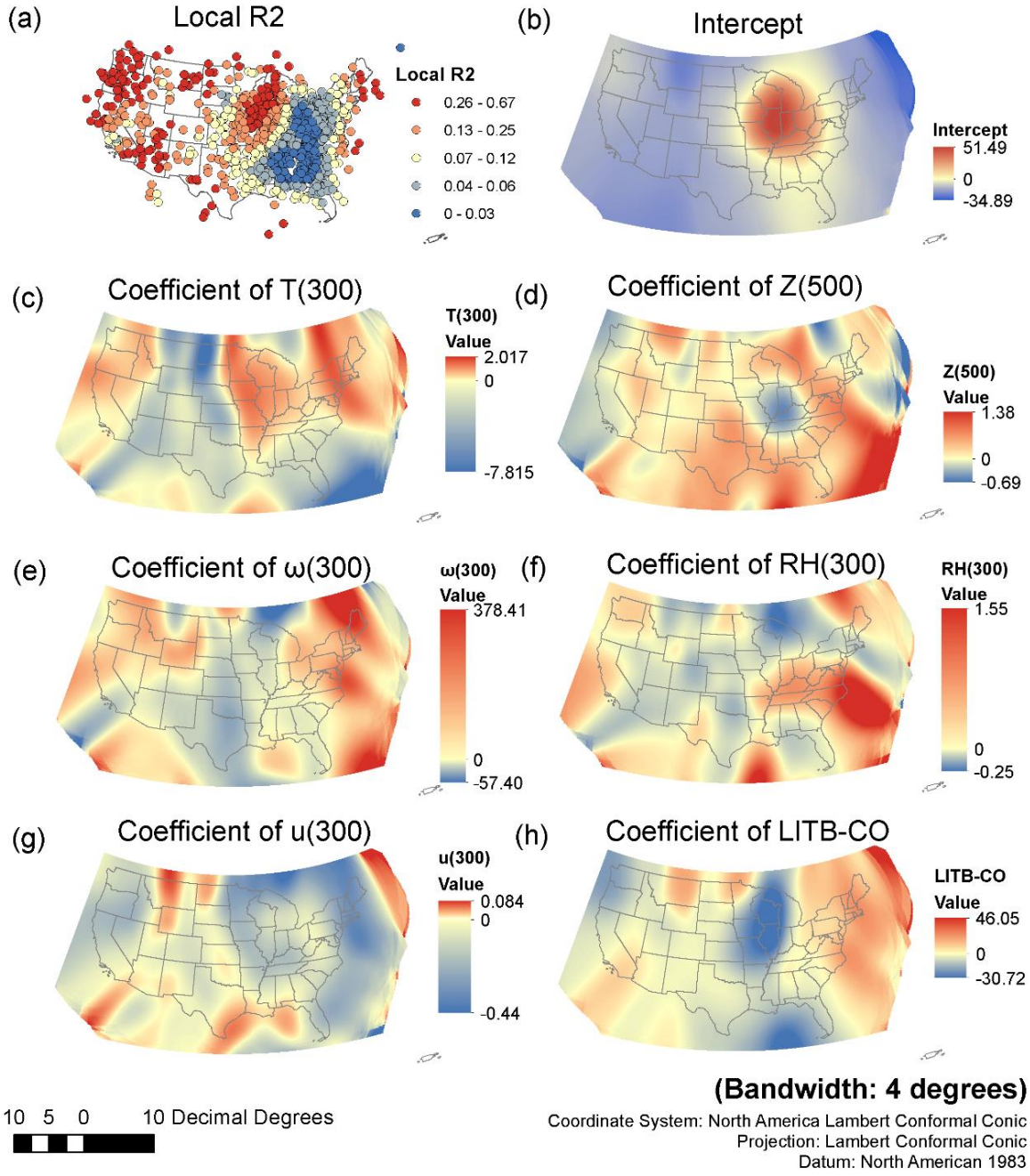


Figure 5-10. Spatial variations in relationships of UT meteorological factors and the local temporal burstiness with the local frequency from GWR model, B - (a) model performance at a local level (Local R2), (b-h) intercept and coefficients of GWR model, B**.**

Summary and conclusions

Comprehending relationships between the spatiotemporal patterns of contrail outbreaks and contributing UT meteorological factors is important for assessing anthropogenic influences of jet aircraft flight activity on climate and its changes at regional scales (Carleton *et al.*, 2013; Carleton & Travis, 2013). Previous studies have focused mainly on the local frequency of contrail outbreaks to examine spatial distributions of contrail outbreaks and the interplay between UT climate conditions and contrail outbreak occurrences. However, a temporal regularity pattern of contrail outbreaks can reveal another aspect of the thermodynamics of contrail outbreak occurrences, but hitherto it had not been explored in contrail research.

To address this research gap, this study applied the LITB proposed in Chapter 4 to characterize the local temporal burstiness of contrail outbreaks in two spatial inventories constructed by Travis *et al.* (2007) and Carleton *et al.* (2013). Then, I explored the spatial variations of relationships among UT meteorological conditions, local frequency, and local temporal burstiness of contrail outbreaks. The results of the burst analysis and GWR analysis can be summarized as follows:

- The local temporal burstiness of contrail outbreaks in CONUS varies with season and region; such seasonal and spatial variations of the local temporal burstiness are highly different from those of the local frequency of contrail outbreaks;
- The local temporal burstiness has a stronger explanatory power for the local frequency than the frequency has for the temporal burstiness;
- The local temporal burstiness and UT meteorological factors (i.e., temperature, relative humidity, zonal wind, vertical motion of air, and geopotential height) are associated with the local frequency;

- Relationships among temporal burstiness, frequency, and UT climate conditions are spatially heterogeneous.

From the results, one can say that the LITB reveals useful attributes of the contrail outbreak climatology. There are several innovations of this study in contrail research: (1) the adoption of the LITB, (2) the application of GWR analysis, (3) automated data collection for meteorological variables by developing Python scripts for web scraping and data processing, and (4) the use of BBOXs and their centers to draw buffers and define the local area. These innovations enabled the analyses presented in this study and can benefit future contrail climatology studies.

This study suggests the following future work to improve the current methodology and explore other characteristics of contrail outbreaks and related environmental factors. First, one could analyze GWR models taking different UT environmental variables as input variables (e.g., the mean values of meteorological variables instead of anomalies) to see which variables result in better model performance. Second, current burst analysis does not address edge effects that are often introduced because the local indicators of spatial association employ the operationally defined local neighborhood structure, such as a spatial buffer drawn from an event location, to measure the local spatial association, and at the edge of the study area, such local neighborhood structure in part covers territory outside of the study area where no observations are presented, potentially leading to biased statistical results (Haase, 1995; Ord & Getis, 1995); one could improve the LITB to address edge effects. Third, the spatial inventories used in this study were constructed manually; in the future, one could develop an advanced automated contrail detection method to build spatial inventories of contrail outbreaks.

Chapter 6

Conclusion

The goal of this dissertation was to develop a suite of methods and tools to characterize temporal burstiness (i.e., frequency-invariant temporal regularity/irregularity) of spatiotemporal events at both global and local scales. The result is a set of exploratory spatio-temporal data analysis (ESTDA) statistics appropriate for use with multiple kinds of spatial containers within which events are aggregated. The potential of these statistics to enhance understanding of the dynamics of geographic phenomena is demonstrated through application in multiple problem domains that include social media posting behavior, wildfire occurrence, and jet aircraft contrails.

The body of this dissertation is presented in four chapters; each chapter has its own goal and objectives, as follows. First, I reviewed literature on burst analysis and ESTDA methods, particularly point pattern analysis methods, and contextualized the concepts and methods for burst analysis in ESTDA (Chapter 2). Second, I proposed a novel burstiness measure to address the small sample size effects of an existing statistical indicator, the Goh & Barabási (2008) burstiness parameter, enabling measurement of temporal burstiness for small data (Chapter 3). Third, I extended the novel burstiness measure proposed in Chapter 3 to the spatial domain by introducing spatial containers into the burstiness measure, developing a statistical significance test for the burstiness measure, implementing algorithms for the proposed methods, and demonstrating their utility through application to wildfire data analysis. Fourth, I conducted a proof-of-concept study on contrail outbreaks in the conterminous United States for mid-season months of 2000-2002 and 2008-2009. Local temporal burstiness was calculated using local indicators of temporal burstiness (LITB) proposed in this research, and its relationships with the contrail outbreak frequency and UT meteorological conditions were identified (Chapter 5).

The studies comprising this dissertation make unique contributions in enriching ESTDA tools by improving and extending an existing burstiness measure from statistical physics for use in analyzing spatiotemporal (geographic) events. The temporal burstiness measure has a unique capability in that it can detect and characterize event sequences with extremely long inactivity periods between events, in comparison to other ESTDA statistics.

Challenges

Although this study advances ESTDA research, there are several challenges in improving and applying the proposed methodology. First, results of the temporal burstiness depend on the definition of events. The proposed burstiness measure assumes that the events are point-based, but a point in space and time cannot represent every event fully. Often, an event lasts long and has its duration with a starting time and ending time. Also, the spatial extent of events can be an area rather than a point, as with the wildfires and contrail outbreaks analyzed in the chapters above. With existing methods that measure the temporal burstiness, as illustrated by the analysis presented in the dissertation, an event is narrowed down to the arbitrary point (e.g., a centroid of an event area, a starting time of an event) for analysis. Ideally, this characteristic needs to be considered in the future improvement of the method. Key questions involve how to take into account extended territories represented by events and determine whether doing so results in a different characterization of burstiness than is obtained by the abstraction of collapsing events to points as done here.

Second, the MAUP is one of most important problems in applying ESTDA statistics, and it is a well-known problem. The MAUP has not yet been explored for the burstiness measure using inter-event times of geographic events as an input variable; thus, doing so remains a future research problem. Some work toward addressing the above issues has, however, been done. In

relation to statistics using inter-event times, Kim *et al.* (2013) explored how the different size of spatial aggregation units affect an exponent of a power-law function fitted to a frequency distribution of fire return intervals (e.g., inter-event times of wildfire events). In this study, my collaborators and I found a pattern that the exponent increases as the size of spatial aggregation units increases. Therefore, in addition to ensuring the functional homogeneity of spatial containers, as suggested in Chapter 4, it is recommended that future research analyzes the sensitivity of the burstiness measure to multiple sizes and forms of spatial containers. Beyond the sensitivity analysis, systematic research on the effect of MAUP on the temporal burstiness measure would contribute to research on burst phenomena.

Similarly, in calculating ESTDA statistics, statistics are distorted in edges of study area, called edge effects, because no data point outside of a study area is included in the analysis. Integrating edge effects into the temporal burstiness measure would reduce the distortion in results of the temporal burstiness.

For a temporal aspect, how to determine the observation duration (i.e., the length of an observed time period of event data) is as critical as the MAUP in the spatial dimension. According to theoretical conditions of the model presented in Chapter 3, event sequences obtained for a highly short duration can be analyzed, if the number of events is equal to or greater than three and the duration is equal to or longer than the sum of all the inter-event times. For empirical data, neither too short nor too long duration is practical.

Most phenomena have their own temporal scale as well as a spatial scale. On one side, the observation duration must exceed the minimum inter-event time of the phenomenon. For tweeting behaviors, it is barely possible to post a subsequent tweet within a millisecond. In the case of the climatological phenomenon of a contrail outbreak, if the outbreak is observed in the same spot within 4-6 hours, it is treated as the same outbreak; thus, the observation duration must at least be longer than 4-6 hours. On the other side, if the duration gets longer, there is a higher

chance of increasing the non-stationarity of behaviors of phenomenon. At the same time, if one can ensure the stationarity, it is better to observe across a duration that is as long as possible because the burstiness measure is specialized in detecting a bursty pattern with both highly short and extremely long inter-event times. One would fail at observing a pattern having extremely long inter-event times if the duration is only moderately long.

Last but not least, the proposed burstiness measure helps detect bursty patterns but it does not confirm any underlying mechanisms that generate bursty patterns. Hence, to infer underlying mechanisms, mathematical or computational modeling or additional data analysis needs to be conducted to support further statements on underlying processes.

Avenues for further study

This research opens many opportunities for future work. The most important immediate area is to extend the temporal burstiness measure to further consider the spatial dimension. The proposed burstiness metric measures the 'temporal' regularity/irregularity of events. In the near future, the concept of bursts needs to be extended to 'spatial' bursts in distance (1D), two-dimensional space (2D), and three-dimensional space or spatiotemporal space (3D). In doing so, the concept of 'spatial bursts' will be established statistically. Kim & MacEachren (2014) made an initial step toward using the Goh & Barabási (2008) burstiness parameter to measure the burstiness in distance by replacing inter-event times by inter-event distances; however, that earlier work needs to be updated with the novel burstiness measure proposed in this study.

Second, another aspect that defines a bursty pattern is *a long-term memory effect* (Goh & Barabási, 2008). This effect implies a long-term interdependence between subsequent events, so the occurrences of events are affected by events that happened a long time ago. In statistics, this

property is measured by temporal autocorrelation coefficients. It is important to integrate this aspect into the proposed burstiness measure in detecting bursty patterns (Goh & Barabási, 2008).

Third, in Chapter 3 the bursty event sequences were modeled with an assumption that there is a single burst—a group of events in a very short time—in the sequence. In reality, there can be multiple bursts. Modeling multi-burst event sequences can potentially achieve a better adjustment of the burstiness measure for small sample size effects.

Four, integrating a de-seasoning method with the current burstiness measure will increase its utility. The de-seasoning method indicates a method of detecting seasonal or cyclic patterns and removing them from data to assist one to find more non-trivial patterns (e.g., Jo *et al.*, 2012a). The seasonality in event occurrences is often treated as a trivial and expected pattern (e.g., Malmgren *et al.*, 2008, 2009) and the focus of exploratory analysis is rather detecting anomalies beyond seasonal or cyclic patterns. The current burstiness measure cannot distinguish the effect of seasonality from bursty patterns. For continuous interval or ratio variables (as was the case for meteorological variables focused on in Chapter 5), adopting composite anomaly variables can eliminate monthly fluctuations in value. For point-based discrete event data, the same de-seasoning approach cannot be applied because the event data do not represent the intensity. Jo *et al.* (2012a) proposed an appropriate de-seasoning method for discrete event data and found that bursty patterns of cell phone corresponding activities still appeared even after human circadian and weekly behavioral patterns were eliminated from input event data. Likewise, integrating an advanced de-seasoning method with the burst analysis can help explore bursty patterns in event sequences.

One potentially fruitful line of follow on research is to leverage resent research on measuring "interestingness" as a strategy to sort out potentially unimportant (but statistically significant) instances of burstiness from those that are meaningful in a particular context. 'Interesting' patterns have been characterized as patterns that are novel, unexpected, non-trivial,

and plausible in the context of data mining (McGarry, 2005). In the field of data mining, interestingness measures have been developed to characterize how interesting a pattern is and there are two categories: objective and subjective measures; the former is on a basis of characteristics of patterns, but the latter is based on one's beliefs or knowledge (Silberschatz & Tuzhilin, 1995; McGarry, 2005). In a visual analytics context, Sacha *et al.* (2018) have recently demonstrated that incorporation of interestingness measures can be effective in providing recommendation cues to analysts when doing exploration of complex relationships.

Time is essential to understanding the dynamics of human and natural phenomena. The methodology proposed by this research benefits researchers who need to explore the temporal regularity of spatiotemporal events that can potentially be bursty, random, or regular in time. This study particularly contributes to exploration of a spatial variation of such temporal regularity. This method can be applied to any kind of events across human and natural phenomena that can be represented as a point in space and time. Potential applications of this approach include cell phone calls, check-ins of places, vehicle accidents, crimes, terrorist attacks, infectious disease outbreaks, rainstorms, and earthquakes, in addition to the applications presented in the dissertation: social media posting behaviors, wildfires, and jet contrail outbreaks.

References

Adèr, H. J., Mellenbergh, G. J., & Hand, D. J. (2008). *Advising on research methods: A consultant's companion*. Johannes van Kessel Publishing.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. & Caski, F. (Eds.). *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Albrecher, H., Ladoucette, S. A., & Teugels, J. L. (2010). Asymptotics of the sample coefficient of variation and the sample dispersion. *Journal of Statistical Planning and Inference*, *140*(2), 358-368.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7), 16-07.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., & Pölitz, C. (2010, October). Discovering bits of place histories from people's activity traces. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (pp. 59-66). IEEE.

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

Bakan, S., Betancor, M., Gayler, V., & Grassl, H. (1994, October). Contrail frequency over Europe from NOAA-satellite images. In *Annales Geophysicae* (Vol. 12, No. 10-11, pp. 962-968). Springer-Verlag.

Banik, S., Kibria, B. M., & Sharma, D. (2012). Testing the population coefficient of variation. *Journal of Modern Applied Statistical Methods*, *11*(2), 5.

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207-211.

Bartier, P. M., & Keller, C. P. (1996). Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences*, *22*(7), 795-799.

Beard, K., Deese, H., & Pettigrew, N. R. (2008). A framework for visualization and exploration of events. *Information Visualization*, *7*(2), 133-151.

Bernhardt, J., & Carleton, A. M. (2015). The impacts of long-lived jet contrail 'outbreaks' on surface station diurnal temperature range. *International Journal of Climatology*, *35*(15), 4529-4538.

Bottiglieri, M., Lippiello, E., Godano, C., & de Arcangelis, L. (2009). Identification and spatiotemporal organization of aftershocks. *Journal of Geophysical Research: Solid Earth*, *114*(B3).

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Briant, A., Combes, P. P., & Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?. *Journal of Urban Economics*, *67*(3), 287-302.

Breunig, R. (2001). An almost unbiased estimator of the coefficient of variation. *Economics Letters*, *70*(1), 15-19.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, *28*(4), 281-298.

Burkhardt, U., & Kärcher, B. (2011). Global radiative forcing from contrail cirrus. *Nature climate change*, *1*(1), 54.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. Statistics & Probability Letters, 33(2), 201-208.

Carleton, A. M., & Lamb, P. J. (1986). Jet contrails and cirrus cloud: A feasibility study employing high-resolution satellite imagery. *Bulletin of the American Meteorological Society*, *67*(3), 301-309.

Carleton, A. M., Silva, A. D., Aghazarian, M. S., Bernhardt, J., Travis, D. J., & Allard, J. (2013). Mid-season climate diagnostics of jet contrail 'outbreaks' and implications for eastern US sky-cover trends. *Climate Research*, *56*(3), 209-230.

Carleton, A. M., Silva, A. D., Bernhardt, J., VanderBerg, J., & Travis, D. J. (2015). Subregion-Scale Hindcasting of Contrail Outbreaks, Utilizing Their Synoptic Climatology. *Journal of Applied Meteorology and Climatology*, *54*(8), 1733-1755.

Carleton, A. M., & Travis, D. J. (2013). Aviation-contrail impacts on climate and climate change: A ready-to-wear research mantle for geographers. *The Professional Geographer*, *65*(3), 421-432.

Carleton, A. M., Travis, D. J., Master, K., & Vezhapparambu, S. (2008). Composite atmospheric environments of jet contrail outbreaks for the United States. *Journal of Applied Meteorology and Climatology*, *47*(2), 641-667.

Carlson, J. M., & Doyle, J. (1999). Highly optimized tolerance: A mechanism for power laws in designed systems. *Physical Review E*, *60*(2), 1412.

Cenek, M., & Franklin, M. (2018). Developing high fidelity, data driven, verified agent based models of coupled socio-ecological systems of Alaska fisheries. In: Perez L., Kim E.-K., Sengupta R. (Eds.). *Agent-Based Models and Complexity Science in the Age of Geospatial Big Data. Advances in Geographic Information Science* (pp. 1-16). Springer, Cham.

Changnon, S. A. (1981). Midwestern cloud, sunshine and temperature trends since 1901: possible evidence of jet contrail effects. *Journal of Applied Meteorology*, *20*(5), 496-508.

Chen, J., Roth, R. E., Naito, A. T., Lengerich, E. J., & MacEachren, A. M. (2008). Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. *International journal of health geographics*, *7*(1), 57.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661-703.

Cui, W., & Perera, A. H. (2008). What do we know about forest fire size distribution, and why is this knowledge useful for forest management?. *International Journal of Wildland Fire*, *17*(2), 234-244.

David, H. A., & Nagaraja, H. N. (2003). *Order Statistics*, John Wiley & Sons. Inc., New York.

DeGrand, J. Q., Carleton, A. M., Travis, D. J., & Lamb, P. J. (2000). A satellite-based climatic description of jet aircraft contrails and associations with atmospheric conditions, 1977–79. *Journal of Applied Meteorology*, *39*(9), 1434-1459.

Delvenne, J. C., Lambiotte, R., & Rocha, L. E. C. (2015). Diffusion on networked systems is a question of time or structure. *Nature communications*, *6*, 7366.

Detwiler, A., & Pratt, R. (1984). Clear-air seeding: Opportunities and strategies. *The Journal of Weather Modification*, *16*(1), 46-60.

Duda, D. P., & Minnis, P. (2009). Basic diagnosis and prediction of persistent contrail occurrence using high-resolution numerical weather analyses/forecasts and logistic regression. Part II:

Evaluation of sample models. *Journal of applied meteorology and climatology*, *48*(9), 1790-1802.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics* (pp. 569-593). Springer New York.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters 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* (pp. 226-231). AAAI Press.

Forkman, J. (2009). Estimator and tests for common coefficients of variation in normal distributions. *Communications in Statistics—Theory and Methods*, *38*(2), 233-251.

Forster, P., Jain, A., Ponater, M., Schumann, U., Wang, W. C., Wigley, T. M. L., Wuebbles, D. J., & Yihui, D. (1999). Potential Climate Change from Aviation. *Aviation and the Global Atmosphere: A Special Report of the Intergovernmental Panel on Climate Change*, 185.

Gabriel, E., & Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, *63*(1), 43-51.

Gandica, Y., Carvalho, J., Aidos, F. S. D., Lambiotte, R., & Carletti, T. (2016). On the origin of burstiness in human behavior: The wikipedia edits case. *arXiv preprint arXiv:1601.00864*.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, *5*(3), 115-146.

Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, *37*(4), 491-496.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189-206.

Gierens, K., Spichtinger, P., & Schumann, U. (2012). Ice supersaturation. In *Atmospheric Physics* (pp. 135-150). Springer Berlin Heidelberg.

Goh, K. I., & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4), 48002.

Haase, P. (1995). Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction. *Journal of Vegetation Science*, *6*(4), 575-582.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Iribarren, J. L., & Moro, E. (2011). Branching dynamics of viral information spreading. *Physical Review E*, *84*(4), 046116.

Jayakumar, G. D. S., & Sulthan, A. (2015). Exact sampling distribution of sample coefficient of variation. *Journal of Reliability and Statistical Studies*, *8*(1), 39-50.

Jensen, E. J., Toon, O. B., Kinne, S., Sachse, G. W., Anderson, B. E., Chan, K.R., Twohy, C.H., Gandrud, B., Heymsfield, A., & Miake-Lye, R.C. (1998). Environmental conditions required for contrail formation and persistence. *Journal of Geophysical Research: Atmospheres*, *103*(D4), 3929-3936.

Jo, H.-H., Karsai, M., Kertész, J., & Kaski, K. (2012a). Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, *14*(1), 013055.

Jo, H.-H., Moon, E., & Kaski, K. (2012b). Optimized reduction of uncertainty in bursty human dynamics. *Physical Review E*, *85*(1), 016102.

Jo, H. H., Pan, R. K., Perotti, J. I., & Kaski, K. (2013). Contextual analysis framework for bursty dynamics. *Physical Review E*, *87*(6), 062131.

Jo, H.-H., Perotti, J. I., Kaski, K., & Kertész, J. (2014). Analytically solvable model of spreading dynamics with non-Poissonian processes. *Physical Review X*, *4*(1), 011041.

Jo, H.-H., Perotti, J. I., Kaski, K., & Kertész, J. (2015). Correlated bursts and the role of memory range. *Physical Review E*, *92*(2), 022814.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., & Zhu, Y. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, *77*(3), 437-471.

Karsai, M., Kaski, K., Barabási, A.-L., & Kertész, J. (2012). Universal features of correlated bursty behaviour. *Scientific reports*, *2*, 397.

Karsai, M., Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A.-L., & Saramäki, J. (2011). Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, *83*(2), 025102.

Keeley, J. E., Safford, H., Fotheringham, C. J., Franklin, J., & Moritz, M. (2009). The 2007 southern California wildfires: lessons in complexity. *Journal of Forestry*, *107*(6), 287-296.

Kim, E.-K., & Jo, H.-H. (2016). Measuring burstiness for finite event sequences. *Physical Review E*, *94*(3), 032311.

Kim, E.-K., Jo, H.-H., & MacEachren, A. M. (2013). Bursts of Wildfires: A Novel Approach to Fire Return Intervals. Paper presented at *the Association of American Geographers (AAG) Annual Meeting, Los Angeles CA, April 2013*.

Kim, E.-K. and MacEachren, A.M. (2014). An index for characterizing spatial bursts of movements: A case study with geo-located Twitter data. In *GIScience 2014 Workshop: Analysis of Movement Data, Vienna, Austria*. URL: http://sites.utexas.edu/amd2014/.

Kistler, R., Collins, W., Saha, S., White, G., Woollen, J., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., & van den Dool, H. (2001). The NCEP–NCAR 50–year reanalysis: Monthly means CD–ROM and documentation. *Bulletin of the American Meteorological society*, *82*(2), 247-267.

Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Saramäki, J., & Karsai, M. (2012). Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, *2012*(03), P03005.

Kivelä, M., & Porter, M. A. (2015). Estimating interevent time distributions from finite observation periods in communication networks. *Physical Review E*, *92*(5), 052813.

Kleban, S. D., & Clearwater, S. H. (2003, May). Fair share on high performance computing systems: What does fair really mean?. In *Cluster Computing and the Grid, 2003. Proceedings. CCGrid 2003. 3rd IEEE/ACM International Symposium on* (pp. 146-153). IEEE.

Knox, E. G., & Bartlett, M. S. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *13*(1), 25-30.

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

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. (1999). Spatial scan statistics: models, calculations, and applications. In Glaz J. & Balakrishnan N. (Eds.), *Scan Statistics and Applications*. Statistics for Industry and Technology. Birkhäuser, Boston, MA.

Kumari, M., Singh, C. K., Bakimchandra, O., & Basistha, A. (2017). Geographically weighted regression based quantification of rainfall–topography relationship and rainfall gradient in Central Himalayas. *International Journal of Climatology*, *37*(3), 1299-1309.

Lamquin, N., Stubenrauch, C. J., Gierens, K., Burkhardt, U., & Smit, H. (2012). A global climatology of upper-tropospheric ice supersaturation occurrence inferred from the Atmospheric Infrared Sounder calibrated by MOZAIC. *Atmospheric Chemistry and Physics*, *12*(1), 381-405.

Lebecki, K. M., Donahue, M. J., & Gutowski, M. W. (2008). Periodic boundary conditions for demagnetization interactions in micromagnetic simulations. *Journal of Physics D: Applied Physics*, *41*(17), 175005.

Lee, D. S., Fahey, D. W., Forster, P. M., Newton, P. J., Wit, R. C., Lim, L. L., Owen, B., & Sausen, R. (2009). Aviation and global climate change in the 21st century. *Atmospheric Environment*, *43*(22), 3520-3537.

Levine, N. (2004). *CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0)*. Houston (TX): Ned Levine & Associates/Washington, DC, National Institute of Justice.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119-133.

Li, Z. (2014). Spatiotemporal pattern mining: algorithms and applications. In *Frequent Pattern Mining* (pp. 283-306). Springer International Publishing.

Liou, K. N., Ou, S. C., & Koenig, G. (1990). An investigation on the climatic effect of contrail cirrus. In Schumann, U. (Ed.). *Air Traffic and the Environment—Background, Tendencies and Potential Global Atmospheric Effects* (pp. 154-169). Springer, Berlin, Heidelberg.

Loveland, T. R., & Merchant, J. M. (2004). Ecoregions and ecoregionalization: geographical and ecological perspectives. *Environmental management*, *34*(1), S1.

Machta, L., & Carpenter, T. (1971). Trends in high cloudiness at Denver and Salt Lake City. *Man's Impact on the Environment*, 410-415.

Malamud, B. D., Morein, G., & Turcotte, D. L. (1998). Forest fires: an example of self-organized critical behavior. *Science*, *281*(5384), 1840-1842.

Malmgren, R. D., Stouffer, D. B., Motter, A. E., & Amaral, L. A. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, *105*(47), 18153-18158.

Malmgren, R. D., Stouffer, D. B., Campanharo, A. S., & Amaral, L. A. N. (2009). On universality in human correspondence activity. *science*, *325*(5948), 1696-1700.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209-220.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, *20*(1), 39-61.

Medewar, P. B. (1964). Is the scientific paper fraudulent. *Saturday Review. August*, *1*, 42-43.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, *94*(2), 284-289.

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

Minnis, P., Ayers, J. K., Nordeen, M. L., & Weaver, S. P. (2003). Contrail frequency over the United States from surface observations. *Journal of Climate*, *16*(21), 3447-3462.

Minnis, P., Bedka, S. T., Duda, D. P., Bedka, K. M., Chee, T., Ayers, J. K., Palikonda, R., Spangenberg, D. A., Khlopenkov, K. V., & Boeke, R. (2013). Linear contrail and contrail cirrus properties determined from satellite data. *Geophysical Research Letters*, *40*(12), 3220-3226.

Minnis, P., Schumann, U., Doelling, D. R., Gierens, K. M., & Fahey, D. W. (1999). Global distribution of contrail radiative forcing. *Geophysical Research Letters*, *26*(13), 1853-1856.

Minnis, P., Young, D. F., Garber, D. P., Nguyen, L., Smith, W. L., & Palikonda, R. (1998). Transformation of contrails into cirrus during SUCCESS. *Geophysical Research Letters*, *25*(8), 1157-1160.

Miritello, G., Moro, E., & Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, *83*(4), 045102.

Moninger, W. R., Benjamin, S. G., Jamison, B. D., Schlatter, T. W., Smith, T. L., & Szoke, E. J. (2010). Evaluation of regional aircraft observations using TAMDAR. *Weather and Forecasting*, *25*(2), 627-645.

Moran, P. A. P. (1947, July). Random associations on a lattice. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 43, No. 3, pp. 321-328). Cambridge University Press.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243-251.

Moritz, M. A., Morais, M. E., Summerell, L. A., Carlson, J. M., & Doyle, J. (2005). Wildfires, complexity, and highly optimized tolerance. *Proceedings of the National Academy of Sciences of the united States of America*, *102*(50), 17912-17917.

Moritz, M. A., Hessburg, P. F., & Povak, N. A. (2011). Native fire regimes and landscape resilience. In *The landscape ecology of fire* (pp. 51-86). Springer Netherlands.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Nelson, T. A. (2012). Trends in spatial statistics. *The Professional Geographer*, *64*(1), 83-94.

Nelson, J. K., & Brewer, C. A. (2017). Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. *Cartography and Geographic Information Science*, *44*(1), 35-50.

Openshaw, S. (1984) *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography No. 38. Geobooks, Norwich, England.

O'Sullivan, D., & Unwin, D. (2003). *Geographic Information Analysis*. John Wiley & Sons.

Palikonda, R., Minnis, P., Duda, D. P., & Mannstein, H. (2005). Contrail coverage derived from 2001 AVHRR data over the continental United States of America and surrounding areas. *Meteorologische Zeitschrift*, *14*(4), 525-536.

Penner, J. E., Lister, D. H., Griggs, D. J., Dokken, D. J., & McFarland, M. (Eds.) (1999). *Aviation and the Global Atmosphere*. Cambridge University Press.

Perotti, J. I., Jo, H. H., Holme, P., & Saramäki, J. (2014). Temporal network sparsity and the slowing down of spreading. *arXiv preprint arXiv:1411.5553*.

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, *84*(3), 441-461.

Radicchi, F. (2009). Human activity in the web. *Physical Review E*, *80*(2), 026118.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability*, *13*(2), 255-266.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 172-212.

Rocha, L. E. C., Liljeros, F., & Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology*, *7*(3), e1001109.

Rodrigues, A. M., & Tenedório, J. A. (2016). Sensitivity analysis of spatial autocorrelation using distinct geometrical settings: Guidelines for the quantitative geographer. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *7*(1), 65-77.

Sacha, D., Kraus, M., Bernard, J., Behrisch, M., Schreck, T., Asano, Y., & Keim, D. A. (2018). SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. *IEEE transactions on visualization and computer graphics*, *24*(1), 120-130.

Sassen, K. (1997). Contrail-cirrus and their potential for regional climate change. *Bulletin of the American Meteorological Society*, *78*(9), 1885-1903.

Schumann, U. (2000). Influence of propulsion efficiency on contrail formation. *Aerospace Science and Technology*, *4*(6), 391-401.

Shalizi, C. (2014). *Power Law Distributions, 1/f Noise, Long-Memory Time Series*. Retrieved on October 10, 2017, from http://bactra.org/notebooks/power-laws.html.

She, B., Zhu, X., & Xiao, W. (2012). Building an integrated web-based environment for exploratory spatiotemporal data analysis. *ISPRS Annals of Photogrammetry, Remote Sensing, and Spatial Information Science*, 169-74.

Silberschatz, A., & Tuzhilin, A. (1995, August). On subjective measures of interestingness in knowledge discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 275-281). AAAI Press.

Slocum, T. A. (1999). *Thematic Cartography and Visualization*, Prentice-Hall, Upper Saddle River, NJ.

Smithwick, E. A., Harmon, M. E., & Domingo, J. B. (2007). Changing temporal patterns of forest carbon stores and net ecosystem carbon balance: the stand to landscape transformation. *Landscape Ecology*, *22*(1), 77-94.

Sokal, R. R., & Braumann, C. A. (1980). Significance tests for coefficients of variation and variability profiles. *Systematic Biology*, *29*(1), 50-66.

Starnini, M., Baronchelli, A., Barrat, A., & Pastor-Satorras, R. (2012). Random walks on temporal networks. *Physical Review E*, *85*(5), 056115.

Stuber, N., Forster, P., Rädel, G., & Shine, K. (2006). The importance of the diurnal and annual cycle of air traffic for contrail radiative forcing. *Nature*, *441*(7095), 864.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*, 234-240.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).

Travis, D. J., Carleton, A. M., & Changnon, S. A. (1997). An empirical model to predict widespread occurrences of contrails. *Journal of Applied Meteorology*, *36*(9), 1211-1220.

Travis, D. J., Carleton, A. M., Johnson, J. S., & DeGrand, J. Q. (2007). US jet contrail frequency changes: influences of jet aircraft flight activity and atmospheric conditions. *International journal of climatology*, *27*(5), 621-632.

Travis, D. J., Carleton, A. M., & Lauritsen, R. G. (2002). Climatology: Contrails reduce daily temperature range. *Nature*, *418*(6898), 601-601.

Travis, D. J., Carleton, A. M., & Lauritsen, R. G. (2004). Regional variations in US diurnal temperature range for the 11–14 September 2001 aircraft groundings: Evidence of jet contrail influence on climate. *Journal of climate*, *17*(5), 1123-1134.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

The United States Geological Survey (USGS). (2014, December). Federal wildland fire occurrence data. In The Federal Fire Occurrence Website. Retrieved January 24, 2015 from http://wildfire.cr.usgs.gov/firehistory/data.html.

Utsu T., Ogata, Y., and Matsu'ura, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, *43*(1), 1-33.

Van Mieghem, P., & Van de Bovenkamp, R. (2013). Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. *Physical review letters*, *110*(10), 108701.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K. I., Kondor, I., & Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*(3), 036127.

Vázquez, A., Rácz, B., Lukács, A., & Barabási, A.-L. (2007). Impact of non-Poissonian activity patterns on spreading processes. *Physical review letters*, *98*(15), 158702.

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. (2016). CyberGIS and spatial data science. *GeoJournal*, *81*(6), 965-968.

Wheatland, M. S., Sturrock, P. A., & McTiernan, J. M. (1998). The waiting-time distribution of solar flare hard X-ray bursts. *The Astrophysical Journal*, *509*(1), 448.

Wiegand, T., & Moloney, K. A. (2004). Rings, Circles, and Null-Models for Point Pattern Analysis in Ecology. *Oikos*, *104*(2), 209-229.

Wiegand, T., & Moloney, K. A. (2013). *Handbook of spatial point-pattern analysis in ecology*. CRC Press.

Wilson, F. (2014). John Stuart Mill. In Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), URL = <http://plato.stanford.edu/archives/spr2014/entries/mill/>.

Yamada, I., & Thill, J. C. (2007). Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis*, *39*(3), 268-292.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, *7*(6), e38869.

Zhang, T., Zhang, Z., & Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in medicine*, *31*(8), 762-774.

Zhao, B., Wang, W., Xue, G., Yuan, N., & Tian, Q. (2015, June). An empirical analysis on temporal pattern of credit card trade. In Y. Tan, Y. Shi, F. Buarque, A. Gelbukh, S. Das, & A. Engelbrecht (Eds.), *International Conference in Swarm Intelligence*, *Lecture Notes in Computer Science* (Vol. 9141, pp. 63-70). Springer International Publishing.

VITA

Eun-Kyeong Kim

Eun-Kyeong Kim earned a Bachelor of Science degree in Geography and a Bachelor of Business Administration degree in e-Business from Kyung Hee University in Seoul, Korea in 2007. She then received a Master of Science degree in Geography, specialized in Geographic Information Systems (GIS) and Cartography, in the same university in 2009. She joined GIS Lab at Kyung Hee University and participated in several research projects as an undergraduate research assistant or a post-master researcher since 2004. In 2012 Eun-Kyeong left GIS Lab and moved to State College, PA in the U.S. to work on her doctoral degree full time at Penn State.

At Penn State, Eun-Kyeong worked on several research projects sponsored by the U.S. government agencies or the National Science Foundation. In 2015-2017, she served as a graduate project manager of the NSF-funded research project on "Building a Big Data Analytics Workforce in iSchools."

Eun-Kyeong has advised numerous undergraduate students during her time at Penn State and served as a co-coordinator of the Undergraduate Research Opportunities Connections (UROC) program. In 2014–2016, she served the College of Earth and Mineral Science (EMS) Graduate Student Council as a Geography representative. During the 2016–2017 academic year she served the Penn State Department of Geography as an elected graduate student representative to the faculty.

Eun-Kyeong has co-organized academic events including seminars, symposia, and workshops at both local and international levels. In February of 2017, she and her colleagues held a Geospatial Data Science Workshop at Penn State.

Eun-Kyeong started working as a postdoctoral fellow in the Department of Geography at the University of Zurich in Switzerland in Spring 2018.