Spring 5-5-2023

Development and Applications of Similarity Measures for Spatial-Temporal Event and Setting Sequences

Fuyu Xu
University of Maine, fuyu.xu@maine.edu

DEVELOPMENT AND APPLICATIONS OF SIMILARITY MEASURES FOR SPATIAL-TEMPORAL EVENT AND SETTING SEQUENCES

By

FUYU XU

M.S. Chinese Academy of Sciences, 1990

M.S. University of Maine, 2003

Ph.D. Michigan Technological University, 2009

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

May 2023

Advisory Committee:

Kate Beard-Tisdale, Professor of Spatial Information Science and Engineering, Advisor

Silvia Nittel, Associate Professor in Spatial Information Science and Engineering

Torsten Hahmann, Associate Professor in Spatial Information Science and Engineering

Kathleen P. Bell, Professor in Environmental Economics and Sustainability

Sean M.C. Smith, Associate Professor in Watershed Process and Sustainability Science

DEVELOPMENT AND APPLICATIONS OF SIMILARITY MEASURES FOR SPATIAL-TEMPORAL EVENT AND SETTING SEQUENCES

By Fuyu Xu

Dissertation Advisor: Dr. M. Kate Beard-Tisdale

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Spatial Information Science and Engineering)

May 2023

Similarity or distance measures between data objects are applied frequently in many fields or domains such as geography, environmental science, biology, economics, computer science, linguistics, logic, business analytics, and statistics, among others. One area where similarity measures are particularly important is in the analysis of spatiotemporal event sequences and associated environs or settings. This dissertation focuses on developing a framework of modeling, representation, and new similarity measure construction for sequences of spatiotemporal events and corresponding settings, which can be applied to different event data types and used in different areas of data science.

The first core part of this dissertation presents a matrix-based spatiotemporal event sequence representation that unifies punctual and interval-based representation of events. This framework supports different event data types and provides support for data mining and sequence classification and clustering. The similarity measure is based on the modified Jaccard index with temporal order constraints and accommodates different event data types. This approach is demonstrated through simulated data examples and the

performance of the similarity measures is evaluated with a k-nearest neighbor algorithm (k-NN) classification test on synthetic datasets. These similarity measures are incorporated into a clustering method and successfully demonstrate the usefulness in a case study analysis of event sequences extracted from space time series of a water quality monitoring system.

This dissertation further proposes a new similarity measure for event setting sequences, which involve the space and time in which events occur. While similarity measures for spatiotemporal event sequences have been studied, the settings and setting sequences have not yet been considered. While modeling event setting sequences, spatial and temporal scales are considered to define the bounds of the setting and incorporate dynamic variables along with static variables. Using a matrix-based representation and an extended Jaccard index, new similarity measures are developed to allow for the use of all variable data types. With these similarity measures coupled with other multivariate statistical analysis approaches, results from a case study involving setting sequences and pollution event sequences associated with the same monitoring stations, support the hypothesis that more similar spatial-temporal settings or setting sequences may generate more similar events or event sequences.

To test the scalability of STES similarity measure in a larger dataset and an extended application in different fields, this dissertation compares and contrasts the prospective space-time scan statistic with the STES similarity approach for identifying COVID-19 hotspots. The COVID-19 pandemic has highlighted the importance of detecting hotspots or clusters of COVID-19 to provide decision makers at various levels with better information for managing distribution of human and technical resources as the

outbreak in the USA continues to grow. The prospective space-time scan statistic has been used to help identify emerging disease clusters yet results from this approach can encounter strategic limitations imposed by the spatial constraints of the scanning window. The STES-based approach adapted for this pandemic context computes the similarity of evolving normalized COVID-19 daily cases by county and clusters these to identify counties with similarly evolving COVID-19 case histories. This dissertation analyzes the spread of COVID-19 within the continental US through four periods beginning from late January 2020 using the COVID-19 datasets maintained by John Hopkins University, Center for Systems Science and Engineering (CSSE). Results of the two approaches can complement with each other and taken together can aid in tracking the progression of the pandemic.

Overall, the dissertation highlights the importance of developing similarity measures for analyzing spatiotemporal event sequences and associated settings, which can be applied to different event data types and used for data mining, sequence classification, and clustering.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Kate Beard, for her unwavering support and guidance throughout my academic journey. Her dedication to collaborative scholarship, generous spirit, and mentorship have been instrumental in helping me achieve my academic and personal goals. I cannot thank her enough for her patience, persistence, and belief in me, even during the most challenging times in the past several years. I also want to extend my heartfelt appreciation to the members of my dissertation committee, Dr. Silvia Nittel, Dr. Torsten Hahmann, Dr. Kathleen P. Bell, and Dr. Sean M.C. Smith, for their invaluable feedback and insights that shaped the direction and quality of my research. Their expertise and commitment to excellence have inspired me to strive for academic excellence.

I am grateful to other faculty and staff from the School of Computing and Information Science for their support and valuable resources that helped me navigate through the various stages of my academic career. Furthermore, I would like to thank Dr. Marcella H. Sorg, Dr. Daniel Soucier and other colleagues in Margaret Chase Smith Policy Center (MCSPC) for their insightful discussions and collaborations that enriched my research experience.

I am also deeply grateful to the Senator George J. Mitchell Center for Sustainability Solutions for their financial support in the first three years and the Graduate School of the University of Maine for offering me another two years' Data Science & Engineering Fellowship, which provided me with the resources to pursue my research goals.

Lastly, I would like to acknowledge the unwavering support and encouragement from my family, including my wife, Qiuli (Julia) Wang, and my two sons, Keji Xu and

William Xu. Their love, patience, and understanding have been my source of strength throughout my academic journey. I owe them my deepest gratitude for making this achievement possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The analysis of spatial-temporal data important to many fields, including environmental science, transportation, climate, ecology, and economics. Spatial-temporal data can be extracted and represented by event sequences or trajectory data, which can be further used to identify meaningful patterns and relationships between different events or event sequences. Several spatial-temporal analyses rely on the development of efficient and effective similarity measures to compare and cluster the event sequences or trajectories. In addition, the analysis of corresponding event settings, which are places and related influencing factors in which events occur, can facilitate insights into how events evolve. Similarity measures between sequences of event settings provides contextual information that allows researchers to compare and analyze the similarity of different event sequences in a standardized and quantitative way. The combination of event sequence similarity and setting similarity metrics offers an expanded approach for variety of fields where researchers need to analyze event sequences that occur in space and time, such as sociology, criminology, and public health, along with ecology, biology, and many others.

1.1. Motivation

Wireless sensor networks (WSN) and other traditional well-established monitoring systems in many fields have been generating large volumes of time series data, and thus there is an increased need for more efficient processing and management of such data. Examples include habitat monitoring (Mainwaring et al., 2002), environmental monitoring

for air pollution, water quality, and weather forecasting (Nittel, 2009; Oliveira and Rodrigues, 2011; Othman and Shazali, 2012), active volcano monitoring (Werner-Allen et al., 2006), security monitoring (Bartariya and Rastogi, 2016), spatiotemporal risk assessment through monitoring urban hazard events (Wang et al., 2016b), and near real time disaster monitoring (Hu, 2016), etc. The information in a time series is not all of equal value. Assuming a time series represents some underlying process, various changes in the time series can indicate changes of state or effectively "events" that signal some information worthy of greater attention (Beard et al., 2008; Beard et al., 2011; Fu, 2011; Rude and Beard, 2012). Extracting "events" from time series thus corresponds to the identification of important changes of state (Andrienko et al., 2010). When the sensor systems are distributed in space, the set of time series and events from the sensor locations can convey information on changes of state in a spatiotemporal process.

All events inherently have the context of location and time, or spatial and temporal attributes. Many events of interest, particularly abnormal events, such as fraud transactions in banking activities, earthquakes, outbreaks of diseases, air and water pollution, and forest fires, have tremendous impact on our everyday lives, the environment in which we live, and many require critical decision making by various organizations, communities, and business sectors. Many events interact with each other and show certain spatiotemporal patterns which may lead us to better understand the mechanisms behind them. Therefore, it is very important for scientists to be able to quickly and efficiently detect events, event patterns and their relationships that are critical in specific domains.

The challenges created by massive event data volumes of different sources and complex relations between events, require quick, efficient and effective approaches for

knowledge discovery. In addition, in many domains, pattern-based knowledge or spatiotemporal context information is explicitly needed. Research in spatial epidemiology, for example, has been increasing rapidly in the past twenty years with the introduction of spatial and spatiotemporal hierarchical models (Banerjee et al., 2014; Elliot et al., 2000; Elliott and Wartenberg, 2004). By monitoring epidemic outbreaks based on reported disease data and characterization of disease spread patterns, spatial epidemiology can provide epidemic spread characterization and alerts to the public. Big data analytics tools for real time processing systems are increasingly needed across a wide range of domains (Dutta and Jayapal, 2015). Event-based analytics contribute to analysis and refinement of hypotheses about what happens in a specific time period or in a specific region. Quick and flexible detection of event data and event sequence analysis are of special interest to many users and experienced analysts as they provide information for further aggregation, visualization, and analysis. Event or event sequence pattern-based knowledge can serve as a strong basis for decision making.

Spatiotemporal event sequences (STES) are temporally ordered events of one or different types distributed over space. Awareness of the similarities between these event sequences can be important for many fields and organizations because many situations, operations, and activities are rich in different events that follow a sequential or certain order. With the appropriate methods and tools, we can compute pairwise similarities between event sequences and detect some patterns or rules of interest or importance in event datasets so that we can make certain predictions or recommendations for business operations and environmental monitoring based on detected similar event sequences. Examples of event sequence-based data can be found in medical records, traffic incident

data, historical and biographical data, administrative process data (Vrotsou and Forsell, 2011), telecommunications data, web access data (Mannila and Salmenkivi, 2001), and environmental monitoring data (Cardell-Oliver et al., 2004; Padhy et al., 2005).

While discovering patterns from event sequences is important, there are many application domains that require understanding the event settings or contextual factors. For instance, spatial context strongly influences the transport disadvantage that in turn affects social exclusion and well-being (Delbosc and Currie, 2011). In a travel behavior research, spatial context is strongly related to the household travel patterns in an international scale (Timmermans et al., 2003). A person's health-related problems are strongly affected by different types of spatial context, such as environmental exposures (Cutter, 1996; Roux and Mair, 2010), social environment (characteristics of communities and neighborhoods) (Roux and Mair, 2010; Sampson, 2003), and ease of access to health services (Yang et al., 2006). Spatial context greatly influences the potential of getting a disease, the adoption of healthy lifestyle, and the ease of access to medical services for disease diagnosis and treatment. Consideration of the spatial event settings and different types of contextual variables is of particular importance in ecology and environmental application areas (Vanderbilt et al., 2015). Therefore, it is also important to further study the representations and similarity measures of event settings when comparing the similarity or distance measures between spatiotemporal event sequences.

We are also motivated by a case study on investigation and monitoring of impaired coastal water quality conditions. Impaired water quality can be viewed as episodic events triggered by the occurrence of heavy rainfall, harmful algal blooms, wastewater treatment facility malfunctions, etc. These episodic events in impaired water quality can in turn

trigger management events such as closures of shellfish growing areas and posting of beach advisories. The case study setting for this thesis is aimed at improving our understanding of the spatial and temporal dynamics of fecal pollution events in Maine coastal waters. The relationship between fecal pollution and other spatiotemporal events along the coastal region is not well understood. We envision an integrated conceptual framework and eventually a pipeline for extracting event sequence data from time series whether as incoming data streams from WSN or other mobile devices or from historical databases. While identification of specific events as higher-level abstractions of time series is useful, the central data object of interest in this research is a spatial temporal event sequence.

1.2. Research Setting and Problems of Interest

1.2.1. Research Setting

Spatiotemporal event sequences (STES), as understood for this thesis, are temporally ordered events of one or more different event types collected from fixed locations in space as illustrated in Figure 1.1. Wireless sensor network (WSN) or other monitoring systems with sensors deployed regularly or irregularly in geographic space provide the problem setting. Each WSN node or platform can have many sensors measuring different variables and producing a time series on each variable, which we refer to as a space-time series. Abstracting these space-time series to event sequences leads to on-going production of STES at each monitoring station. Collectively we can imagine the STES at monitoring stations forming a field of event sequences of a type related to each observed variable. This conceptualization of STES differs from other types of event sequences such as genomic sequences (Darling et al., 2004), industrial process monitoring sequences (Maurya et al., 2007), patient symptom sequences (Tao et al., 2012) or consumer purchasing sequences

(Prinzie and Van den Poel, 2011) in that STES derive from specific fixed geospatial locations.

Within this setting, we can view an individual STES or the collection of STES across a set of monitoring stations as reflecting an evolving underlying process (Yang et al., 2014).

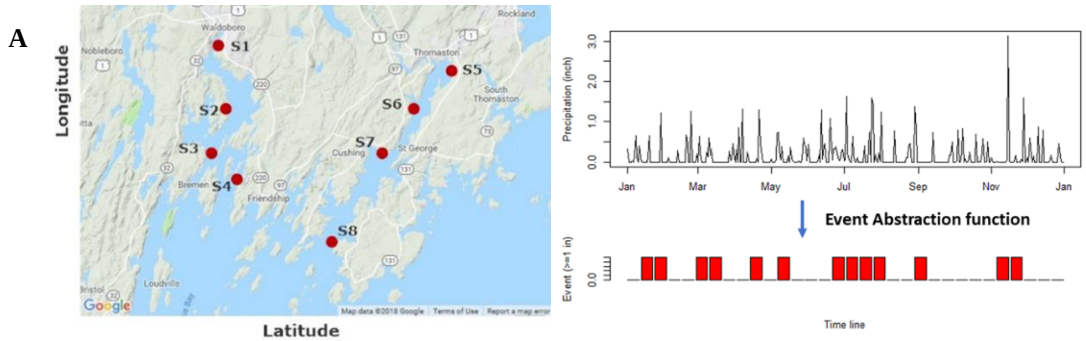


Figure 1.1. Thesis problem setting. (A) An example of fixed locations of interest or observation sites distributed along the coast. (B) An example of a spatiotemporal event sequence extracted from a spatial time series of precipitation with the threshold of ≥1.0 inch/24hr. Red bars stand for individual events. The sequence consists of the events along with the interval spacing between events.

In other words, an STES is viewed as a data abstraction representing a realization of a process. For example, we could say a precipitation event sequence observed at station S1 (Figure 1.1) represents a local realization of a meteorological process.

Extending Tobler's first law of geography, often interpreted as values near-by in space being more similar than values which are more distant, this thesis proposes to examine the following general scientific questions: Are processes that are occurring near-by in space more similar than those occurring at a distance? and Are processes that occur in similar spatial settings more similar than those that occur in less similar spatial settings? We propose to investigate these questions of process similarity through an examination of STES similarity. To this end, the thesis proposes to develop a suite of analytical approaches for similarity assessment of STES that takes into consideration the temporal, thematic, and

spatial dimensions of the STES. Direct questions to be examined are: Do STES (obtained from monitoring stations) that are closer in space tend to be more similar temporally and thematically? Do STES (obtained from monitoring stations) that have similar spatial setting, or a similar collection of contextual parameters tend to be more similar than STES from less similar spatial settings?

1.2.2 Problems of Interest

This study is not just about events in isolation which are of interest but the sequence and pattern of their occurrences (the interval spacing between events) in conjunction with their patterns of occurrence in space. We seek to discover and understand patterns in these event sequences. This study considers the problem of comparing the similarity between sequences of events obtained at fixed spatial locations such as the situation in WSN. The goal of the study is to develop a suite of analytical approaches for similarity assessment of spatial temporal event sequences and corresponding spatial event setting parameters that supports discovery of process interactions.

In this dissertation, we propose a new similarity measure for spatial-temporal event sequences and investigate its scalability and effectiveness in comparison to other existing methods. Additionally, we explore the development of similarity measures for spatiotemporal setting sequences and conduct a case study to compare its performance with event sequence similarity.

We propose to first examine the temporal and thematic similarity between STES at locations as depicted in Figure 1.1. Since each sequence is associated with a location, the similarity metric provides the basis for similarity assessment across locations. i.e., we have

a measure of how similar location S1 is to S2 in terms of their sequences as illustrated in Figure 1.2. While we might expect event sequences, generated from locations that are close in space, to be similar, we also want to consider and test the possibility that event sequence similarity may not be just a function of spatial proximity but of spatial setting similarity. We address spatial similarity further through development of a spatial setting similarity metric (Figure 1.2). We investigate the spatial setting similarity of monitoring stations in terms of both static and dynamic variables. The variables for consideration are domain dependent. For a water quality monitoring application, for example, the static spatial setting variables of interest could include land cover, topography, and soils. What constitutes a spatial setting has scale implications. A setting is envisioned as some area of influence around a monitoring station with different sizes, or configurations of these settings dependent on the analysis domain and objective.

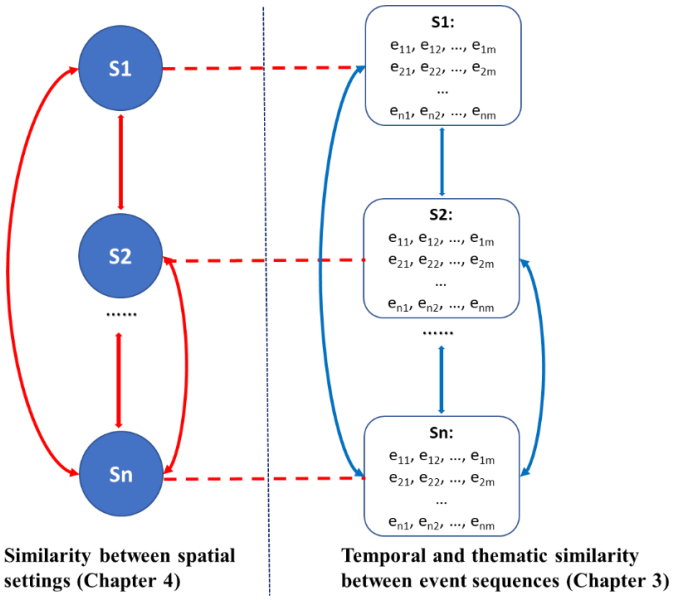


Figure 1.2. Illustration of different types of similarity measures covered in Chapter 3 and 4. Chapter 3 sets up basis for examining the temporal and thematic similarity between event sequences for fixed locations (S1, S2, …, Sn). Chapter 4 examines the similarity between spatial settings or contextual parameters of corresponding fixed locations.

There has been substantial research on methods for measuring similarity between event sequences (Gundersen, 2012; Mannila and Moen, 1999; Mannila and Ronkainen, 1997; Obweger et al., 2010; Wongsuphasawat et al., 2012). However, to date, this event sequence analysis research has not addressed event sequence similarity in a spatial setting without considering contextual factors. In addition, there exist many important issues that need to be solved in current available similarity measures for event sequences such as high computational expense, difficulty of handling different types of events occurring at the same timestamp within one sequence, and consideration of both point and interval events. This thesis aims to address these gaps.

The identified problem setting also differs from previous event sequence similarity research. The context for most sequence similarity problems has generally been posed as a database query problem where a target event sequence is compared against a (potentially very large) database of event sequences to find the closest match or k matches. For the problem setting of this thesis, we assume a set of STES associated with some geographic region and the problem is to determine measures of similarity between this fixed set. We propose a matrix data structure for this problem setting as a fast and efficient method for computing STES similarity. We adopt a matrix structure for first computing the temporal and thematic similarity between STES. Similarly, a matrix data structure will be applied to compute a multivariate similarity measure for spatial setting similarity. In a third extension of the matrix structure, we propose to apply it to a multivariate sequence setting.

With the development of these similarity measures, the thesis aims to address the following research questions:

1. Given a set of STES composed of either point events or interval events, how similar are these sequences in space and time?
2. Given a set of Event Settings represented by a group of static variables and the other group of dynamic variables over time, how similar are these sequences of Event Settings in space?
3. Are similar STES correlated with similar sequences of spatial settings?
4. How well are these similarity measures for STES and sequences of spatial settings fit into existing clustering methods (or classification and data mining, etc.)? And, can these results from clustering methods be validated with ground truth dataset?

1.3. Research Contributions

The main purpose of this dissertation research is to develop more effective and more efficient similarity measures of ST-event sequences and corresponding spatial settings. The expected contributions are as follows:

1. Novel approaches of measuring similarities suitable for STES of both point and interval events with the option of considering quantitative levels of individual events and filling the gap for investigating the similarities between STES of different types of events.
2. Efficient and faster matrix-based similarity measures for STES of different types of events compared to existing similarity measures.
3. Proposed the concept of "Setting Sequences" or "Sequences of Spatial-temporal Settings": Setting sequences refer to the ordered list of event settings that events occur over time, corresponding to STES. Setting sequences can be used to

analyze the context in which events occur and can provide insights into how the environment influences events.

4. Formalization of how to quantify spatial setting or contextual variables of corresponding STES of interest and set-up of guidelines for selecting parameters of spatial settings.
5. Expected better or complementary results of classification and clustering analysis from STES datasets of real world compared with using existing methods.
6. Novel approach of combining similarity measures of both STES of interest and corresponding spatial context for knowledge discovery from STES datasets.

1.4. Intended Audience

The intended audience of this dissertation primarily includes researchers, application system developers, environmentalists, and decision makers from business sectors and government organizations who are interested in studying spatiotemporal information from event sequences.

The audience also covers researchers from the fields of computer science whose research focus on semantics, formal spatial models, and computational algorithms.

1.5. Organization of the Remaining Chapters

The remaining chapters of this dissertation are organized as follows. Chapter 2 gives a brief review and background information on events and event sequences and in particular emphasizes sequence similarity measures and roles of corresponding spatial context and setting. Chapter 3 defines primitive ST-events in a spatiotemporal setting, demonstrates

how to construct spatiotemporal event sequences (STES) from spatial time series, and develops a generic method to compute similarity between STES of different types of events. This chapter also set up a solid foundation for calculating similarity between STES from different data types. Chapter 4 extends the sequence similarity measures from Chapter 3 to address spatial settings and contextual variables in a quantitative sense for event sequences and proposes a similarity measure for event setting sequences that incorporates dynamic variables alongside static variables. This chapter also aligns the similarity metric development with a case study involving setting sequences and pollution event sequences associated with the same monitoring stations. This case study supports the hypothesis that more similar spatial-temporal settings or setting sequences may generate more similar events or event sequences. Chapter 5 explores the scalability and extended application of STES similarity measures developed in Chapter 3 through a case study of COVID-19 surveillance based on clustering results and evaluates its advantages with a well-established space-time scan statistic approach. Chapter 6 summarizes the studies conducted in this dissertation and discusses some future directions.

Chapter References

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., and Poelitz, C., Extracting events from spatial time series, *in* Proceedings Information Visualisation (IV), 2010 14th International Conference2010, IEEE, p. 48-53.

Banerjee, S., Carlin, B. P., and Gelfand, A. E., 2014, Hierarchical modeling and analysis for spatial data, Crc Press.

Bartariya, S., and Rastogi, A., 2016, Security in wireless sensor networks: Attacks and solutions: environment, v. 5, no. 3.

Beard, K., Deese, H., and Pettigrew, N. R., 2008, A framework for visualization and exploration of events: Information Visualization, v. 7, no. 2, p. 133-151.

Beard, K., Emerson, J., Deese, H. E., Rude, A., Scott, M., and Pettigrew, N. R., 2011, Use of the EventViewer for visualizing and exploring events extracted from Ocean Observing System Data: Marine Technology Society Journal, v. 45, no. 1, p. 112-124.

Cardell-Oliver, R., Smettem, K., Kranz, M., and Mayer, K., 2004, Field testing a wireless sensor network for reactive environmental monitoring.

Cutter, S. L., 1996, Vulnerability to environmental hazards: Progress in human geography, v. 20, no. 4, p. 529-539.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T., 2004, Mauve: multiple alignment of conserved genomic sequence with rearrangements: Genome research, v. 14, no. 7, p. 1394-1403.

Delbosc, A., and Currie, G., 2011, The spatial context of transport disadvantage, social exclusion and well-being: Journal of Transport Geography, v. 19, no. 6, p. 1130-1137.

Dutta, K., and Jayapal, M., Big data analytics for real time systems, *in* Proceedings Big Data Analytics Seminar2015, p. 1-13.

Elliot, P., Wakefield, J. C., Best, N. G., and Briggs, D., 2000, Spatial epidemiology: methods and applications, Oxford University Press.

Elliott, P., and Wartenberg, D., 2004, Spatial epidemiology: current approaches and future challenges: Environmental health perspectives, v. 112, no. 9, p. 998.

Fu, T.-c., 2011, A review on time series data mining: Engineering Applications of Artificial Intelligence, v. 24, no. 1, p. 164-181.

Gundersen, O. E., Toward measuring the similarity of complex event sequences in real-time, *in* Proceedings International Conference on Case-Based Reasoning2012, Springer, p. 107-121.

Hu, H., 2016, Online Near Real-time Mine Disaster Monitoring System Based on Wireless Sensor Networks: International Journal of Online Engineering, v. 12, no. 3.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., and Anderson, J., Wireless sensor networks for habitat monitoring, *in* Proceedings Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications2002, Acm, p. 88-97.

Mannila, H., and Moen, P., Similarity between event types in sequences, *in* Proceedings International Conference on Data Warehousing and Knowledge Discovery1999, Springer, p. 271-280.

Mannila, H., and Ronkainen, P., Similarity of event sequences, *in* Proceedings Temporal Representation and Reasoning, 1997.(TIME'97), Proceedings., Fourth International Workshop on1997, IEEE, p. 136-139.

Mannila, H., and Salmenkivi, M., Finding simple intensity descriptions from event sequence data, *in* Proceedings Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining2001, ACM, p. 341-346.

Maurya, M. R., Rengaswamy, R., and Venkatasubramanian, V., 2007, Fault diagnosis using dynamic trend analysis: A review and recent developments: Engineering Applications of artificial intelligence, v. 20, no. 2, p. 133-146.

Nittel, S., 2009, A survey of geosensor networks: Advances in dynamic environmental monitoring: Sensors, v. 9, no. 7, p. 5664-5678.

Obweger, H., Suntinger, M., Schiefer, J., and Raidl, G., Similarity searching in sequences of complex events, *in* Proceedings Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on2010, IEEE, p. 631-640.

Oliveira, L. M., and Rodrigues, J. J., 2011, Wireless Sensor Networks: A Survey on Environmental Monitoring: JCM, v. 6, no. 2, p. 143-151.

Othman, M. F., and Shazali, K., 2012, Wireless sensor network applications: A study in environment monitoring system: Procedia Engineering, v. 41, p. 1204-1210.

Padhy, P., Martinez, K., Riddoch, A., Ong, H., and Hart, J. K., 2005, Glacial environment monitoring using sensor networks.

Prinzie, A., and Van den Poel, D., 2011, Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: an Acquisition Pattern Analysis application: Journal of Intelligent Information Systems, v. 36, no. 3, p. 283-304.

Roux, A. V. D., and Mair, C., 2010, Neighborhoods and health: Annals of the New York Academy of Sciences, v. 1186, no. 1, p. 125-145.

Rude, A., and Beard, K., High-Level Event Detection in Spatially Distributed Time Series, Berlin, Heidelberg, 2012, Springer Berlin Heidelberg, p. 160-172.

Sampson, R. J., 2003, The neighborhood context of well-being: Perspectives in biology and medicine, v. 46, no. 3, p. S53-S64.

Tao, C., Wongsuphasawat, K., Clark, K., Plaisant, C., Shneiderman, B., and Chute, C. G., Towards event sequence representation, reasoning and visualization for EHR data, *in* Proceedings Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium2012, ACM, p. 801-806.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A. S., Kurose, S., and Zandee, R., 2003, Spatial context and the complexity of daily travel patterns: an international comparison: Journal of Transport Geography, v. 11, no. 1, p. 37-46.

Vanderbilt, K. L., Lin, C.-C., Lu, S.-S., Kassim, A. R., He, H., Guo, X., San Gil, I., Blankman, D., and Porter, J. H., 2015, Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network: Ecosphere, v. 6, no. 10, p. 1-18.

Vrotsou, K., and Forsell, C., A qualitative study of similarity measures in event-based data, *in* Proceedings Symposium on Human Interface2011, Springer, p. 170-179.

Wang, W., Hu, C., Chen, N., Xiao, C., and Jia, S., 2016, Spatio-Temporal Risk Assessment Process Modeling for Urban Hazard Events in Sensor Web Environment: ISPRS International Journal of Geo-Information, v. 5, no. 11, p. 203.

Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J., and Welsh, M., 2006, Deploying a wireless sensor network on an active volcano: IEEE internet computing, v. 10, no. 2, p. 18-25.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M., and Shneiderman, B., 2012, Querying event sequences by exact match or similarity search: Design and empirical evaluation: Interacting with computers, v. 24, no. 2, p. 55-68.

Yang, D.-H., Goerge, R., and Mullner, R., 2006, Comparing GIS-based methods of measuring spatial accessibility to health services: Journal of medical systems, v. 30, no. 1, p. 23-32.

Yang, J., McAuley, J., Leskovec, J., LePendu, P., and Shah, N., Finding progression stages in time-evolving event sequences, *in* Proceedings Proceedings of the 23rd international conference on World wide web2014, ACM, p. 783-794.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews state of the art in areas starting with basic concepts related to events, event sequences, event patterns, and event processing or event-based analytics. It introduces concepts of events, settings, contexts, and their sequences. It further reviews related work on similarity measures between sequences and their applications.

2.1. Event, Setting and Context

The following event concepts and terms are synthesized from the literature and presented here to clarify terms used throughout the remainder of the thesis.

Concepts of event: There exist many definitions for events in different domains and events have been defined and described in numerous research papers. By the definition in the Merriam-Webster dictionary, an event is an occurrence, "something that happens or a noteworthy happening". Within the area of computer science, events and their properties have been defined differently by different researchers. Some researchers (Gehani et al., 1992) consider events as happening "instantaneously at specific points in time." An event can be defined as anything that happened or is contemplated as happening with a significant change of state (Luckham and Schulte, 2011) or some detectable condition (Gunderson 2012). Luckham et al. have also defined an event as an object that is a record of an activity or an occurrence of significance in a system (Luckham, 2016; Luckham, 2001; Luckham and Frasca, 1998). Chandy & Schulte (2010) define events as any occurrences over space and time. For example, disease outbreaks in a region, earthquakes and value changes in sensors at monitoring stations can be considered as events. In addition, events can be

defined as either real or abstract, i.e. not occurring in a physical world and only modeled, imagined or simulated as virtual events (Luckham and Schulte, 2011). In an event-based system, (Obweger 2009) defines an event as any notable state change. Any event by its nature has space and time dimensions either explicitly or implicitly. In this thesis we take the definition of an event as a significant change of status specified by the users or experts in a specific domain.

Events are assumed to have attributes or properties which can include an assigned event type or a more detailed set of information that describes the state change. An event type is a class of event objects (Luckham and Schulte, 2011). (Luckham and Schulte, 2011; Luckham, 2001). Event type classifies the structure and properties of an event. It emphasizes the essential factors that uniquely identify the occurrence of an event of that type (Paschke and Boley, 2009). An event instance (also known as **event object**, **event individual**, and **raw event**) is a concrete instantiation of an event type. Event instances can have attributes of a timestamp or duration, spatial location, as well as attributes specific to the event type.

Spatial-Temporal Events: Spatial-temporal events are events that occur in a particular location and time. These events are described using spatial and temporal attributes, such as latitude, longitude, altitude, time, and duration. Spatial-temporal events can be used to represent various phenomena such as weather patterns, disease outbreaks, and transportation routes.

Primitive event: A primitive event (also known as simple event or atomic event) is defined as an instantaneous and significant occurrence of a thing, emphasizing the indivisible property as a lowest level of event. A primitive or simple event is an event that

is not viewed as summarizing, representing, or denoting a set of other events (Luckham and Schulte, 2011).

Complex event and composite event: A complex event can be defined as an event that summarizes, represents, or denotes a set of other events (Luckham and Schulte, 2011). A complex event is composed or derived from atomic or other complex events, and the included events are called components (Paschke and Boley, 2009). A composite event is derived from a complex event that is created by combining base events using a specific set of event algebra or constructors like disjunctions, conjunctions, sequences. So, a composite event is also known as a compound event.

Event sequence: There are many definitions of event sequences or "sequences of events". Event sequences can refer to either sequences of event instances or sequences of event types depending on the specific context. In this thesis we borrow an event sequence definition from (Moen, 2000), as an ordered collection of events from a finite set of event types, with each event of the sequence having an occurrence time. An event sequence can thus be denoted as (e_i, t_i), where $i = 1, \ldots, n$, and for each i, $e_i \in E$ is an event type and t_i is the occurrence time. Event sequences can be analyzed using various techniques such as pattern mining and clustering to identify patterns and trends. In this thesis we consider sequences of event instances of the same type unless otherwise specified.

Event Settings: An event setting is a type of geographic context that describes the physical and social environment in which an event takes place (Worboys, 2005; Worboys and Hornsby, 2004). Event settings can be described using various attributes such as location, time, and social environment. Event settings can influence the perception and

interpretation of events and can be used to provide additional context and meaning to events. According to there are three key components that define an event setting:

Spatial extent: This refers to the physical boundaries of the setting, which can range from very small (e.g., a single room) to very large (e.g., a city or region).

Temporal extent: This refers to the duration of the event setting, which can range from a few seconds (e.g., a car accident) to many years (e.g., the lifespan of a city).

Social context: This refers to the social and cultural factors that shape the event setting, including the people, institutions, and practices that are present in the setting.

Worboys and Hornsby argue that understanding event settings is important for a range of applications in GIScience, including emergency response, urban planning, and cultural heritage management. By analyzing the spatial and temporal characteristics of event settings, researchers can gain insights into how people interact with their environment and how events are shaped by social and cultural factors.

Contextual Factors or Contexts: Contextual factors refer to the various factors that influence the perception and interpretation of events. Contextual factors can include social, cultural, and historical factors that shape the way events are perceived and understood. Contextual factors can be used to provide additional context and meaning to events and can help to explain why events occur.

2.2. Event detection and pattern discovery

Primitive or simple event extraction or detection can be grouped into three categories according to different event abstraction functions: 1) threshold-based approaches (Abadi et al., 2016) in which an event is regarded to occur when sensor readings exceed some

predefined thresholds, 2) pattern-based approaches (Hogenboom et al., 2016) in which an event is represented as spatiotemporal patterns and event detection is performed using pattern matching techniques; and 3) learning-based approaches (Wang et al., 2016a) in which selected modeling methods are used to model spatiotemporal dependencies of sensor data and make probabilistic inference about events.

Event pattern discovery is the process of identifying novel relationships between events. These relationships can be temporal, spatial, and causal, etc. Event pattern discovery has been used to identify sequences of event types in a specific domain (Hasan et al., 2015).

2.3. Similarity between event sequences

Similarity is an important concept in many research areas including biology, computer science, linguistics, logic, mathematics, philosophy and statistics (Moen, 2000). Similarity or distance measures between data objects form a basic building block for several computational tasks such as clustering, classification, and anomaly detection. Across difference fields, similarity or distance measures have a wide variety of definitions. A similarity metric measures how alike two data objects or sequences are, based on features of the objects. Different similarity measures can reflect different facets of the data and no single similarity measure can capture all dimensions of the data or serve all purposes. Similarity is typically measured in the range 0 to 1 [0, 1] with 1 indicating complete similarity and 0 indicating complete dissimilarity. Similarity measures are often defined in measures of distance where a high degree of similarity means a short separation distance,

and vice versa. These distances measures are expected to satisfy a set of mathematical conditions:

Given a set of objects O for all a, b, c belonging to O

Not negative: $d(a, b) \geq 0$ and $d(a, b) = 0$ if and only if $a = b$

Symmetric: $d(a, b) = d(b, a)$

Triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$

The main interest of this thesis is similarity metrics for event sequences. Substantial research exists on sequence similarity measures, but these measures depend very much on the type of sequence. Measures have been defined for numeric sequences and time series (Berndt and Clifford, 1994; Goldin and Kanellakis, 1995; Guralnik and Srivastava, 1999), text sequences (Levenshtein, 1966) and biological sequences (Smith and Waterman, 1981) but event sequences are notably different.

Event sequences are a common form of data that can contain important knowledge to be discovered. Using event sequences is practical in many applications or scenarios, particularly those that aim to detect patterns in the occurrence of events or to retrieve past event sequences that are similar to a current one (Lupiani et al., 2013). A similarity measure is important for a variety of reasoning tasks used by many applications and data mining algorithms. Measuring similarity or distance between event sequences is a basic task for knowledge discovery from these temporally ordered or sequential event-based data in many domains. In the following sections, we review three main similarity measures for sequences that have been adapted to or apply to event sequences.

2.3.1. Jaccard similarity

To find the similarity measure between objects is to compare some metrics associated with these objects. The Jaccard similarity measure, originally proposed by Paul Jaccard (Jaccard, 1912; Leskovec et al., 2014; Niwattanakul et al., 2013) is one of the oldest similarity measures but one which still gets wide usage. This measure concerns the similarity between sets by looking at the relative size of their intersection. A set here refers to the mathematic term, which has no order and only a collection of elements, such as $\{a, b, c, d, e\}$. The sets $\{a, b, c, d, e\}$ and $\{\{e, b, d, c, a\}$ are equivalent. The Jaccard similarity between two sets $S1$ and $S2$ is the ratio of the size of their intersection to the size of their union, which is given below:

$$sim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|}$$

Where, two vertical bars for a set represent the cardinality of the set, such as |S1|, |S2| which calculates the number of elements in S1 and S2. The intersection between two sets S1 and S2 denoted by S1∩S2 reveals all common items in both sets S1 and S2. The union between two sets S1 and S2 denoted by S1∪S2 means all elements in either set. So, the Jaccard similarity is the similarity between finite sample sets and can be defined as the cardinality of the intersection of two sets divided by the cardinality of their union. Calculation of the Jaccard similarity for the following sets is shown below.

Two sets: $S1 = \{a, b, c, d, e, f, g, h\}$

$S2 = \{b, d, i, j\}$

Jaccard similarity:

$$sim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{2}{10} = 0.2$$

For this thesis we adapt the Jaccard measure to incorporate temporal alignment.

2.3.2. Edit distance

Edit distance was first developed for comparing similarity or distance between strings. It refers to the total number of editing operations needed to transform from one string into another string. The lower the number, the more similar the strings are. Some examples of developing and using edit distance can be seen in the early works such as Hamming distance (Hamming, 1950), Levenshtein distance (Levenshtein, 1966), Jaro-Winkler distance (Jacobs and Walczak, 1983), and Longest Common Subsequence (LCSS) distance (André-Jönsson and Badal, 1997). The edit distance measure was first extended to a similarity measure between event sequences with the lowest cost of three types of editing operations: insert, delete and move (Mannila and Moen, 1999; Mannila and Ronkainen, 1997). The move operation was included to incorporate the occurrence time of the events. As noted by Wongsuphasawat et al. (2012) this approach allows only monotonic mapping, which means that the matched events in the target and candidate sequences must be in similar order.

2.3.3. Explicit Event Sequence Similarity Measures

Several measures for event sequence similarity have been presented in the literature. Most of the event sequence similarity research has focused on querying for similar event sequences. In this context, the assumption is that collected event sequences are stored in a database, a query or target event sequence is supplied, and the task is to find all similar event sequences to the target sequence within the database.

Match and Mismatch (M&M) measure is an explicit event sequence similarity measure. The first version of this measure called M&M measure v.1 was initially developed by Wongsuphasawat and Shneiderman (2009), to calculate similarity scores

between a query event sequence and stored event sequences in the database based on the number and time difference of matched and mismatched events. To overcome some limitations in M&M measure v.1, an improved version called M&M measure v.2 was developed (Wongsuphasawat et al., 2012). This second version improves the matching algorithm with dynamic programming instead of the Hungarian Algorithm. In addition, it considers more types of differences in event similarity measures including the number of missing events, the number of extra events, and the number of swaps. Also following the idea of M&M measure, Vrotsou (2010) and Vrotsou and Forsell (2011) proposed and discussed nine measures to cover several aspects regarding similarity of event sequences, which confirms the theoretical applicability of these measures and provides a solid basis for further evaluation in practical applicability. With the context of complex events, Obweger et al. (2010) proposed approaches for both single event and event sequence based similarity measures. Single-event similarity is obtained based on event attribute values through the relative positioning of two events in an n-dimensional space. Therefore, the similarity between two individual events is calculated from weighted attribute-level similarities. The proposed event-sequence similarity is computed considering individual event-level similarities, order, and relative and absolute temporal structures.

2.4. Spatial context

In this research, we want to associate the spatial context with occurrences of events and event sequences. Context can be defined in many ways, most often with location as the most important emphasis, namely spatial context. Context is defined as "location and the identity of nearby people and objects." (Jiang and Yao, 2007). A broader definition is “context is any information that can be used to characterize the situation of an entity, where

entity means a person, place, or object, which is relevant to the interaction between a user and an application, including the user and the applications themselves." (Dey, 2001).

Spatial context is an important factor in many domains and applications. For instance, spatial context strongly influences the transport disadvantage that in turn affects social exclusion and well-being (Delbosc and Currie, 2011). In a travel behavior research, spatial context is strongly related to the household travel patterns in an international scale (Timmermans et al., 2003). A person's health-related problems are strongly affected by different types of spatial context, such as environmental exposures (Cutter, 1996; Roux and Mair, 2010), social environment (characteristics of communities and neighborhoods) (Roux and Mair, 2010; Sampson, 2003), and ease of access to health services (Yang et al., 2006). Spatial context greatly influences the potential of getting a disease, the adoption of healthy lifestyle, and the ease of access to medical services for disease diagnosis and treatment. An early psychological behavior research study indicates that decision behavior is affected by spatial context or spatially varied factors (Wolpert, 1964). A farming population was selected to study the effect of spatial context in decision processes because the outcomes of decision behavior are easily observable over the landscape. The decision making in farming is dispersed spatially among many farmers due to the uneven diffusion of market and technical information. With the strong emphasis and integration of spatial context, a new area of ecological studies called spatial ecology emerged (Gripenberg and Roslin, 2007; Tilman and Kareiva, 2018).

Spatial context is also very important in recognition of objects in images. In a content-based image retrieval experiment, incorporating the spatial context models dramatically reduced the misclassification and increased the accuracy of material detection by 13%

(Singhal et al., 2003). In order to better recognize or identify defined objects (e.g. cars, rivers, sky) in an image, combining the naturally classified texture or colors as spatial context greatly improved detection accuracy (Heitz and Koller, 2008).

Context, especially spatial context in this research, plays an important role when measuring the similarity of two entities or event sequences. Very little research effort has been focused on this area. In one study, the authors explored the effect of context on existing similarity measurement approaches in the geospatial domain (Keßler, 2007; Keßler et al., 2007). They defined context for similarity measurement as " *A similarity measurement's context is any information that helps to specify the similarity of two entities more precisely concerning the current situation. This information must be represented in the same way as the knowledge base under consideration, and it must be capturable at maintainable cost*." In combining a generic set of characteristics of context into similarity measurement, they also pointed out that developers in specific domains should focus on parameters that influence an application-specific context model considering the impact, representation and capture. Therefore, problems still remain for practical formalization and applications in specific domains or systems, for which we propose more practical methods in this research and assessment with a case study.

2.5. Discovery of spatiotemporal event patterns

In this section, we present various methods from different perspectives to find novel and relatively general relationships or associations among many different spatiotemporal event types where the relationships or associations among these event types are not known in advance or not explicitly represented in the data. We especially focus on the event sequence-based knowledge discovery.

2.5.1. Spatial and spatiotemporal analysis approach

Spatial analysis combining conventional geostatistics and modeling can be used to discover knowledge of spatiotemporal patterns of various phenomena or complex events occurring in nature or human society. Extracting interesting and useful patterns from spatial or spatiotemporal datasets is more difficult than discovering patterns from conventional numeric and categorical data due to the complexity of spatial data types, spatial relationships, spatial autocorrelation, and nonlinearity (Shekhar et al., 2011), which spatial analysis approaches must consider. The Event-based SpatioTemporal Data Model (ESTDM) was first proposed to explicitly represent change over space relative to time (Peuquet and Duan, 1995). ESTDM has set up a foundation for facilitating procedures for answering queries relating to temporal relationships, as well as analytical tasks for comparing different event sequences.

As one important spatial analysis approach, spatial scan statistics have been widely applied to event-based data analysis in many domains. The space-time permutation model (STPM) in scan statistics, a spatiotemporal clustering method, was successfully applied to analyze historical fire event sequence datasets from 1969 to 2008 for hotspots detection with different spatial context in Canton Ticino, Switzerland (Orozco et al., 2012). It was also used to detect active fire events in the state of Florida (US) identified by MODIS (Moderate Resolution Imaging Spectroradiometer) during the period 2003–06 (Tonini et al., 2009). STPM was effectively applied to early detection of events of disease outbreaks with only case numbers (Kulldorff et al., 2005b). It was further evaluated using daily analyses of hospital emergency department visits in New York City and identified four of

the five strongest potential event signals associated to citywide outbreaks due to rotavirus, norovirus, and influenza.

We also apply this STPM model in spatial scan statistics to our case study and compare the results with this method to those based on the approaches developed in this thesis.

2.5.2. Machine learning centered data mining

Machine learning centered data mining is a powerful approach for mining spatiotemporal patterns. Existing approaches for pattern mining use state-of-the-art approaches from machine learning to extract complex events and detect patterns in the form of association rules or sequential rules. More options from machine learning approaches provide appropriate selection for solving a specific problem. The combination of Support Vector Machine (SVM) with Conditional Random Field (CRF) has been successfully used to identify spatiotemporal activity patterns of animal movements (Behmann et al., 2016).

Different algorithms have been developed and integrated in machine learning systems to solve some specific event prediction problems. For instance, *timeweaver*, a generic algorithm based machine learning was developed to predict rare events from historical sequences of events by identifying predictive temporal and sequential patterns (Weiss and Hirsh, 1998). An episode can be defined as a subset of events that occur within time intervals of a given size in a given partial order, based on which we can produce rules or event patterns for describing or predicting the behavior of entire event sequence or a system. An efficient algorithm was developed for the discovery of all frequent episodes

from a given class of episodes, and was evaluated with promising experimental results (Mannila et al., 1995). It is difficult to differentiate anomalous event sequences from normal network traffic. An application was built to enhance domain knowledge with machine learning techniques to create rules for intrusion detection (Sinclair et al., 1999). In this application, genetic algorithms and decision trees were used to automatically generate rules for classifying network connections.

Automatic integration of a set of rules or event sequence patterns into Complex Event Processing (CEP) system is a recent trend. Contributing to this work, a Sequence Clustering-based Automated Rule Generation (SCARG) was proposed to automatically generate rules by mining decision-making history of domain experts based on event sequence clustering and probabilistic graphical modeling (Lee and Jung, 2017). This model-based system can make self-adaptive CEP system possible by combining the rule generation method and the existing dynamic CEP systems, which is verified by a case study of an automated stock trading system. As a branch of machine learning, a deep learning method of a neural network has been used to predict events of interest. For example, in order to predict purchasing intent in an ecommerce setting, with the input event data comprising categories, quantities and unique instances, multi-layer recurrent neural networks were established to automatically capture both session-local and dataset-global event dependencies and relationships for user sessions of any length (Sheil et al., 2018).

Integrating the similarity measures of event sequences and spatial context into machine learning systems is expected to improve the capacity and efficiency of event sequence-based data analysis. No or little work has been found from current publications. We incorporate the similarity measures developed in this research into some selected

machine learning algorithms to test the availability and verify the effectiveness with a case study of coastal fecal pollution.

Chapter References

Abadi, D., Madden, S. and Lindner, W. (2016) Data Stream Management, pp. 409-428, Springer.

André-Jönsson, H. and Badal, D.Z. 1997 Using signature files for querying time-series data, pp. 211-220, Springer.

Behmann, J., Hendriksen, K., Mцller, U., Bцscher, W. and Plцmer, L. 2016. Support Vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors. Geoinformatica 20(4), 693-714.

Berndt, D.J. and Clifford, J. 1994 Using dynamic time warping to find patterns in time series, pp. 359-370, Seattle, WA.

Cutter, S.L. 1996. Vulnerability to environmental hazards. Progress in human geography 20(4), 529-539.

Delbosc, A. and Currie, G. 2011. The spatial context of transport disadvantage, social exclusion and well-being. Journal of Transport Geography 19(6), 1130-1137.

Dey, A.K. 2001. Understanding and using context. Personal and ubiquitous computing 5(1), 4-7.

Gehani, N.H., Jagadish, H.V. and Shmueli, O. 1992. Event specification in an active object-oriented database. ACM sigmod record 21(2), 81-90.

Goldin, D.Q. and Kanellakis, P.C. 1995 On similarity queries for time-series data: constraint specification and implementation, pp. 137-153, Springer.

Gripenberg, S. and Roslin, T. 2007. Up or down in space? Uniting the bottom-up versus top-down paradigm and spatial ecology. Oikos 116(2), 181-188.

Guralnik, V. and Srivastava, J. 1999 Event detection from time series data, pp. 33-42, ACM.

Hamming, R.W. 1950. Error detecting and error correcting codes. Bell System technical journal 29(2), 147-160.

Hasan, A., Teymourian, K. and Paschke, A. 2015 Probabilistic event pattern discovery, pp. 241-257, Springer.

Heitz, G. and Koller, D. 2008 Learning spatial context: Using stuff to find things, pp. 30-43, Springer.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. and Caron, E. 2016. A survey of event extraction methods from text for decision support systems. Decision Support Systems 85, 12-22.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. New phytologist 11(2), 37-50.

Jacobs, B.E. and Walczak, C.A. 1983. A generalized query-by-example data manipulation language based on database logic. IEEE Transactions on Software Engineering (1), 40-57.

Jiang, B. and Yao, X. (2007) Location based services and telecartography, pp. 27-45, Springer.

Keßler, C. 2007 Similarity measurement in context, pp. 277-290, Springer.

Keßler, C., Raubal, M. and Janowicz, K. 2007 The effect of context on semantic similarity measurement, pp. 1274-1284, Springer.

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

Lee, O.-J. and Jung, J.E. 2017. Sequence clustering-based automated rule generation for adaptive complex event processing. Future Generation Computer Systems 66, 100-109.

Leskovec, J., Rajaraman, A. and Ullman, J.D. (2014) Mining of massive datasets, Cambridge university press.

Levenshtein, V.I. 1966 Binary codes capable of correcting deletions, insertions, and reversals, pp. 707-710.

Luckham, D. 2016. Event Processing Glossary-Version 2.0, Event Processing Technical Society. http://www. ep-ts. com/component/option, com_docman/task, doc_download/gid, 66/Itemid, 84/.

Luckham, D. and Schulte, R. 2011. EPTS Event Processing Glossary v2. 0. Event Processing Technical Society.

Luckham, D.C. (2001) The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley Longman Publishing Co., Inc.

Luckham, D.C. and Frasca, B. 1998. Complex event processing in distributed systems. Computer Systems Laboratory Technical Report CSL-TR-98-754. Stanford University, Stanford 28.

Lupiani, E., Sauer, C., Roth-Berghofer, T., Juarez, J.M. and Palma, J. 2013 Implementation of similarity measures for event sequences in myCBR.

Mannila, H. and Moen, P. 1999 Similarity between event types in sequences, pp. 271-280, Springer.

Mannila, H. and Ronkainen, P. 1997 Similarity of event sequences, pp. 136-139, IEEE.

Mannila, H., Toivonen, H. and Verkamo, A.I. 1995 Discovering frequent episodes in sequences extended abstract.

Moen, P. 2000. Attribute, event sequence, and event type similarity notions for data mining. PhD thesis, University of Helsinki.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S. 2013 Using of Jaccard coefficient for keywords similarity.

Obweger, H., Suntinger, M., Schiefer, J. and Raidl, G. 2010 Similarity searching in sequences of complex events, pp. 631-640, IEEE.

Orozco, C.V., Tonini, M., Conedera, M. and Kanveski, M. 2012. Cluster recognition in spatial-temporal sequences: the case of forest fires. Geoinformatica 16(4), 653-673.

Paschke, A. and Boley, H. (2009) Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, pp. 215-252, IGI Global.

Peuquet, D.J. and Duan, N. 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International journal of geographical information systems 9(1), 7-24.

Roux, A.V.D. and Mair, C. 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186(1), 125-145.

Sampson, R.J. 2003. The neighborhood context of well-being. Perspectives in biology and medicine 46(3), S53-S64.

Sheil, H., Rana, O. and Reilly, R. 2018. Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks. arXiv preprint arXiv:1807.08207.

Shekhar, S., Evans, M.R., Kang, J.M. and Mohan, P. 2011. Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3), 193-214.

Sinclair, C., Pierce, L. and Matzner, S. 1999 An application of machine learning to network intrusion detection, pp. 371-377, IEEE.

Singhal, A., Luo, J. and Zhu, W. 2003 Probabilistic spatial context models for scene content understanding, pp. I-I, IEEE.

Smith, T.F. and Waterman, M.S. 1981. Comparison of biosequences. Advances in applied mathematics 2(4), 482-489.

Tilman, D. and Kareiva, P. (2018) Spatial ecology: the role of space in population dynamics and interspecific interactions (MPB-30), Princeton University Press.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A.S., Kurose, S. and Zandee, R. 2003. Spatial context and the complexity of daily travel patterns: an international comparison. Journal of Transport Geography 11(1), 37-46.

Tonini, M., Tuia, D. and Ratle, F. 2009. Detection of clusters using space–time scan statistics. International journal of wildland fire 18(7), 830-836.

Vrotsou, K. (2010) Everyday mining: Exploring sequences in event-based data, Linköping University Electronic Press.

Vrotsou, K. and Forsell, C. 2011 A qualitative study of similarity measures in event-based data, pp. 170-179, Springer.

Wang, T.-Y., Yang, M.-H. and Wu, J.-Y. 2016. Distributed Detection of Dynamic Event Regions in Sensor Networks With a Gibbs Field Distribution and Gaussian Corrupted Measurements. IEEE Transactions on Communications 64(9), 3932-3945.

Weiss, G.M. and Hirsh, H. 1998 Learning to Predict Rare Events in Event Sequences, pp. 359-363.

Wolpert, J. 1964. The decision process in spatial context. Annals of the Association of American Geographers 54(4), 537-558.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. and Shneiderman, B. 2012. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with computers 24(2), 55-68.

Wongsuphasawat, K. and Shneiderman, B. 2009 Finding comparable temporal categorical records: A similarity measure with an interactive visualization, pp. 27-34, IEEE.

Worboys, M. 2005. Event-oriented approaches to geographic phenomena. International Journal of Geographical Information Science 19(1), 1-28.

Worboys, M. and Hornsby, K. 2004 From objects to events: GEM, the geospatial event model, pp. 327-343, Springer.

Yang, D.-H., Goerge, R. and Mullner, R. 2006. Comparing GIS-based methods of measuring spatial accessibility to health services. Journal of medical systems 30(1), 23-32.

CHAPTER 3

A UNIFYING FRAMEWORK FOR ANALYSIS OF SPATIAL-TEMPORAL EVENT SEQUENCE SIMILARITY AND ITS APPLICATIONS

The draft in this chapter is the reformatted version from the published research paper:

Xu, F.; Beard, K. A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. ISPRS Int. J. Geo-Inf. 2021, 10, 594. https://doi.org/10.3390/ijgi10090594

Chapter Abstract

Measures of similarity or differences between data objects are applied frequently in geography, biology, computer science, linguistics, logic, business analytics, and statistics, among other fields. This work focuses on event sequence similarity among event sequences extracted from time series observed at spatially deployed monitoring locations with the aim of enhancing the under-standing of process similarity over time and geospatial locations. We present a framework for a novel matrix-based spatiotemporal event sequence representation that unifies punctual and interval-based representation of events. This unified representation of spatiotemporal event sequences (STES) supports different event data types and provides support for data mining and sequence classification and clustering. The similarity measure is based on the Jaccard index with temporal order constraints and accommodates different event data types. The approach is demonstrated through simulated data examples and the performance of the similarity measures is evaluated with a k-nearest neighbor algorithm (k-NN) classification test on synthetic datasets. As a case study, we demonstrate the use of these similarity measures in a spatiotemporal analysis of event sequences extracted from space time series of a water quality monitoring system.

Keywords: spatiotemporal event sequences (STES); matrix representation; similarity measures; time locked Jaccard similarity; K-NN/1-NN

3.1. Introduction

Wireless sensor networks (WSN) or other monitoring systems, deployed regularly or irregularly in geographic space, have become commonly used for environmental data collection and monitoring. Each monitoring station or node can have one or more sensors producing time series on variables of interest for monitoring. Within this setting, we may be interested in the similarity among the time series observed across a set of monitoring stations. For example, we might want to ask, how similar are water quality monitoring variables within an estuary or across different estuaries? Several prior studies have researched time series similarity measures but time series can contain substantial data redundancy making similarity computations inefficient and expensive (Bollobas et al., 1997; Fu, 2011). Converting time series to event sequences can reduce the data volume while retaining key information (Du et al., 2016; Shurkhovetskyy et al., 2018; Yeh et al., 2018). In this paper we report on development of an approach for measuring the similarity among event sequences associated with monitoring stations distributed within some geographic space. We refer to these as spatiotemporal event sequences (STES) because of the pertinence of their distribution in space. The approach aims to address two basic questions. Firstly, how similar are event sequences within a defined geospatial region? Secondly, within the region, do event sequences that are closer in space tend to be more similar? Answers to these questions can contribute to insights on patterns in spatial processes that can be helpful for environmental monitoring.

Figure 3.1A illustrates an instance of an STES as a set of temporally ordered events observed at a fixed location in space. An STES differs from other types of event sequences such as genomic sequences (Darling et al., 2004), industrial process monitoring sequences (Maurya et al., 2007), patient symptom sequences (Tao et al., 2012), political event sequences (Stehle and Peuquet, 2015), or consumer purchasing sequences (Prinzie and Van den Poel, 2011) in that STES derive from time series observed at fixed geospatial locations and each sequence consists of events of the same type (e.g., high temperature events, heavy precipitation events, impaired water quality events, drought events).

Converting time series to event sequences leads to on-going production of STES at each monitoring station as illustrated in Figure 3.1B. An individual STES conceptually represents a realization of a process at the location and the set of STES deployed in a region conceptually forms a field of event sequences representing an evolving underlying process (Yang et al., 2014). As an example, a precipitation event sequence observed at station S1 (Figure 3.1A) represents a local realization of a meteorological process. Through similarity

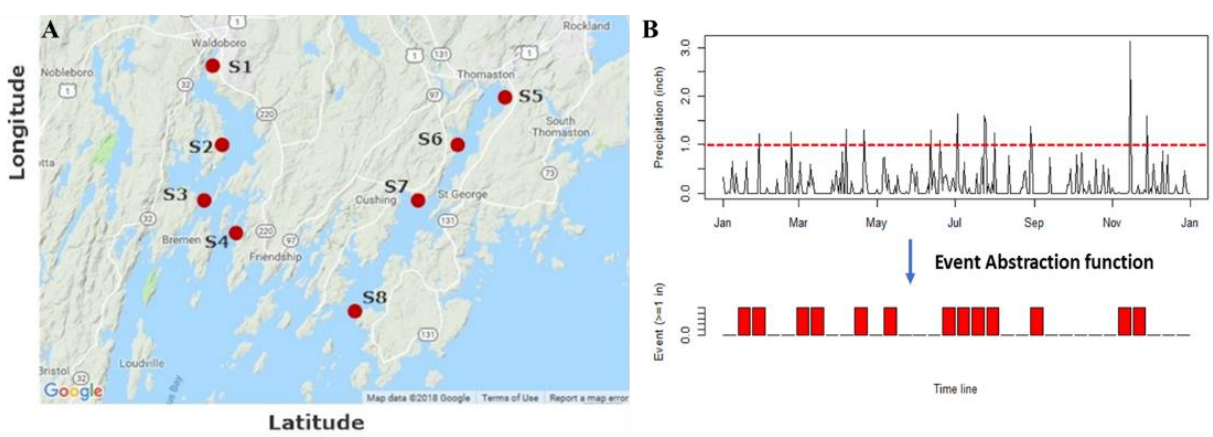


Figure 3.1. The STES problem setting. (**A**) An example of fixed locations of interest or observation sites distributed along the coast. (**B**) An example of a spatiotemporal event sequence extracted from a space-time series of precipitation with the threshold of ≥1.0 inch/24 h. Red bars represent events.

measures among event sequences in geographic space we can extend Tobler's First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things", to an assessment of process similarity in space.

Related work on several similarity measures can be found for event sequences, but not directly STES as we define them. Edit distance is a measure of similarity first developed for comparing strings (a type of sequence). It refers to the total number of editing operations needed to transform one string into another string. The lower the number, the more similar the strings. Some examples of edit distance include Hamming distance (Hamming, 1950), Levenshtein distance (Levenshtein, 1966), Jaro–Winkler distance (Jacobs and Walczak, 1983), and Longest Common Subsequence (LCSS) distance (André-Jönsson and Badal, 1997). The edit distance measure was first extended to measure event sequence similarity using the lowest cost of three types of editing operations: insert, delete and move (Mannila and Moen, 1999; Mannila and Ronkainen, 1997). The move operation was included to incorporate the occurrence time of the events. As noted by Wongsuphasawat et al. (2012) this approach allows only monotonic mapping, which means that the matched events in the target and candidate sequences must be in similar order. The Jaccard similarity coefficient is a classic measure of similarity between two sets that continues to be applied in several application domains, for example in comparing biological sequence data (Chung et al., 2019; Vorontsov et al., 2013) and in web usage mining (Luu et al., 2020). More recent event sequence similarity measures have been proposed to take into consideration temporal order and temporal duration in addition to assessing event type similarity (Obweger et al., 2010). While most similarity metrics treat events as points in time, Kotsifakos et al. (2013) and Mirbagheri and Hamilton (2020) propose approaches for interval based event sequence similarity (Andrienko et al., 2010; Mirbagheri and Hamilton, 2020). Their event representation includes an event label and start and end time, and the event sequence is a

list of these arranged in ascending order. Their concept of similarity between two event sequences includes the presence of event intervals with the same labels, the order of occurrences of the event intervals, the duration of the event intervals, and the temporal relations among the event intervals. To our knowledge, none of the currently available similarity measures for event sequences address both time stamped and interval based events and consider the spatial dimension. Our event sequence similarity approach builds on the Jaccard index and integrates interval and time stamped events.

The paper is organized as follows. Section 2, Materials and Methods, describes the process of eventization and generation of STES, the proposed methods for transforming STES to matrices based on various measurement characteristics, and the development of similarity measures for different levels of event representation (qualitative vs. quantitative), as applied to entire sequences or user defined moving windows. Section 3, Results and Discussion, demonstrates construction of STES similarity matrices and implementation of the similarity measures on synthetic mini datasets, further evaluates the performance of the similarity measures on execution speed and classification accuracy and provides a real-world application on classification of the Maine coastal regions based on cluster analysis of precipitation event sequences. Finally, Section 4 concludes this study, considering the remaining issues and future work.

3.2. Materials and Methods

3.2.1. Eventization and Spatiotemporal Event Sequences (STES)

Jassby and Powel (1990) describe an event as a short-term, yet substantial, discontinuity in the underlying behavior of a time series (Jassby and Powell, 1990). Eventization is the process of event identification from observations or measured raw data according to user definitions applied in a specific domain. In this paper it refers to the

process of event identification from space-time series and formation of timestamped, ordered event sequences. Briefly, primitive or simple event extraction (Rude and Beard, 2012) or detection can be grouped into three categories: (1) threshold-based approaches (Abadi et al., 2016) in which an event is regarded to occur when observations exceed some predefined thresholds, (2) pattern-based approaches (Hogenboom et al., 2016) in which an event is represented as a spatiotemporal pattern and event detection is performed using pattern matching techniques; and (3) learning-based approaches (Wang et al., 2016a) in which selected modeling methods are used to model spatiotemporal dependencies of sensor data and make probabilistic inference about events.

In environmental applications, we are interested in the spatiotemporal context of the sequences. The expressions of space and time components capture different granularities. Temporal entities have two types of time expression, timestamps and time intervals (Shahar, 1997). Timestamps can express different granularities as in what time, what date, what day of the week, what week, and what year, etc. Time intervals can also be of different granularities, such as seconds, minutes, hours, days, months, seasons, and years. Given these two temporal concepts, we identify two general types of STES: timestamped and interval events as illustrated in Figure 3.2.

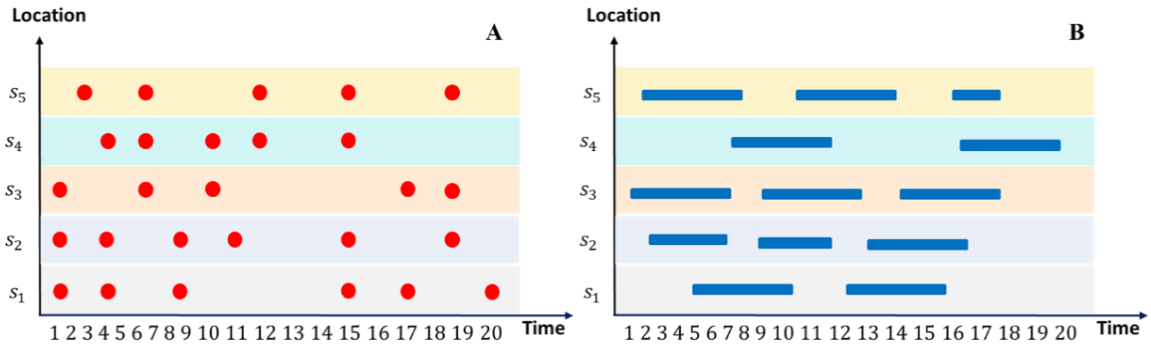


Figure 3.2. Graphical illustration of spatiotemporal event sequences (STES). (**A**) An example of spatiotemporal timestamped event sequences where rows represent locations each with 20-time units. (**B**) An example of interval event sequences for 5 locations and 20-time units.

For eventization, we need to consider the level of measurement of an observed time series variable. A real valued level of measurement may for example be retained in an event representation (as illustrated in Figure 3.3A). Alternatively, an observed real value at a time stamp may be transformed to an ordinal or binary value (as illustrated in Figure 3.3D). Interval events can be divided into as many timestamps as determined by an event definition and user defined granularity, within which the full range of observed values satisfying the event definition may be retained (see Figure 3.3B, C). Alternatively, all observed values within an interval that satisfy an event definition may be transformed to ordinal or binary values (as illustrated in Figure 3.3E, F).

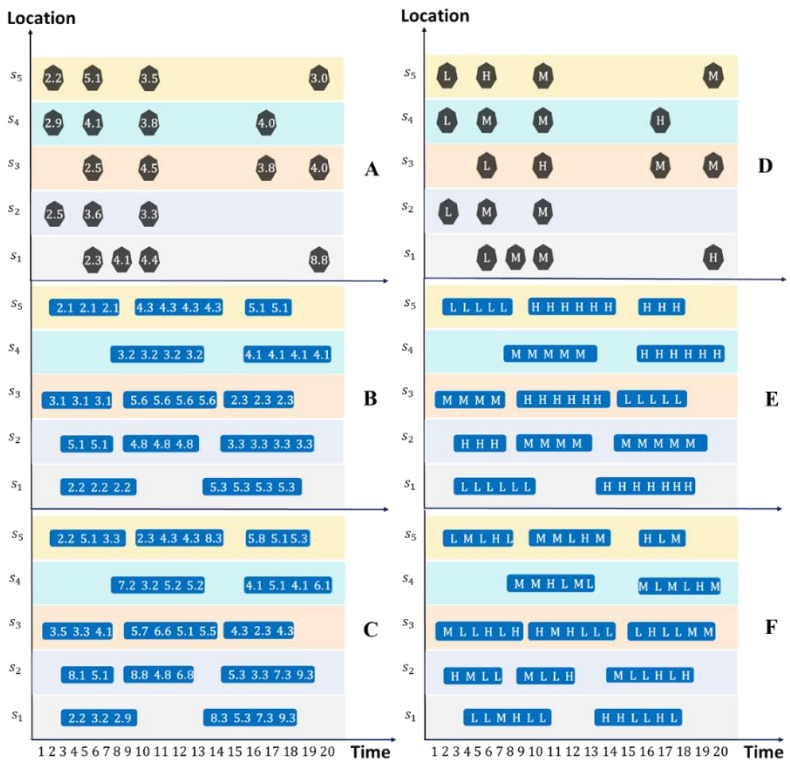


Figure 3.3. Graphical illustration of spatiotemporal event sequences (STES) with consideration for level of measurement and variation within a single event. STES in (**A**–**C**) are extracted from space-time series with interval/ratio values, and (**D**–**F**) are extracted as ordinal values from space-time series. (**A**, **D**) are punctual event sequences. (**B**, **E**) are interval event sequences with no internal variation over the interval. (**C**, **F**) are interval event sequences with bounded variation within the interval consistent with an event definition. H, M and L represent high, medium and low, respectively.

3.2.2. Matrix Representation of STES

For a regularly sampled time series, the set of timestamps T forms a discrete set, with observations spaced at uniform time intervals. Given s locations and t timestamps, a space-time series dataset can be represented with a $s \times t$ matrix where locations correspond to rows and timestamps to columns and ν represents an observed variable.

$$\begin{array}{cc} & G_0\text{-Timestamps } (1, 2, 3, \ldots, t) \\ \text{Spatial locations } (1, 2, 3, \ldots, s) & \begin{pmatrix} v_{11}, v_{12}, v_{13}, \ldots, v_{1t} \\ v_{21}, v_{22}, v_{23}, \ldots, v_{2t} \\ v_{31}, v_{32}, v_{33}, \ldots, v_{3t} \\ \ldots \\ v_{s1}, v_{s2}, v_{s3}, \ldots, v_{st} \end{pmatrix} \end{array} \quad (1)$$

G_0-Timestamps represent the finest temporal granularity as described by Shahar (1997), here corresponding to the time series sampling rate. Each value potentially corresponds to a status change, which could define a timestamped event or the start or end of an interval event. As noted above, events are identified based on different user defined functions such as threshold based, pattern-based, or learning based (Yin et al., 2009). For simplicity, in the following definitions, we assume use of a threshold, but the approach is generalizable to other event detection approaches (Guralnik and Srivastava, 1999). A temporal granularity in integer unit G_i scaled from G_0 (e.g., hour to day, day to week) is specified by a user based on application domain considerations. At each observation location s, an event sequence is formed at the Gi scale after eventization. The event sequences for all locations form an initial STES matrix. In the eventization process, the dimension can be further reduced through removing rows and columns without events in

locations across all G_i-timestamps or G_i-timestamps across all locations. Following this data reduction, we may have n locations and G_i granularity of m timestamps, in which the STES are represented as $n \times m$ matrix ($n \leq s$ and $m \leq t$).

$$\begin{array}{cc} & G_i \text{ Timestamps } (1, 2, 3, \ldots, m) \\ \text{Spatial locations } (1, 2, 3, \ldots, n) & \begin{pmatrix} e_{11}, e_{12}, e_{13,} \ldots, e_{1m} \\ e_{21}, e_{22}, e_{23,} \ldots, e_{2m} \\ e_{31}, e_{32}, e_{33,} \ldots, e_{3m} \\ \ldots \\ e_{n1}, e_{n2}, e_{n3,} \ldots, e_{nm} \end{pmatrix} \end{array} \quad (1)$$

We identify four different cases corresponding to timestamped versus interval events and qualitative versus quantitative. For the case of nominal values, appearance of a user specified nominal category or label at a timestamp indicates the occurrence of an event. For this case the event value is defined as follows:

$$e_{ij} = \begin{cases} na & if\ v_{ij} = missing\ data \\ 1 & if\ v_{ij} \geq threshold, or \\ & if\ v_{ij} = defined\ nominal\ value \\ 0 & otherwise \end{cases} \quad (2)$$

$$i = 1, 2, 3, \ldots, n; j = 1,2,3, \ldots, m$$

where all timestamped observations (v_{ij}) are ordinal, interval or ratio values, the corresponding event value e_{ij} may retain the original observation value or be subjected to some data transformation such as logarithm, percentage or normalization. Given a threshold for defining an event instance, sequences in this case can be represented as follows:

$$e_{ij} = \begin{cases} na \quad if\ v_{ij} = missing\ data \\ 0 \quad if\ v_{ij} < threshold \\ v_{ij}\ or\ v'_{ij} \quad if\ v_{ij} \geq threshold \\ \quad v'_{ij}\ is\ transformed\ v_{ij} \end{cases} \tag{3}$$

$$i = 1, 2, 3, \dots, n; j = 1,2,3, \dots, m$$

An interval event occurs when the defining event conditions persist for more than one G_1 timestamp. As long as we determine the smallest temporal granularity in a specific study or system, we can represent an interval event sequence through the same timestamped event matrix as described above. The case for interval events with categorical values can be defined according to Equation (5):

$$e_{ij,}e_{ij+1,} \dots, e_{ij+\Delta t} = \begin{cases} na \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} = missing\ data \\ 1 \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} \geq threshold \\ or, \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} = defined\ nomial\ scale \\ \quad \Delta t \geq 1 \\ 0 \quad otherwise \end{cases} \tag{4}$$

$$i = 1, 2, 3, \dots, n; j = 1,2,3, \dots, m$$

The case for interval events with ordinal or interval/ratio values can be defined according to Equation (6):

$$e_{ij,}e_{ij+1,} \dots, e_{ij+\Delta t} = \begin{cases} na \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} = missing\ data \\ 0 \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} < threshold \\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t}\ or\ v'_{ij,}v'_{ij+1,} \dots, v'_{ij+\Delta t} \\ \quad if\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} \geq threshold \\ \quad v'_{ij,}v'_{ij+1,} \dots, v'_{ij+\Delta t}: transformed\ v_{ij,}v_{ij+1,} \dots, v_{ij+\Delta t} \\ \quad \Delta t \geq 1 \\ 0 \quad otherwise \end{cases} \tag{5}$$

$$i = 1, 2, 3, \dots, n; j = 1,2,3, \dots, m$$

For all interval events with no internal variation within the interval, i.e., with a constant event class level, the defined interval events are as described in Equation (6):

$$e_{ij} = e_{ij+1} = \cdots = e_{ij+\Delta t}$$
$$from\ v_{ij} = v_{ij+1} = \cdots = v_{ij+\Delta t}\ or\ v'_{ij} = v'_{ij+1} = \cdots = v'_{ij+\Delta t}$$

3.2.3. Development of Similarity Measures for Spatiotemporal Event Sequences

The matrix framework presented above provides a flexible method to investigate sequence similarity over space for the same time windows. In this context, we consider the event sequence similarity as the level of co-occurring timestamped events for a certain time period for two or more locations. We can vary the selection of a time window based on the sampling frequency of the observation data and a target event granularity (e.g., drought events which may be defined as over 10 days of no rain need a larger time window relative to heavy precipitation events). We present similarity measures for five situations: (a) binary timestamped events (no consideration of variable class levels or magnitude), (b) timestamped events with variable class levels or magnitude, (c) interval events considering time overlaps only, (d) interval events with constant nominal or ordinal labels and time overlaps, and (e) interval events with a range of real values and time overlaps.

We follow the concept of Jaccard similarity (Jaccard, 1901) but consider the order of individual event elements within each event sequence. The intersection between two sets of spatiotemporal event sequences means the common events must "co-occur" in both sequences, and the union refers to all events in either sequence. The measure of co-occurrence is demonstrated by the following example:

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
S1	e	e	e	e	e	e	e			e
	|	|	|	|	|	|	|	|	|	|
S2	e		e		e		e	e	e	e

Given the two spatiotemporal timestamped event sequences with 10 timestamps, we compute the similarity between the two spatiotemporal event sequences as:

$$sim(es_1, es_2) = \frac{|es_1 \cap es_2|}{|es_1 \cup es_2|} = \frac{5}{10} = 0.5$$

where, es_1, es_2 are two spatiotemporal event sequences from two locations S1, S2; t1, t2, …, t10 are 10 timestamps. The intersection between two event sequences is the number of co-occurring events between them. We discuss this similarity measure in more detail for five different situations in the following sections.

3.2.3.1 Similarity Measures between Event Sequences without Considering Event Magnitude

First, we compute the level of pairwise co-occurrence between two event sequences es_1 and es_2, $co\_occur(es_1, es_2)$, by simply counting the number of punctual events with the same occurrence time appearing in both es_1 and es_2. So, the global (long duration) similarity between event sequences can be calculated as below:

$$sim_{globlal}(es_1, es_2) = \frac{co\_occur(es_1, es_2)}{|es_1 \cup es_2|} \quad (6)$$

where $sim_{globlal}(es_1, es_2)$—global similarity between event sequences $es_1\ and\ es_2$, meaning overall similarity between two event sequences over a long user specified

duration, $co\_occur(es_1, es_2)$—co-occurring number of events between sequences es_1 and es_2,

$|es_1 \cup es_2|$—cardinality of the union of two event sequences es_1 and es_2.

In contrast to global similarity, we introduce a user defined local comparison temporal window (*ctw*) (equivalent to a moving window), for which local (short duration) similarity is calculated as:

$$sim_{local}(es_1\_ctw_i,\ es_2\_ctw_i) = \frac{co\_occur(es_1\_ctw_i, es_2\_ctw_i)}{|es_1\_ctw_i \cup es_2\_ctw_i|} \quad (7)$$

where, $i = 1, 2, 3, \ldots, k$; $k = \frac{Temporal\ Length\ of\ Event\ Sequence}{ctw}$, the number of time window chunks in an event sequence; $|es_1\_ctw_i \cup es_2\_ctw_i|$, cardinality of the union of two corresponding subsequences of two event sequences in the same ctw. For each pair of spatiotemporal event sequences, we have k local similarities in an ordered list, represented as $\left(sim^1_{local,}\ sim^2_{local,}\ sim^3_{local,}\ \ldots, sim^k_{local}\right)$.

3.2.3.2. Similarity Measures between Event Sequences Considering Event Magnitude

We first find all co-occurrence time points between two event sequences es_1 and es_2, and then we calculate the similarity between two individual events at the co-occurrence timestamp based on their level of measurement. We have two similarity calculation situations. First, if event values are interval or ratio level, the global similarity can be calculated as below:

$$sim_{globlal}(es_1, es_2) = \frac{\sum_{j=1}^{C}(1 - Abs(lev(es_{1j}) - lev(es_{2j})))}{|es_1 \cup es_2|} \quad (8)$$

Second, if event levels are ordinal attribute based, the formula becomes:

$$sim_{globlal}(es_1, es_2) = \frac{\sum_{j=1}^{C}(1-\frac{Abs(lev(es_{1j})-lev(es_{2j}))}{n-1})}{|es_1 \cup es_2|} \quad (9)$$

where,

$sim_{globlal}(es_1, es_2)$—global similarity between event sequences es_1 and es_2

es_{1j}, es_{2j}—the event levels of two corresponding co-occurring events in es_1 and es_2 at timestamp j, inherited from original measurements,

$lev(es_{1j}), lev(es_{2j})$—the relative event levels of two corresponding co-occurring events in es_1 and es_2 at timestamp j, respectively:

$$lev(es_{1j}) = \frac{es_{1j}}{es_{1j}+es_{2j}} \text{ and } lev(es_{2j}) = \frac{es_{2j}}{es_{1j}+es_{2j}}$$

where,

C—the total number of co-occurring timestamps,

$Abs(lev(es_{1j}) - lev(es_{2j}))$—absolute value of difference between relative event levels of two corresponding co-occurring events in es_1 and es_2 at time stamp j,

$|es_1 \cup es_2|$—cardinality of the union of two event sequences es_1 and es_2,

n—the number of ordinal attribute-based event levels.

Similarly, we can characterize the local similarity between event sequences by the following Equation (11) for interval/ratio attribute-based events and (12) for ordinal attribute-based events:

$$sim_{local}(es_1\_ctw_i,\ es_2\_ctw_i) = \frac{\sum_{j=1}^{c}(1-Abs\left(lev\left(es_{1ctw_{ij}}\right)-lev\left(es_{2ctw_{ij}}\right)\right))}{\left|es_{1ctw_i} \cup es_{2ctw_i}\right|} \quad (10)$$

and

$$sim_{local}(es_{1\_}ctw_i,\ es_{2\_}ctw_i) = \frac{\sum_{j=1}^{C}\left(1-\frac{Abs(lev\left(es_{1ctw_{ij}}\right)-lev\left(es_{2ctw_{ij}}\right)}{n-1}\right)}{\left|es_{1ctw_{ij}} \cup es_{2ctw_{ij}}\right|} \quad (11)$$

where, $i = 1, 2, 3, \ldots, k$; $k = \frac{Temporal\ Length\ of\ Event\ Sequence}{ctw}$, c is the number of co-occurring time points in ctw, $|es_{1ctw_{ij}} \cup es_{2ctw_{ij}}|$, cardinality of the union of two corresponding subsequences of two event sequences in the same ctw. As before for each pair of spatiotemporal event sequences, we have k local similarities in an ordered list, represented as $\left(sim^1_{local,}\ sim^2_{local,}\ sim^3_{local,}\ \ldots, sim^k_{local}\right)$.

We note that the approaches for measuring sequence similarity as described above apply also to interval event sequences. We simply transform interval event sequences to punctual event vectors to form a matrix.

3.3. Results and Discussion

3.3.1. Implementation Examples

In this section we use simulated precipitation and temperature datasets to demonstrate the transformation of raw space- time series observations to event sequence matrices based on the event definitions described in the previous section. We calculate global and local pairwise event sequence similarities according to the steps described above. The transformation to STES matrices and the similarity measure calculations have been developed as R functions (see the link for software availability). The first two experiments cover timestamped events based on simulated precipitation measurements for 5 locations and 20 timestamps as shown in Table 3.1. We note that these timestamps could apply to different temporal granularities, but some minimum granularity is considered a punctual timestamp.

Table 3.1. Simulated precipitation measurements in 5 locations with 20 timestamps.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20
s1	0.22	0.35	1.20	0.56	3.10	2.20	1.30	1.77	0.30	0.00	1.00	0.55	2.10	0.50	1.55	0.80	0.20	1.20	1.50	2.20
s2	0.25	2.50	0.40	1.67	2.80	2.10	1.50	0.60	0.20	0.00	1.00	0.44	2.00	0.33	1.23	1.80	0.10	0.10	1.80	2.10
s3	0.28	2.10	0.45	1.45	2.40	1.80	0.44	0.80	0.10	0.00	1.00	0.70	1.50	0.80	1.50	1.20	0.00	0.00	1.60	2.00
s4	0.31	1.70	0.50	1.23	0.50	0.60	0.55	2.10	0.20	0.00	0.00	1.50	0.50	2.10	0.22	1.60	0.10	0.22	0.10	1.90
s5	0.34	1.60	0.55	1.01	0.60	0.67	1.66	1.80	0.10	0.00	0.00	1.40	0.70	2.50	0.52	1.90	1.15	0.30	0.50	1.80

Situation 1. We define precipitation ≥ 1 inch as events from the dataset in Table 3.1 and based on Equation (3) we transform the measurements to a matrix of binary punctual events:

$$\begin{array}{c} \textbf{Temporal points (1, 2, 3, ..., 20)} \\ \textbf{Spatial locations (1, 2, 3, 4, 5)} \end{array} \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

In an alternate view of this matrix seen in Figure 3.4. we show local comparative temporal windows based on 4 timestamps, i.e., $ctw = 4$:

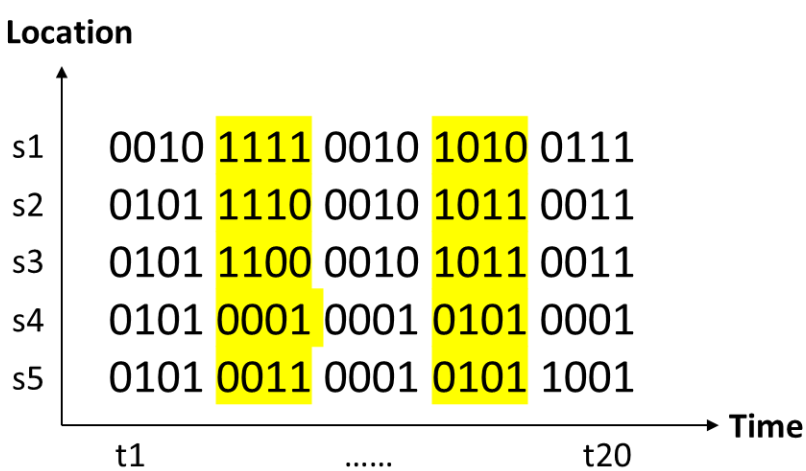


Figure 3.4. A schematic view of the punctual event matrix of Situation 1 with 5 local comparison temporal windows. Blocking 2 columns in yellow is intended to improve visual separation of the local windows.

The pairwise global similarity and local similarity between event sequences were calculated based on Equations (7) and (8). Here, $ctw = 4$, so we have 5 chunks of subsequences for each original event sequence. The pairwise similarity measures between event sequences of 5 locations is shown in Figure 3.5.

	Global similarity	Local similarity with 5 windows				
	$\boldsymbol{entire}$	$\boldsymbol{ctw}1$	$\boldsymbol{ctw}2$	$\boldsymbol{ctw}3$	$\boldsymbol{ctw}4$	$\boldsymbol{ctw}5$
$\boldsymbol{s1-s2}$	0.57	0.00	0.75	1.00	0.67	0.67
$\boldsymbol{s1-s3}$	0.50	0.00	0.50	1.00	0.67	0.67
$\boldsymbol{s1-s4}$	0.13	0.00	0.25	0.00	0.00	0.33
$\boldsymbol{s1-s5}$	0.18	0.00	0.50	0.00	0.00	0.25
$\boldsymbol{s2-s3}$	0.91	1.00	0.67	1.00	1.00	1.00
$\boldsymbol{s2-s4}$	0.29	1.00	0.00	0.00	0.25	0.50
$\boldsymbol{s2-s5}$	0.33	1.00	0.25	0.00	0.25	0.33
$\boldsymbol{s3-s4}$	0.31	1.00	0.00	0.00	0.25	0.50
$\boldsymbol{s3-s5}$	0.27	1.00	0.00	0.00	0.25	0.33
$\boldsymbol{s4-s5}$	0.78	1.00	0.50	1.00	1.00	0.50

Figure 3.5. Output matrix of local similarity with five temporal windows and global similarity between five spatiotemporal event sequences from Situation 1.

By intuition, the event sequences in locations *s*2 and *s*3 are more similar than other pairs with only one mismatch, which is reflected in the global similarity matrix with the highest score of 0.91. The lowest similarity score is between *s*1 and *s*4 event sequences with only two co-occurring events. The rest of the similarity scores for other pairwise comparisons reflect their closeness in terms of co-occurrences.

Situation 2. We again extract precipitation ≥ 1 inch-events from the dataset in Table 3.1 but now consider the magnitude of the precipitation ≥ 1 inch by retaining the original

observation values. Based on the transformation rules described in Equation (4) we obtain the event matrix with event levels as follows:

$$
\begin{array}{c}
\textbf{Temporal points (1, 2, 3, ..., 20)} \\
\textbf{Spatial locations (1, 2, 3, 4, 5)} \quad \begin{pmatrix}
0 & 0 & 1.2 & 0 & 3.1 & 2.2 & 1.3 & 1.77 & 0 & 0 & 1 & 0 & 2.1 & 0 & 1.55 & 0 & 0 & 1.2 & 1.5 & 3.2 \\
0 & 2.5 & 0 & 1.67 & 2.8 & 2.1 & 1.5 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 1.23 & 1.8 & 0 & 0 & 1.8 & 2.1 \\
0 & 2.1 & 0 & 1.45 & 2.4 & 1.8 & 0 & 0 & 0 & 0 & 1 & 0 & 1.5 & 0 & 1.5 & 1.2 & 0 & 0 & 1.6 & 2 \\
0 & 1.7 & 0 & 1.23 & 0 & 0 & 0 & 2.1 & 0 & 0 & 0 & 1.5 & 0 & 2.1 & 0 & 1.6 & 0 & 0 & 0 & 1.9 \\
0 & 1.6 & 0 & 1.01 & 0 & 0 & 1.66 & 1.8 & 0 & 0 & 0 & 1.4 & 0 & 2.5 & 0 & 1.9 & 1.15 & 0 & 0 & 1.8
\end{pmatrix}
\end{array}
$$

The alternate view of this event matrix is shown in Figure 3.6. Where Equations (9) and (11) are used to calculate the global and local similarity respectively:

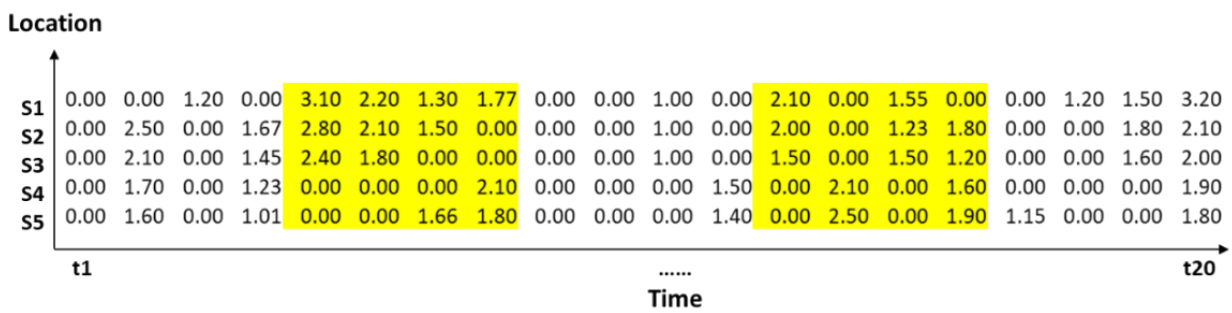


Figure 3.6. A schematic view of the punctual event matrix of **Situation 2** while considering varying event levels with 5 temporal comparison windows. Blocking 2 columns in yellow is for better visual separation of 5 local windows.

From the similarity matrix in Figure 3.7 we can see the change of similarity scores from the results shown in Figure 3.5 that do not take event magnitude into consideration. While all scores in Figure 3.7 decrease compared to Figure 3.5, the overall rankings of these scores are the same. This indicates that refinement of event levels and additional attributes of events incorporated into the similarity measure can affect the similarity values but rankings between event sequences remain stable.

$$
\begin{array}{c}
\\ s1-s2 \\ s1-s3 \\ s1-s4 \\ s1-s5 \\ s2-s3 \\ s2-s4 \\ s2-s5 \\ s3-s4 \\ s3-s5 \\ s4-s5
\end{array}
\begin{array}{c}
\underline{\text{Global similarity}} \\
\begin{Bmatrix} entire \\ 0.53 \\ 0.46 \\ 0.11 \\ 0.16 \\ 0.84 \\ 0.26 \\ 0.30 \\ 0.28 \\ 0.23 \\ 0.73 \end{Bmatrix}
\end{array}
\begin{array}{c}
\underline{\text{Local similarity with 5 windows}} \\
\begin{Bmatrix} ctw1 & ctw2 & ctw3 & ctw4 & ctw5 \\
0.00 & 0.72 & 1.00 & 0.62 & 0.58 \\
0.00 & 0.45 & 1.00 & 0.61 & 0.59 \\
0.00 & 0.23 & 0.00 & 0.00 & 0.27 \\
0.00 & 0.47 & 0.00 & 0.00 & 0.20 \\
0.93 & 0.62 & 1.00 & 0.87 & 0.96 \\
0.85 & 0.00 & 0.00 & 0.24 & 0.48 \\
0.81 & 0.24 & 0.00 & 0.24 & 0.31 \\
0.91 & 0.00 & 0.00 & 0.22 & 0.49 \\
0.86 & 0.00 & 0.00 & 0.20 & 0.32 \\
0.94 & 0.46 & 0.97 & 0.92 & 0.49 \end{Bmatrix}
\end{array}
$$

Figure 3.7. Output matrix including local similarity with five temporal windows and global similarity with consideration of events with variable class levels between five spatiotemporal event sequences from **Situation 2**.

The following examples for interval events are based on the temperature graph for five locations shown in Figure 3.8.

Situation 3. Here we identify interval events ≥ 10 ℃ from high frequency temperature measurements at 5 locations. Assume that a minimum temporal granularity is specified (e.g., day, hour) such that we can obtain the measurements at all time points (t1, t2, …, t20) as in the dataset shown in Table 3.2. Using **Equation** (**5**), we obtain interval events as a sequence of contiguous 1s in a binary event matrix.

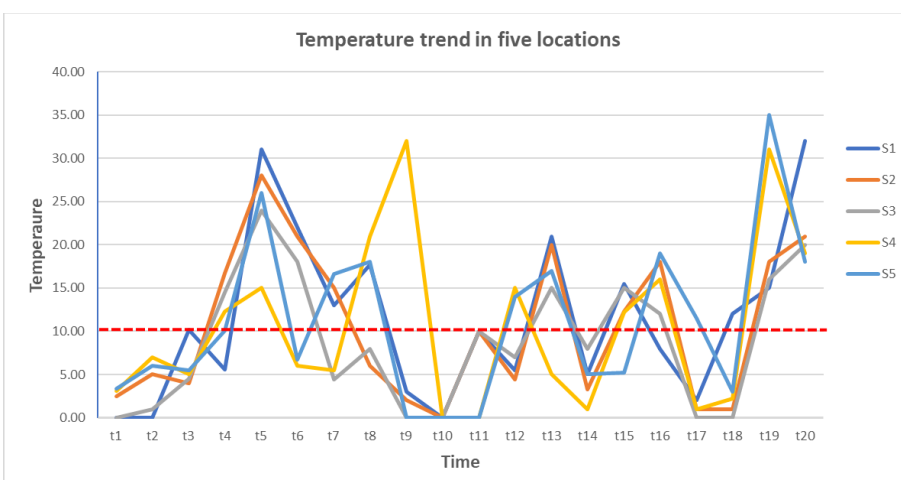


Figure 3.8. Simulated temperature trend in 5 locations over 20 time units. Notice that the red dashed horizontal line is the applied threshold value of 10℃.

Table 3.2. Extracted temperature measurements at 20 time points from continuous data.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20
S1	0.00	0.00	10.20	5.60	31.00	22.00	13.00	17.70	3.00	0.00	10.00	5.50	21.00	5.00	15.50	8.00	2.00	12.00	15.00	32.00
S2	2.50	5.00	4.00	16.70	28.00	21.00	15.00	6.00	2.00	0.00	10.00	4.40	20.00	3.30	12.30	18.00	1.00	1.00	18.00	21.00
S3	0.00	1.00	4.50	14.50	24.00	18.00	4.40	8.00	0.00	0.00	10.00	7.00	15.00	8.00	15.00	12.00	0.00	0.00	16.00	20.00
S4	3.10	7.00	5.00	12.30	15.00	6.00	5.50	21.00	32.00	0.00	0.00	15.00	5.00	1.00	12.20	16.00	1.00	2.20	31.00	19.00
S5	3.40	6.00	5.50	10.10	26.00	6.70	16.60	18.00	0.00	0.00	0.00	14.00	17.00	5.00	5.20	19.00	11.50	3.00	35.00	18.00

The sequence of contiguous 1's represents interval events, but these are processed as punctual events in the event sequence matrix:

$$\begin{array}{c} \textbf{Temporal points (1, 2, 3, ..., 20)} \\ \textbf{Spatial locations (1, 2, 3, 4, 5)} \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

The alternative view of the interval event matrix in the example of Situation 3 can be seen in Figure 3.9. In this figure, we also assume that the comparative temporal window has 10 timestamps, i.e., *ctw* = 10 such that we have only 2 subsequences.

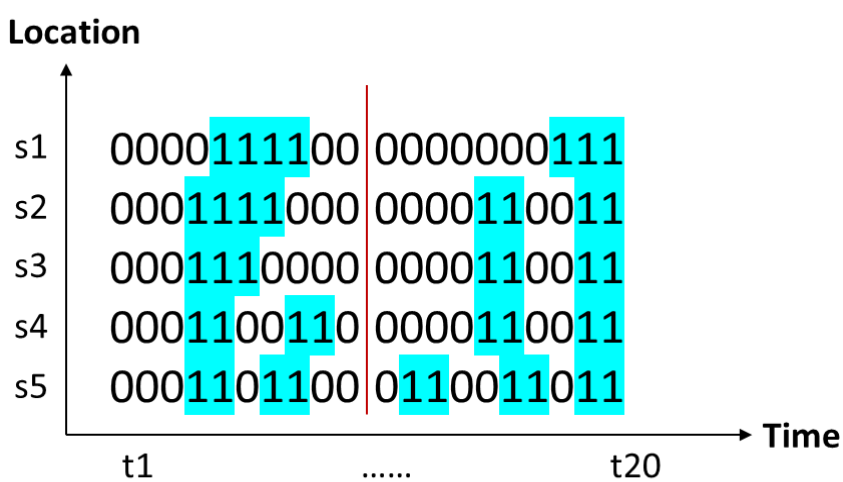


Figure 3.9. A schematic view of the interval event matrix of Situation 3 with binary events and with two temporal windows separated by a red vertical line. Notice that the chunks blocked with blue color in horizontal orientation represent interval events.

The pairwise global similarity and local similarity between event sequences is calculated with the Formulas (7) and (8). Here, $ctw = 10$, so we have 2 chunks of subsequences for each original event sequence. The pairwise similarity matrices between event sequences for the 5 locations is shown in Figure 3.10.

	Global similarity	Local similarity with 2 windows	
	$entire$	$ctw1$	$ctw2$
$s1 - s2$	0.50	0.60	0.40
$s1 - s3$	0.40	0.40	0.40
$s1 - s4$	0.36	0.33	0.40
$s1 - s5$	0.42	0.60	0.29
$s2 - s3$	0.88	0.75	1.00
$s2 - s4$	0.60	0.33	1.00
$s2 - s5$	0.50	0.60	0.43
$s3 - s4$	0.67	0.40	1.00
$s3 - s5$	0.42	0.40	0.43
$s4 - s5$	0.50	0.60	0.43

Figure 3.10. Output matrix of local similarity with two temporal windows and global similarity between five spatiotemporal event sequences from Situation 3.

The event sequences for locations *s*2 and *s*3 in Figure 3.10 are more similar than other pairs with only one mismatch at one timepoint, which is reflected in the global similarity matrix with the highest score of 0.88. The lowest similarity is between *s*1 and *s*4 event sequences with only four co-occurring timepoints and a relatively long union of events. The rest of the similarity scores reasonably reflect their actual closeness.

Situation 4. For the interval events of Situation 3 with the consideration of event level, i.e., the variation of event values within the interval, we obtain a matrix of interval events based on Equation (7) as below:

The sequence of contiguous values represents interval events, but these are processed as punctual events in the event sequence matrix:

$$\begin{array}{c} \\ \textbf{Spatial locations} \\ \textbf{(1, 2, 3, 4, 5)} \end{array} \begin{array}{c} \textbf{Temporal points (1, 2, 3, ..., 20)} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 31.0 & 22.0 & 13.0 & 17.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12.0 & 15.0 & 32.0 \\ 0 & 0 & 0 & 16.7 & 28.0 & 21.0 & 15.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12.3 & 18.0 & 0 & 0 & 18.0 & 21.0 \\ 0 & 0 & 0 & 14.5 & 24.0 & 18.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15.0 & 12.0 & 0 & 0 & 16.0 & 20.0 \\ 0 & 0 & 0 & 12.3 & 15.0 & 0 & 0 & 21.0 & 32.0 & 0 & 0 & 0 & 0 & 0 & 12.2 & 16.0 & 0 & 0 & 31.0 & 19.0 \\ 0 & 0 & 0 & 10.1 & 26.0 & 0 & 16.6 & 18.0 & 0 & 0 & 0 & 14.0 & 17.0 & 0 & 0 & 19.0 & 11.5 & 0 & 35.0 & 18.0 \end{pmatrix} \end{array}$$

The alternative view of the interval event matrix in the example of Situation 4 is represented in Figure 3.11. In this figure, we assume that the comparative temporal window has 10 timestamps, i.e., *ctw* = 10.

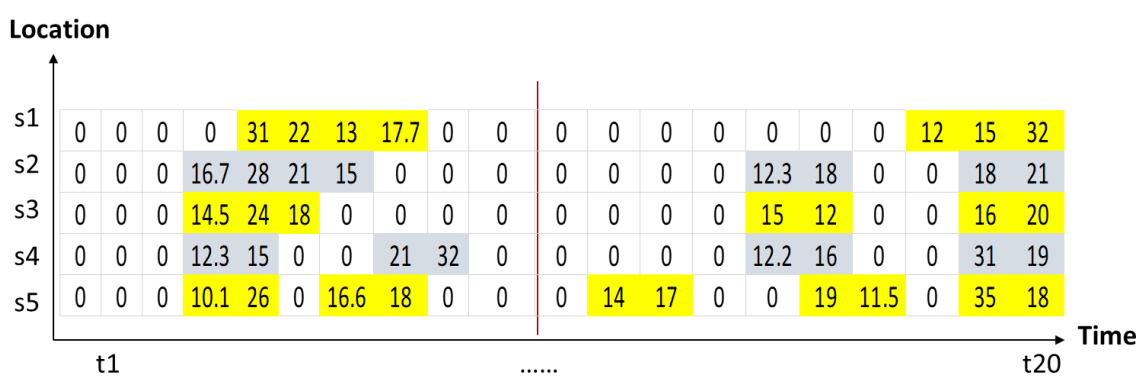


Figure 3.11. A schematic view of the interval event matrix of Situation 4 with consideration of event level and variation between starting and ending time points with 2 temporal comparison windows. Notice that the chunks blocked with colors in the horizontal orientation represent interval events.

The pairwise global similarity and local similarity between event sequences is calculated with Equations (9) and (11) for this situation. Here, $ctw = 10$, so we have 2 chunks of subsequences for each original event sequence. The pairwise similarity matrices between event sequences of 5 locations is shown in Figure 3.12.

	Global similarity	Local similarity with 2 windows	
	$entire$	$ctw1$	$ctw2$
$s1 - s2$	0.46	0.57	0.35
$s1 - s3$	0.36	0.36	0.36
$s1 - s4$	0.29	0.28	0.31
$s1 - s5$	0.36	0.56	0.21
$s2 - s3$	0.81	0.70	0.92
$s2 - s4$	0.53	0.27	0.92
$s2 - s5$	0.45	0.54	0.38
$s3 - s4$	0.58	0.35	0.88
$s3 - s5$	0.36	0.36	0.36
$s4 - s5$	0.46	0.53	0.41

Figure 3.12. Output matrix of global similarity and local similarity with two temporal windows considering events with ratio level values between five spatiotemporal event sequences from Situation 4.

This situation considers the internal variation within an interval event along with co-occurrences. We can compare the similarity scores in Figure 3.12 with those in Figure 3.10. Like Situation 2, the overall similarity values decrease compared to the situations without considering event magnitude. We can see here the event sequences at locations *s*2 and *s*3 in Figure 3.12 are still more similar than other pairs with slight variations of event values

between co-occurring timepoints, which can be reflected in the global similarity matrix with the highest score of 0.81. The lowest similarity (0.29) remains between *s*1 and *s*4 as in Situation 3. The rest of the similarity scores for other pairwise comparisons also reasonably reflect an intuitive sequence closeness.

A special case in Situation 4. If the temperature measurements are recorded as an average value for every four days as shown in Table 3.3.

Table 3.3. Simulated averaged temperature measurements for every 4 time units in 5 locations.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20
S1	9.8	9.8	9.8	9.8	22.0	22.0	22.0	22.0	12.0	12.0	12.0	12.0	8.8	8.8	8.8	8.8	31.0	31.0	31.0	31.0
S2	9.1	9.1	9.1	9.1	28.0	28.0	28.0	28.0	14.0	14.0	14.0	14.0	5.0	5.0	5.0	5.0	26.0	26.0	26.0	26.0
S3	11.0	11.0	11.0	11.0	24.0	24.0	24.0	24.0	11.0	11.0	11.0	11.0	7.0	7.0	7.0	7.0	28.0	28.0	28.0	28.0
S4	14.0	14.0	14.0	14.0	25.0	25.0	25.0	25.0	18.0	18.0	18.0	18.0	9.0	9.0	9.0	9.0	33.0	33.0	33.0	33.0
S5	8.0	8.0	8.0	8.0	18.0	18.0	18.0	18.0	12.0	12.0	12.0	12.0	15.0	15.0	15.0	15.0	24.0	24.0	24.0	24.0

We can transform this dataset to a matrix of interval events with event levels based on Equation (6) as shown in the matrix below:

$$\begin{array}{c} \\ \textbf{Spatial locations} \\ \textbf{(1, 2, 3, 4, 5)} \end{array} \begin{array}{c} \textbf{Temporal points (1, 2, 3, ..., 20)} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 22 & 22 & 22 & 22 & 12 & 12 & 12 & 12 & 0 & 0 & 0 & 0 & 31 & 31 & 31 & 31 \\ 0 & 0 & 0 & 0 & 28 & 28 & 28 & 28 & 14 & 14 & 14 & 14 & 0 & 0 & 0 & 0 & 26 & 26 & 26 & 26 \\ 11 & 11 & 11 & 11 & 24 & 24 & 24 & 24 & 11 & 11 & 11 & 11 & 0 & 0 & 0 & 0 & 28 & 28 & 28 & 28 \\ 14 & 14 & 14 & 14 & 25 & 25 & 25 & 25 & 18 & 18 & 18 & 18 & 0 & 0 & 0 & 0 & 33 & 33 & 33 & 33 \\ 0 & 0 & 0 & 0 & 18 & 18 & 18 & 18 & 12 & 12 & 12 & 12 & 15 & 15 & 15 & 15 & 24 & 24 & 24 & 24 \end{pmatrix} \end{array}$$

The alternative view of the interval event matrix for this example of average temperature can be seen in Figure 3.13. In this figure, we also assume that the comparative temporal window has 10 timestamps, i.e., *ctw* = 10.

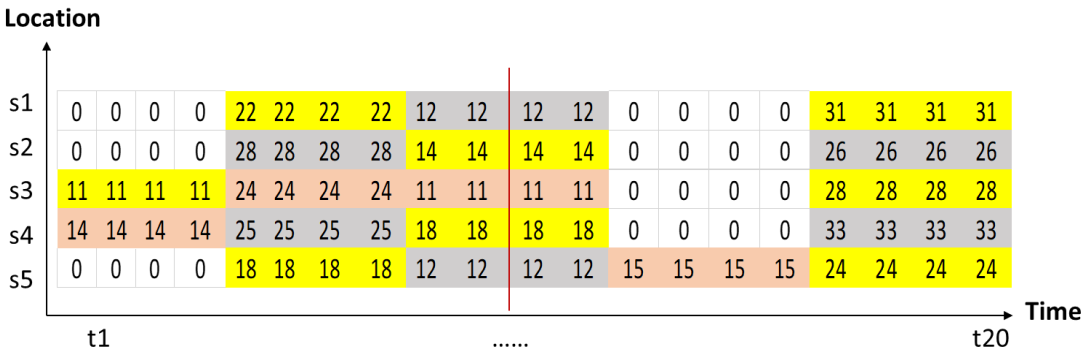


Figure 3.13. A schematic view of the interval event matrix of Situation 4 with consideration of event level and no variation between starting and ending time points with 2 temporal comparison windows. Notice that the chunks blocked with colors in the horizontal orientation represent interval events.

The pairwise global similarity and local similarity between event sequences can be calculated with Equations (9) and (11). Here, *ctw* = 10, so we have 2 chunks of subsequences for each original event sequence. The pairwise similarity matrices between event sequences of 5 locations is shown in Figure 3.14.

	Global similarity	Local similarity with 2 windows	
	$entire$	$\boldsymbol{ctw}1$	$\boldsymbol{ctw}2$
$\boldsymbol{s1-s2}$	0.91	0.90	0.92
$\boldsymbol{s1-s3}$	0.72	0.58	0.95
$\boldsymbol{s1-s4}$	0.69	0.54	0.92
$\boldsymbol{s1-s5}$	0.70	0.94	0.55
$\boldsymbol{s2-s3}$	0.70	0.55	0.94
$\boldsymbol{s2-s4}$	0.68	0.56	0.89
$\boldsymbol{s2-s5}$	0.68	0.86	0.57
$\boldsymbol{s3-s4}$	0.90	0.91	0.88
$\boldsymbol{s3-s5}$	0.55	0.54	0.56
$\boldsymbol{s4-s5}$	0.51	0.51	0.51

Figure 3.14. Output matrix of local similarity with two temporal windows and global similarity considering event magnitude between five spatiotemporal event sequences based on the special case of **Situation 4.**

Intuitively we can see the event sequences between locations *s*1 and *s*2, and between locations *s*3 and *s*4 in Figure 3.14 are more similar than other pairs with all co-occurring events of similar value at most timepoints, which can be reflected in the global similarity matrix with the highest score of 0.91 and 0.90. The lowest similarity is 0.51 between *s*4 and *s*5 event sequences with three co-occurring events of different significance and two mismatched events.

3.3.2. Performance Evaluation

In this section we present our experimental evaluation of the accuracy and speed of different similarity measures with some synthetic datasets. In the first experiment, we compared the speed for computing similarity matrices using the small dataset used in this section. For the second, we used K-nearest neighbor (k-NN) classification with different similarity measures for comparing classification accuracy and efficiency.

3.3.2.1. Execution Speed for a Binary Event Matrix

The purpose of this experiment is to assess processing times for the timestamp locked Jaccard based similarity described in this paper (STES.sim1, see the software availability link). We compared STES.sim1 with generic edit distance in R (EditD Dynamic), and two functions of Edit Distance and original Jaccard similarity from the R package Rstringdist. The dataset contains 20 timestamps and 5 locations so we can generate a 5 × 5 similarity matrix. Microbenchmarks (Bershad et al., 1992) in R was used to record the time elapsed for each similarity algorithm in the same similarity matrix generation function in R. The result indicated that STES.sim1 outperformed edit distance by a factor of 10 (Table 3.4).

3.3.2.2. Accuracy Evaluation with Synthetic Datasets Using 1-NN Classifier

K-NN is a conventional non-parametric classifier, used widely as the baseline

Table 3.4. Evaluation of different similarity measures with STES similarity matrix on example data for 100 times (unit: microseconds).

Algorithm	Min	lq	Mean	Median	uq	Max	n_eval
STES.sim1	503	549	676	587	657	2328	100
EditD Dynamic	4904	5250	5942	5474	6319	12,467	100
EditD_Rstringdist	2064	2280	2591	2408	2625	5501	100
Jaccard_Rstringdist	1863	2021	2651	2167	2556	8504	100

classifier for solving many classification problems (Peterson et al., 2005; Prasath et al., 2017). It is based on measuring the distances or similarities between a test data set and each of the training data to decide the final classification output. When proposing a new distance or similarity measure, 1-NN accuracy was strongly recommended for testing (Wang et al., 2013). Note that this does not exclude the additional other trainings and tests with different K values. Here, however, the 1-NN test has the advantage of having no parameters and allowing comparisons between similarity measures.

Synthetic dataset 1: This dataset contains 100 event sequences (records) with 50 timestamped fields of binary values (0, 1) representing whether the event occurred or not. The test uses 3 different event distribution patterns (groups or classes) labeled by A, B and C. The sample function in R with the *prob* argument was used to control density and order of event occurrences. The first pattern (Label A) is characterized with the first 20 timestamps having a higher probability (0.8) of event occurrence and the remainder with lower probability (0.2). In the second pattern (Label B) the subsequence of higher

probability of event occurrence is placed in the middle, and in the third pattern (Label C), the higher probability occurrence region is placed at the end. The event data structure of these three patterns and the observation number of each pattern are graphically depicted in Figure 3.15.

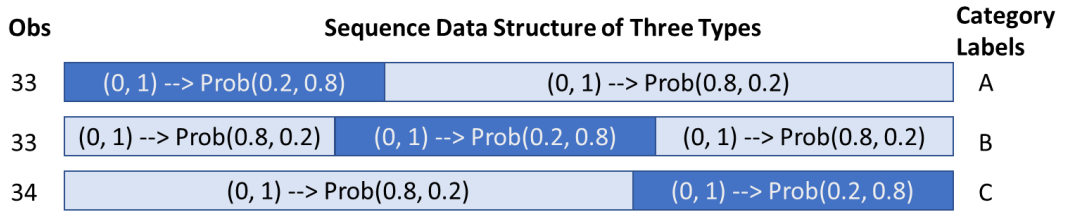
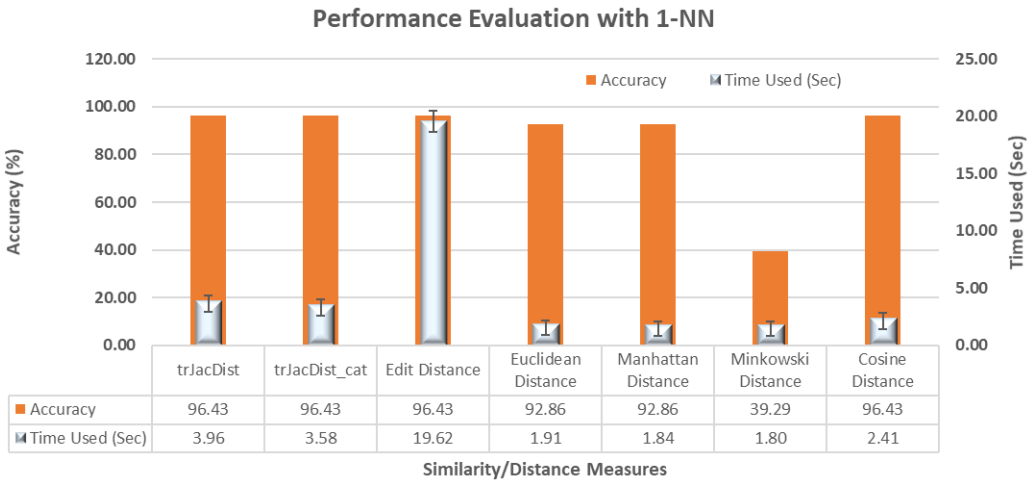


Figure 3.15. Schematic event sequence data structure for synthetic dataset 1 with three different mono-categorical event (0, 1) distribution.

We should note that the binary data (0, 1) can represent either two categories or actual values of 0 and 1. Therefore, both category-based and value-based similarity measures can be applied to this dataset. In this evaluation experiment, the category-based measures include Edit Distance and time-restricted Jaccard Distance for category data (trJacDist-cat) developed in this paper, and the value-based distance measures are Euclidean, Manhattan, Minkowski, and Cosine Distance. When running 1-NN classification test, the dataset with three patterns is first randomized and then divided into 70% training and 30% test set for the experimental setup. Hence, there are 70 training event sequences and 30 test sequences on which classification was performed. The effectiveness of a similarity measure in this experiment is evaluated with accuracy for classifying three patterns of event sequences (Label A, B, and C) and time for completing the task. To capture the fluctuation of time used for each task due to internal computer operation system, we run each 1-NN test for each similarity measure 15 times to compute the error bars.

Using seven similarity measures carried out with 1-NN classification for the dataset mentioned above, Figure 3.16 shows the comparison of accuracy and time elapsed to complete the given task. The effectiveness of different similarity measures can be seen by comparing the accuracy and time required to complete the task. While the same accuracy can be achieved with trJacDist/trJacDist-cat and Edit Distance for classifying this small dataset, the time required with trJacDist measure is about 5 times less than Edit Distance measure. Euclidean, Manhattan and Minkowski Distance algorithms show a time advantage over trJacDist/trJacDist-cat, but slightly lower accuracy. We note that Cosine Distance has similar accuracy but a slightly better time performance.



	trJacDist	trJacDist_cat	Edit Distance	Euclidean Distance	Manhattan Distance	Minkowski Distance	Cosine Distance
Accuracy	96.43	96.43	96.43	92.86	92.86	39.29	96.43
Time Used (Sec)	3.96	3.58	19.62	1.91	1.84	1.80	2.41

Figure 3.16. The bar graph for accuracy and times for 1-NN using seven different similarity measures applied on synthetic dataset 1 with three classes. Note: error bars are based on 15 times of computation for the same task.

Synthetic dataset 2: This dataset contains 100 records (event sequences) with 128 timestamped fields of real values. As shown in Figure 3.17, there are three types of patterns in this dataset: sine, box, and ramp-cliff, each function of which has high level of white noise as the background noise. We excluded the Edit distance in this test as it is inappropriate for real valued data. We compared trJacDist with Euclidean, Manhattan, Minkowski, and Cosine distance-based similarity measures for evaluating the efficiency

and accuracy of classification with 1-NN classifier. The dataset was also randomized and then split into 70% training and 30% test sub-datasets when running 1-NN classification. From the results shown in Figure 3.18 we can see that while trJacDist shows a time disadvantage against these methods it shares the same accuracy with Euclidean, Manhattan, and Cosine distance.

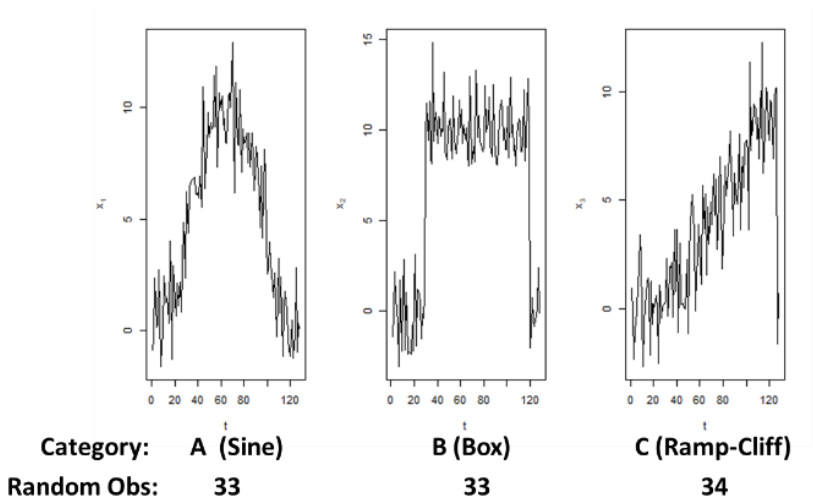


Figure 3.17. Schematic sequence data structure of three types of events (sine, box, and ramp-cliff) with real values for synthetic dataset 2.

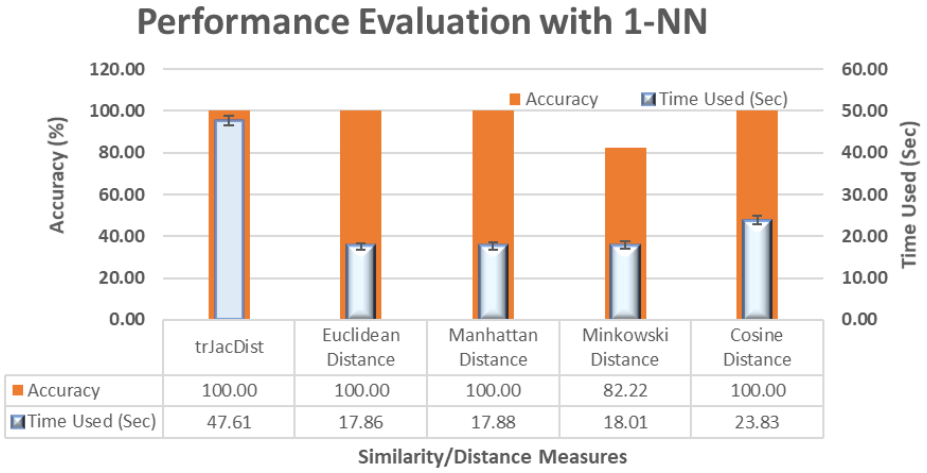


Figure 3.18. The bar graph for accuracy and times for 1-NN using five different similarity measures applied on synthetic dataset 2 with three classes. Note: error bars are based on 15 times of computation for the same task.

3.3.3. Application Example

We examined the feasibility of the proposed framework in the real-world application of monitoring precipitation events obtained from observation stations distributed along the Maine coast. Here we demonstrate the specific steps of eventization and similarity measures developed in this study and we address the question: Do STES that are closer in space show higher similarity measures?

The Maine Department of Marine Resources (DMR) manages the shellfish growing areas in coastal Maine based on the fecal pollution situations observed from more than 2000 monitoring stations. Precipitation events can trigger high levels of fecal coliform in shore waters and are thus of concern. Grouping of similar stations in terms of heavy rain or high precipitation events is useful for allocating the limited labor pool for long term water sampling. We used the similarity measures developed in this study to conduct clustering analysis with the high precipitation event sequences (>=1 in daily) of selected monitoring stations for 5 years.

Considering the daily precipitation is very close between nearby monitoring stations we selected 43 monitoring stations for this experiment in the shellfish growing areas that are well distributed along the Maine coast (Figure 3.19). With daily precipitation data of 5 years, we have an initial 43 × 1826 matrix of precipitation raw data (Table A.1).

The dimensions of the raw data matrix is reduced through the eventization steps developed in this research. In this specific example, we extracted event sequences of either >=1″ or >=2″ precipitation for each monitoring station. Based on Equation (3) we computed the data in Table A.1 with R script (STS.eventize1.R) and created the event sequence matrix of 43 × 192 (>=1″ precipitation) (Table A.2) or 43 × 52 (>=2″ precipitation) (Table A.3). Taking >=1″ precipitation event sequences as an example

(Table A.2) and using the STES similarity measure (STES.sim1.R) from this paper, we created the similarity matrix of 43 × 43 (Table A.4) between selected test

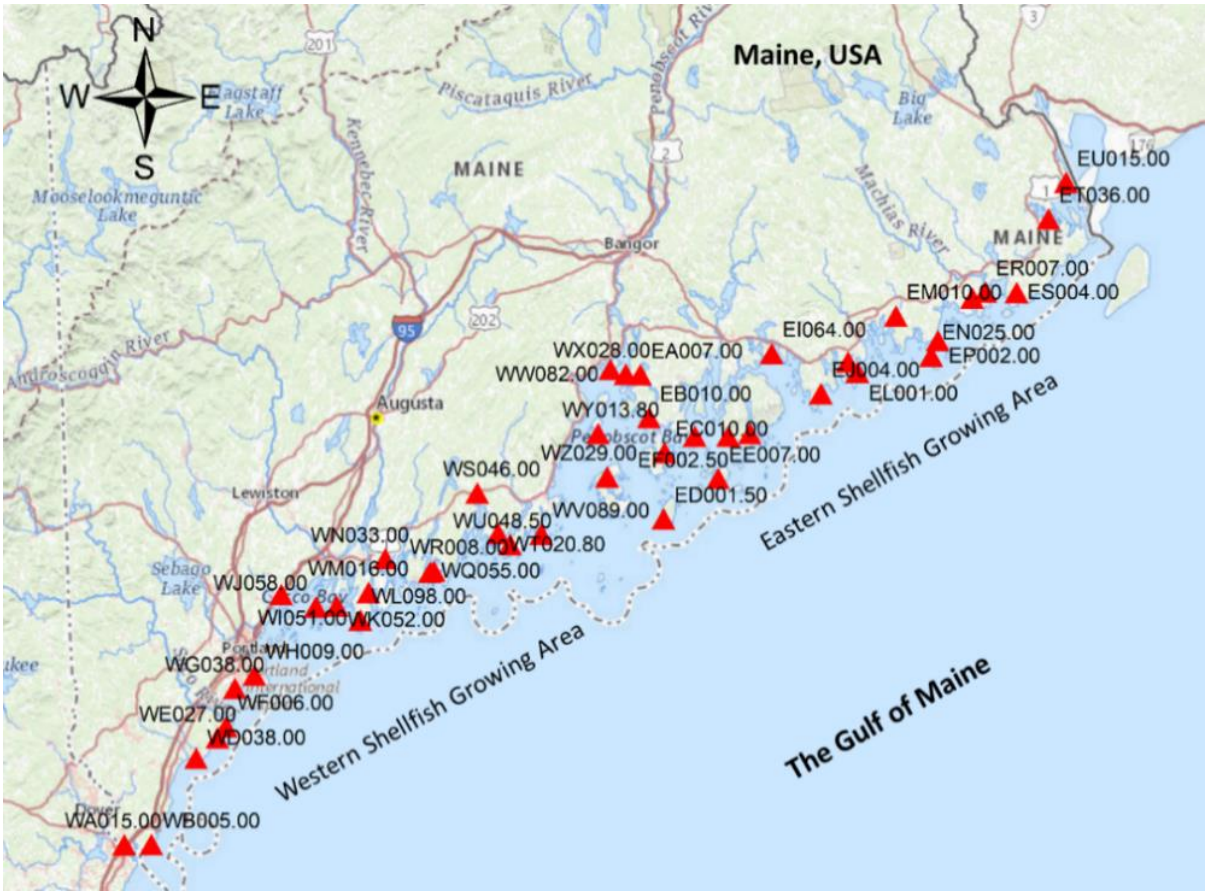


Figure 3.19. Experimental sites along the Maine coast.

monitoring stations. We transformed these similarity data into distance data to conduct hierarchical clustering analysis (Ros and Guillaume, 2019) using the hclust R function with linkage method Ward.D2.

Figure 3.20 shows the clustering results from using STES similarity on event sequences of >=1 in precipitation during 5 years in 43 locations (monitoring stations) as a heatmap and distance-based cluster dendrogram.

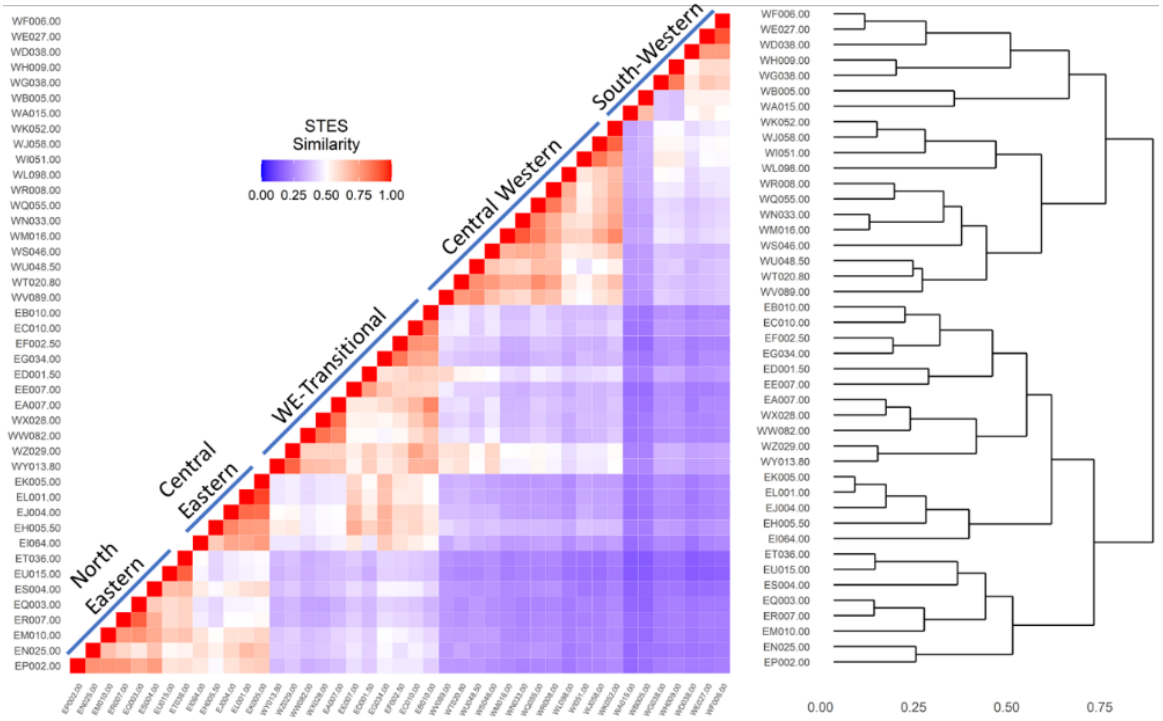


Figure 3.20. STES similarity-based heat map and STES distance based hierarchical clustering between monitoring stations along ME coast in >1 in precipitation events in 5 years (2010–2014).

The results show the emergence of five clusters (groupings of event sequences that are most similar). The heatmap and cluster dendrogram indicate that these clusters are in fact spatial clusters indicating that for this case, sequences that are close in space tend to be more similar. These results can provide decision makers with more information for arranging the labor within each region (cluster) along the Maine coast to collect water samples for fecal coliform measurements from selected stations.

3.4 Conclusions

In this paper, we have demonstrated a matrix-based representation of spatiotemporal event sequences for unifying punctual and interval events. These similarity measures along with the univariate spatiotemporal event matrices for event data storage discussed above

provide a novel method and an alternative foundation for further event sequence pattern discovery. A comparison of event sequence similarity is important for detecting co-occurrence patterns and investigating the influence of event sequences of interest. We assume that similar event sequences indicate a similar process structure and potential shared causal mechanisms.

Based on the analysis of sequence properties for four situations and one special case that consider event co-occurrences and event levels, we have proposed corresponding similarity measures for pairwise comparisons for punctual and interval events and for whole or long duration sequences or their subsequences. The experimental results with simulated datasets showed that these similarity scores between spatiotemporal event sequences reasonably represent perceived closeness.

A comparative evaluation against other similarity algorithms shows the same or better accuracy results. Our method shows a time disadvantage against the real valued methods but a substantial time advantage over the qualitative Edit Distance. Overall, our approach has the advantages of flexibility in that it can accommodate both qualitative and quantitative event values as well as both punctual and interval events.

We recognize some limitations in the current research. This research establishes a framework of matrix representations and similarity development for univariate event sequences of different types. It does not yet handle similarity assessment for multivariate event sequences. Such an extension requires some modification of the matrix representation and similarity measures which will be addressed in future work. In the current work we demonstrate fixed matrix sizes which can be chunked into smaller subsequence sets for local versus global similarity computations. For future work, an

extension that addresses streaming events from monitoring stations would be a useful addition. The addition of temporal logic operations and extensions to consider lagged sequence alignment similarity rather than the time locked case are other considerations for future work. Furthermore, we have not tested the current methods on big data. Future work will focus on the evaluation extensive real datasets from environmental monitoring or other domains. Currently our STES representation includes the intervals between event occurrences. For sequences in which event occurrences may be sparse with long intervening intervals we are considering approaches for sparse matrices. Lastly, we also consider extensions to detect complex events of interest, and incorporation of our methods into Complex Event Processing (CEP) systems.

Chapter References

Abadi, D., Madden, S. and Lindner, W. (2016) Data Stream Management, pp. 409-428, Springer.

André-Jönsson, H. and Badal, D.Z. 1997 Using signature files for querying time-series data, pp. 211-220, Springer.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M. and Poelitz, C. 2010 Extracting events from spatial time series, pp. 48-53, IEEE.

Bershad, B., Draves, R.P. and Forin, A. 1992 Using microbenchmarks to evaluate system performance, pp. 148-153, IEEE.

Bollobas, B., Das, G., Gunopulos, D. and Mannila, H. 1997 Time-series similarity problems and well-separated geometric sets, pp. 454-456.

Chung, N.C., Miasojedow, B., Startek, M. and Gambin, A. 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC bioinformatics 20(15), 1-11.

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 14(7), 1394-1403.

Du, F., Shneiderman, B., Plaisant, C., Malik, S. and Perer, A. 2016. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. IEEE transactions on visualization and computer graphics 23(6), 1636-1649.

Fu, T.-c. 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence 24(1), 164-181.

Guralnik, V. and Srivastava, J. 1999 Event detection from time series data, pp. 33-42, ACM.

Hamming, R.W. 1950. Error detecting and error correcting codes. Bell System technical journal 29(2), 147-160.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. and Caron, E. 2016. A survey of event extraction methods from text for decision support systems. Decision Support Systems 85, 12-22.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37, 547-579.

Jacobs, B.E. and Walczak, C.A. 1983. A generalized query-by-example data manipulation language based on database logic. IEEE Transactions on Software Engineering (1), 40-57.

Jassby, A.D. and Powell, T.M. 1990. Detecting changes in ecological time series. Ecology 71(6), 2044-2052.

Levenshtein, V.I. 1966 Binary codes capable of correcting deletions, insertions, and reversals, pp. 707-710.

Luu, V.-T., Forestier, G., Weber, J., Bourgeois, P., Djelil, F. and Muller, P.-A. 2020. A review of alignment based similarity measures for web usage mining. Artificial Intelligence Review 53(3), 1529-1551.

Mannila, H. and Moen, P. 1999 Similarity between event types in sequences, pp. 271-280, Springer.

Mannila, H. and Ronkainen, P. 1997 Similarity of event sequences, pp. 136-139, IEEE.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V. 2007. Fault diagnosis using dynamic trend analysis: A review and recent developments. Engineering Applications of artificial intelligence 20(2), 133-146.

Mirbagheri, S.M. and Hamilton, H.J. 2020 Similarity Matching of Temporal Event-Interval Sequences, pp. 420-425, Springer.

Obweger, H., Suntinger, M., Schiefer, J. and Raidl, G. 2010 Similarity searching in sequences of complex events, pp. 631-640, IEEE.

Peterson, M.R., Doom, T.E. and Raymer, M.L. 2005 Ga-facilitated knn classifier optimization with varying similarity measures, pp. 2514-2521, IEEE.

Prasath, V., Alfeilat, H.A.A., Lasassmeh, O. and Hassanat, A. 2017. Distance and similarity measures effect on the performance of K-nearest neighbor classifier-a review. arXiv preprint arXiv:1708.04321.

Prinzie, A. and Van den Poel, D. 2011. Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: an Acquisition Pattern Analysis application. Journal of Intelligent Information Systems 36(3), 283-304.

Ros, F. and Guillaume, S. 2019. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Systems with Applications 128, 96-108.

Rude, A. and Beard, K. 2012 High-Level Event Detection in Spatially Distributed Time Series, pp. 160-172, Springer Berlin Heidelberg, Berlin, Heidelberg.

Shahar, Y. 1997. A framework for knowledge-based temporal abstraction. Artificial intelligence 90(1-2), 79-133.

Shurkhovetskyy, G., Andrienko, N., Andrienko, G. and Fuchs, G. 2018 Data abstraction for visualizing large time series, pp. 125-144, Wiley Online Library.

Stehle, S. and Peuquet, D.J. 2015. Analyzing spatio-temporal patterns and their evolution via sequence alignment. Spatial Cognition & Computation 15(2), 68-85.

Tao, C., Wongsuphasawat, K., Clark, K., Plaisant, C., Shneiderman, B. and Chute, C.G. 2012 Towards event sequence representation, reasoning and visualization for EHR data, pp. 801-806, ACM.

Vorontsov, I.E., Kulakovskiy, I.V. and Makeev, V.J. 2013. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms for Molecular Biology 8(1), 1-11.

Wang, T.-Y., Yang, M.-H. and Wu, J.-Y. 2016. Distributed Detection of Dynamic Event Regions in Sensor Networks With a Gibbs Field Distribution and Gaussian Corrupted Measurements. IEEE Transactions on Communications 64(9), 3932-3945.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. and Keogh, E. 2013. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26(2), 275-309.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. and Shneiderman, B. 2012. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with computers 24(2), 55-68.

Yang, J., McAuley, J., Leskovec, J., LePendu, P. and Shah, N. 2014 Finding progression stages in time-evolving event sequences, pp. 783-794, ACM.

Yeh, C.-C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Zimmerman, Z., Silva, D.F., Mueen, A. and Keogh, E. 2018. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Mining and Knowledge Discovery 32(1), 83-123.

Yin, J., Hu, D.H. and Yang, Q. 2009 Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields, pp. 1321-1327.

CHAPTER 4

A NOVEL SIMILARITY MEASURE OF SPATIOTEMPORAL EVENT SETTING SEQUENCES: METHOD DEVELOPMENT AND CASE STUDY

The content in this chapter is the reformatted version of the published research paper:

Xu, F.; Beard, K. A Novel Similarity Measure of Spatiotemporal Event Setting Sequences: Method Development and Case Study. Geographies 2023, 3, 303-320. https://doi.org/10.3390/geographies3020016

Chapter Abstract

Examining the similarity of event environments or surroundings—more precisely, settings—provides additional insight in analyzing event sequences, as it provides information about the context and potential common factors that may have influenced them. This article proposes a new similarity measure for event setting sequences, which involve the space and time in which events occur. While similarity measures for spatiotemporal event sequences have been studied, the settings and setting sequences have not yet been studied. While modeling event setting sequences, we consider spatial and temporal scales to define the bounds of the setting and incorporate dynamic variables alongside static variables. Using a matrix-based representation and an extended Jaccard index, we developed new similarity measures that allow for the use of all variable data types. We successfully used these similarity measures coupled with other multivariate statistical analysis approaches in a case study involving setting sequences and pollution event sequences associated with the same monitoring stations, which validate the hypothesis that more similar spatial-temporal settings or setting sequences may generate

more similar events or event sequences. In conclusion, the developed similarity measures have wide application beyond the case study to other disciplinary contexts and geographical settings. They offer researchers a powerful tool for understanding different factors and their dynamics corresponding to occurrences of spatiotemporal event sequences.

Keywords: event sequence; spatial context; similarity measure; Jaccard index; cluster analysis

4.1. Introduction

An event setting, or more explicitly a spatiotemporal event setting, can be defined as a space and its collective influencing factors which are related to the occurrence of an event or sequence of events at a specific time and location. It can refer to the physical location, such as a specific venue or building, or to the overall atmosphere and environs or surroundings of an event. Similarity measures between events and event sequences have been well studied (Guralnik and Srivastava, 1999; Lupiani et al., 2013; Mannila and Ronkainen, 1997; Moen, 2000; Obweger et al., 2010; Wongsuphasawat et al., 2012; Xu and Beard, 2021). Assessing similarity between event settings adds another dimension to event sequence analysis in that it offers context and information on potential shared influencing factors. We hypothesize that the occurrences of at least some types of events and event sequences are likely to be related to the spatiotemporal settings from which they arise. In other words, spatiotemporal differentials in environmental settings contribute to variations in levels and patterns of event occurrences and event sequences.

While, as noted above, event sequence similarity has been well researched, no such similarity measures for event sequence settings have been found in the literature to date.

This paper addresses this gap by developing similarity measures for event sequence settings. In (Xu and Beard, 2021), we established similarity measures for comparing event sequences and demonstrated their potential applications. In this study, we question whether similar patterns of event sequences reflect similarity in the spatiotemporal settings of the event sequences. A working hypothesis is that more similar spatial settings may generate more similar event sequences.

Measures of similarity among event sequence settings have several potential applications in real world contexts. First, in predicting future events or phenomena, similarity measures can be used to identify patterns in the spatial-temporal settings of past events or phenomena, which can help predict the likelihood of similar events occurring in the future. For example, a similarity measure could be used to predict the likelihood of a hurricane occurring in a particular region based on the spatial-temporal settings of past hurricanes in the region. Second, for better understanding the spread of diseases or other public health concerns, similarity measures can identify patterns in the spatial-temporal settings of disease outbreaks or other public health concerns. Such information can help public health officials understand how diseases or other health concerns spread and take steps to prevent or mitigate their impact. In analyzing the impact of climate change, similarity measures can identify patterns in the spatial-temporal settings of natural disasters or other events that may be influenced by climate change. Setting similarity information in this context can help policymakers and researchers understand the potential impacts of climate change and take steps to mitigate those impacts. In analyzing the distribution of resources or services, similarity measures can help identify patterns in the spatial-temporal settings of resource distribution or service delivery, which can help policymakers and

service providers understand where resources or services are most needed and how to allocate them effectively. Similarity measures can also help identify patterns in the spatial-temporal settings of human activities, such as economic development or land use planning, among other areas covering the natural and social sciences.

To develop an event sequence setting similarity metric, we propose to characterize the environs where a time series has been observed, and, by extension, event sequences have been derived. How to describe and determine what constitutes such environs raises various challenges for conceptualizing such a space. Recent geographical research has pointed to the need for more careful and critical evaluations of our conceptualizations of space (Malpas, 2012; Paasi, 2004; Simandan, 2020), advocating for concepts that avoid predetermined hierarchies or boundaries. The concept of 'site' (Marston et al., 2005; Schatzki, 2002; Woodward et al., 2012) is one such option. Site has been presented as an organizationally autonomous, subject-independent 'event-space' where something occurs. It has a processual focus of differentiated and differentiating forces at work that contribute to its assembly. The authors clearly state that a site is not a fixed space in the sense of a setting, context, or place for action but dynamic, unbounded, and subject to compositional variation. This process-focused view of site has much conceptual appeal as a dynamic generative source of events. While recognizing the value and applicability of this site concept, in order to proceed with a similarity comparison among a set of environs, we revert to a need for some bounding constraints. The term setting has received less attention but appears in (Worboys and Hornsby, 2004) with reference to events. They present events as situated within a setting which may be spatial, temporal, or spatiotemporal. The setting is assumed to have extent in either or both space and time, and an appropriate scale. What

constitutes a setting, scale, and an appropriate scale yields yet more conceptual turmoil. Recent reflections on scale (Marston et al., 2005; Moore, 2008; Paasi, 2004) suggest attention to processes or practices that are differentially scaled rather than delimited extents. We return to this issue in Section 2.

Context is another term that could apply to an environ. Context has numerous definitions, many with location as the emphasis, namely spatial context. Context has been described as the "location and the identity of nearby people and objects." (Jiang and Yao, 2007). Contexts can include factors such as the natural environment, climate, culture, economic conditions, and population characteristics. Spatial-temporal context can refer to the historical and cultural background of a place, as well as the relationships and interactions that have occurred within that place over time. Context is an important factor in many domains and applications. In computer science, context has referred to any information available for characterizing the situation of an entity, where entity could be a person, place, or object, which is related to the interaction between a user and an application (Brézillon and Gonzalez, 2014; Dey, 2001; Loke, 2006). In geography, it has referred to physical and social conditions that exist in a particular place and time (Gong and Hassink, 2020; Sunley, 1996; Weber and Kwan, 2003; Zolnik, 2009). In human geography, the importance of contexts has given rise to an epistemological framework of situated knowledge (Simandan, 2019), which argues that our knowledge is contextualized by geographical location among other influences we may or may not be aware of.

Spatial context strongly influences the transport disadvantage that in turn affects social exclusion and well-being (Delbosc and Currie, 2011). In travel behavior research, spatial context was shown to be strongly related to household travel patterns at an

international scale (Timmermans et al., 2003). A person's health-related problems can be strongly affected by different types of spatial context, such as environmental exposures (Cutter, 1996; Roux and Mair, 2010), social environment (characteristics of communities and neighborhoods) (Roux and Mair, 2010; Sampson, 2003), and ease of access to health services (Yang et al., 2006). Spatial context greatly influences the potential of getting a disease, the adoption of healthy lifestyle, and the ease of access to medical services for disease diagnosis and treatment. An early psychological behavior research study indicates that decision behavior is affected by spatial context or spatially varied factors (Wolpert, 1964). A farming population was selected to study the effect of spatial context in decision processes because the outcomes of decision behavior are easily observable over the landscape. The decision making in farming is dispersed spatially among many farmers due to the uneven diffusion of market and technical information. With the strong emphasis and integration of spatial context, a new area of ecological studies, called spatial ecology, has emerged (Gripenberg and Roslin, 2007; Tilman and Kareiva, 2018).

Spatial context is also very important in recognition of objects in images. In a content-based image retrieval experiment, incorporating spatial context models dramatically reduced the misclassification and increased the accuracy of material detection by 13% (Singhal et al., 2003). In order to better recognize or identify defined objects (e.g., cars, rivers, sky) in an image, combining the naturally classified texture or colors as spatial context greatly improved detection accuracy (Heitz and Koller, 2008).

Spatial context plays an important role when measuring the similarity of two entities or event sequences. The effect of context on existing similarity measurement approaches has been reported on in the geospatial domain (Keßler, 2007; Keßler et al.,

2007). This work focuses on quantifying the impact of changing contexts on similarity measures, thus recognizing the potential influence of context on similarity measures embedded in that context. This paper focuses on measures of similarity for spatial settings with the expectation that setting similarity is likely influencing the similarity of event sequences observed within a setting.

In general, spatial-temporal contexts describe the general or broader context in which an event or phenomenon occurs. We distinguish spatial-temporal settings as referring to a specific location and time frame in which an event or phenomenon occurs. There is no fixed or natural scale (Levin, 1992) for such a setting, which may be as broad as a particular region or as specific as a particular location within a region. It can also refer to a specific point in time, or a specific time interval.

In this study, we develop similarity measures between individual spatiotemporal settings and sequences of spatiotemporal settings, which may affect or drive the formation of event sequence patterns. Spatiotemporal settings are characterized by a collection of parametric factors within the environs where events or event sequences are situated, with an emphasis on location, time, and circumstances. We discuss the concepts of classification and scale of spatiotemporal settings, followed by representation and variable selection for assessing spatiotemporal setting similarity. We then develop a matrix-based approach for computing similarity measures between spatiotemporal settings at a certain time point or period and sequences of spatiotemporal settings over serial times, which we evaluate through a case study. The developed similarity measure serves as an index that combines a set of quantitative and qualitative factors. The measures have broad application beyond ecological and environmental event settings, to social, cultural and health related contexts.

4.2. Materials and Methods

4.2.1. Model for Event Sequence Settings

As noted above, a conceptual challenge for modeling an event sequence setting lies in the spatial and temporal specification of the setting. The event sequence similarity measure described by Xu and Beard (Xu and Beard, 2021) assume time series and derived events sequences are observed at fixed point locations. Clearly, influences on a time series, and, by extension, a derived event sequence, extend beyond a point location, but a projected extent will be application and scale dependent, the scale dependence here being a function of pertinent event generating processes. As (Marston et al., 2005) note, the space of a site is something that emerges through unfolding event relations. Thus, we assume that the space of an event setting will vary based on the observed process, local environmental circumstances, and monitoring practices and have scale implications for variable selection.

As with most analyses, spatial and temporal scales must be considered in identifying and characterizing spatiotemporal event sequence settings. As a basis for modeling sequences of spatiotemporal event settings, we first model an event-situated setting at a specific temporal scale or time point with different spatial scales. Figure 4.1 illustrates the potential for different spatial boundaries for a setting. Where a boundary is placed has implications for the set of influencing factors. With changes in spatial scale, the influencing factors for a setting may vary and may be both static and dynamic.

To account for the dynamic aspects of setting as relating to an event sequence at a location, we conceptualize the setting as a sequence, i.e., a sequence of settings at ordered time points, as illustrated in Figure 4.2. The measurement of spatial pattern and heterogeneity is dependent on the scale at which the measurements are made. In this study,

we do not consider interactions between scales. For a specific application context, we assume that we have determined the pertinent set of static and dynamic variables for representing all event settings at one spatial scale. For a set of monitored locations generating spatiotemporal event sequences as discussed in *(Xu and Beard, 2021)*, we specify corresponding sequences of spatiotemporal event settings. Figure 4.2 graphically illustrates these conceptual sequences of spatiotemporal event settings with n dynamic and m static representative variables.

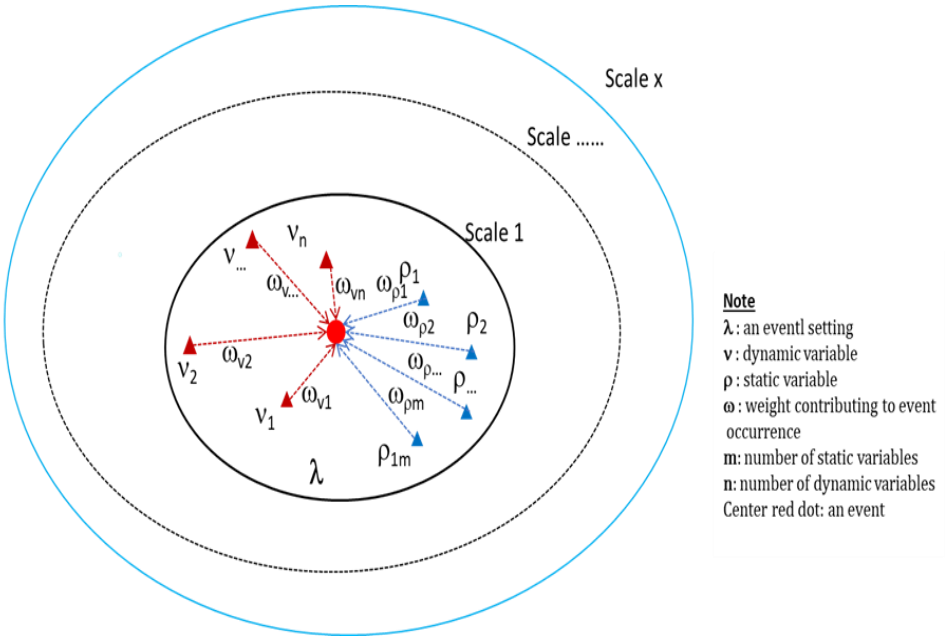


Figure 4.1. Schematic representation of an event-situated setting considering different spatial scales for the setting. Influencing factors with different weights are shown only at Scale 1. More, fewer, or different sets of factors may apply at another scale.

4.2.2. Matrix Representation of Sequences of Spatiotemporal Settings

For a given application context, we assume we have determined the major variables which strongly or satisfactorily represent the spatial settings for a set of sensor locations or

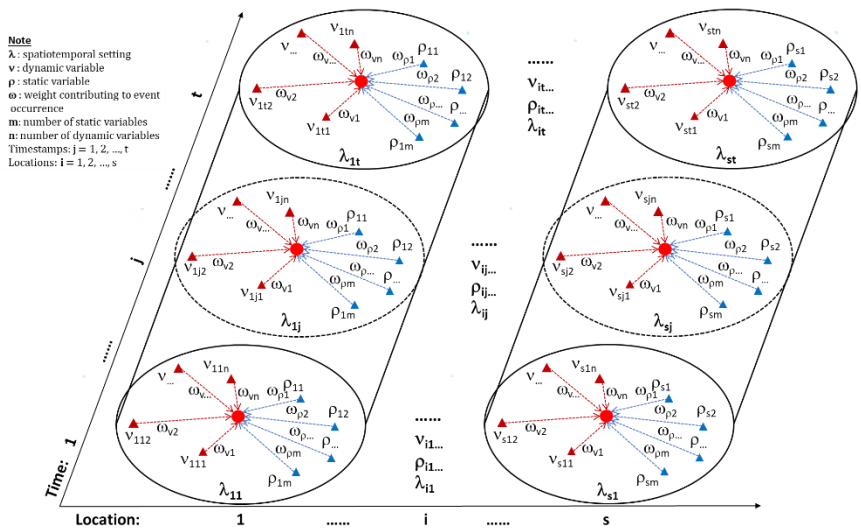


Figure 4.2. Schematic illustration of sequences of spatial-temporal settings with *t* time points and *s* locations.

monitoring stations where event sequences are observed. Given s locations or monitoring stations and t temporal points, we conceptually associate an event sequence with a setting sequence. We then represent these sequences of spatiotemporal event settings with a $s \times t$ matrix, as schematically illustrated in Figure 4.3.

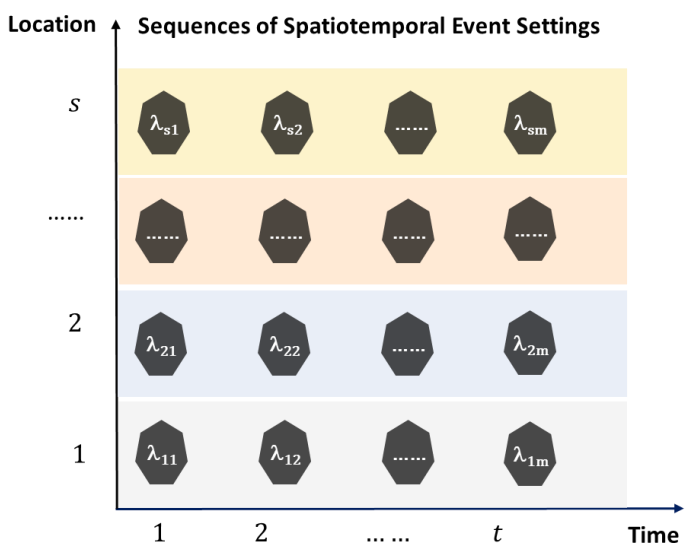


Figure 4.3. Schematic matrix representation of sequences of spatial-temporal settings with *t* time points and *s* locations. λ_{st}—a setting at location *s* and time *t*.

For each setting λ with n dynamic (ν) and m static variables (ρ), i.e.,

$$\lambda: \nu_1, \nu_2, \dots\dots, \ \rho_m,$$

Figure 4.3 can be expanded to Figure 4.4 to become the variables-based matrix representation of the sequences of spatiotemporal event settings.

$$\begin{array}{cc}
 & \textbf{Temporal points} \\
 & \begin{array}{cccc} 1, & 2, & \dots\dots, & t \end{array} \\
\begin{array}{c} \textbf{Spatial locations} \\ \textbf{1} \\ \textbf{2} \\ \textbf{3} \\ \dots \\ \textbf{s} \end{array} &
\begin{pmatrix}
\nu_{111}, \nu_{112}, \dots\dots, \nu_{11n}, \rho_{11}, \rho_{12}, \dots\dots, \rho_{1m}, \nu_{121}, \nu_{122}, \dots\dots, \nu_{12n}, \rho_{11}, \rho_{12}, \dots\dots, \rho_{1m}, \dots\dots, \nu_{1t1}, \nu_{1t2}, \dots\dots, \nu_{1tn}, \rho_{11}, \rho_{12}, \dots\dots, \rho_{1m} \\
\nu_{211}, \nu_{212}, \dots\dots, \nu_{21n}, \rho_{21}, \rho_{22}, \dots\dots, \rho_{2m}, \nu_{221}, \nu_{222}, \dots\dots, \nu_{22n}, \rho_{21}, \rho_{22}, \dots\dots, \rho_{2m}, \dots\dots, \nu_{2t1}, \nu_{2t2}, \dots\dots, \nu_{2tn}, \rho_{21}, \rho_{22}, \dots\dots, \rho_{2m} \\
\nu_{311}, \nu_{312}, \dots\dots, \nu_{31n}, \rho_{31}, \rho_{32}, \dots\dots, \rho_{3m}, \nu_{321}, \nu_{322}, \dots\dots, \nu_{32n}, \rho_{31}, \rho_{32}, \dots\dots, \rho_{3m}, \dots\dots, \nu_{3t1}, \nu_{3t2}, \dots\dots, \nu_{3tn}, \rho_{31}, \rho_{32}, \dots\dots, \rho_{3m} \\
\dots \\
\nu_{s11}, \nu_{s12}, \dots\dots, \nu_{s1n}, \rho_{s1}, \rho_{s2}, \dots\dots, \rho_{sm}, \nu_{s21}, \nu_{s22}, \dots\dots, \nu_{s2n}, \rho_{s1}, \rho_{s2}, \dots\dots, \rho_{sm}, \dots\dots, \nu_{st1}, \nu_{st2}, \dots\dots, \nu_{stn}, \rho_{s1}, \rho_{s2}, \dots\dots, \rho_{sm}
\end{pmatrix}
\end{array}$$

Figure 4.4. Matrix representation of sequences of spatiotemporal event settings with *s* locations and *t* time points.

4.2.3. Similarity Measures of Spatial Settings

4.2.3.1. Pairwise Similarity between Individual Spatial Settings

Pairwise similarity between individual settings is fundamental to further develop similarity measures between sequences of spatiotemporal settings based on certain criteria. In a study of environmental settings, for example, pairwise similarity can be used to measure the similarity between two or more settings based on factors such as temperature, humidity, rainfall, and other environmental variables. By calculating pairwise similarity scores, we can gain insights into how different or how similar settings relate to each other and identify patterns that may be useful in predicting future outcomes.

In this study, we develop a new pairwise similarity measure between spatial settings based on the modifications of the Jaccard index described in (Xu and Beard, 2021). The original Jaccard index is a similarity measure commonly used in the context of sets or binary vectors, where each element can either be present or absent (Choi et al., 2010). To

adapt the Jaccard index for measuring the similarity between spatial settings associated with thematic events, we need to determine a set of common features, including static and dynamic variables, representing each spatial setting. Considering the number of common features for a pair of settings, we have two major considerations, (1) the magnitude or quantitative level of each element from both settings, and (2) that the values of the dynamic variables or elements should be measured at the same timestamps or time intervals.

We first identify the co-existing dynamic variables between two representative dynamic variable sets l_{d1} and l_{d2}, and the co-existing static variables between two representative static variable sets l_{s1} and l_{s2} of two spatial settings, setting 1 and 2. We calculate the relative ratios of individual common variables, and then sum them by dynamic and static variables. The modified Jaccard similarity between two spatial settings at time *k* can be expressed as the sum of relative ratios of all common features/variables divided by the total number of unique features/variables in both sets/settings as in Equation (1):

$$sim_k(l_1, l_2) = \frac{Sd_{k12} + Ss_{12}}{|l_{d1} \cup \ l_{d2}| + |l_{s1} \cup \ l_{s2}|} = \frac{Sd_{k12} + Ss_{12}}{N_d + N_s} \tag{12}$$

where

l_1-set 1 representing spatial setting 1, including the subset 1 of dynamic variables (l_{d1}) and the subset 2 of static variables (l_{s1}),

l_2-set 2 representing spatial setting 2, including the subset 1 of dynamic variables (l_{d2}) and the subset 2 of static variables (l_{s1}),

Sd_{k12}—sum of relative ratios of common dynamic variables between two settings at time *k*,

Ss_{12}—sum of relative ratios of common dynamic variables between two settings, assuming no changes over time during the experiment,

$N_d = |l_{d1} \cup l_{d2}|$—cardinality of union set of l_{d1} and l_{d2},

$N_s = |l_{s1} \cup l_{s2}|$—cardinality of union set of l_{s1} and l_{s2}.

We have two similarity calculation situations dependent on variable types. First, if variable values are interval, ratio, binary and categorical, the pairwise similarity at time *k* can be calculated using Equations (2) and (3). Note that the categorical data can be converted to binary data format based on the number of categories.

If not considering weights or relative importance of individual elements/variables:

$$sim_k(l_1, l_2) = \frac{\sum_{i=1}^{Ckd12}\left(1-Abs\left(lev(d_{k1i})-lev(d_{k2i})\right)\right)+\sum_{j=1}^{Cs12}\left(1-Abs\left(lev(s_{1j})-lev(s_{2j})\right)\right)}{N_d+N_s} \quad (13)$$

If considering weights or relative importance of individual elements/variables:

$$sim_k(l_1, l_2) = \frac{c_{kd12}\sum_{i=1}^{Ckd12}\omega_i\left(1-Abs\left(lev(d_{k1i})-lev(d_{k2i})\right)\right)}{N_d} + \frac{c_{s12}\sum_{j=1}^{Cs12}\omega_j\left(1-Abs\left(lev(s_{1j})-lev(s_{2j})\right)\right)}{N_s} \quad (14)$$

Second, if variables are ordinal valued, the similarity can be calculated using Equations (4) and (5):

If not considering weights or relative importance of individual elements/variables:

$$sim_k(l_1, l_2) = \frac{\sum_{i=1}^{Ckd12}\left(1-\frac{Abs\left(lev(d_{k1i})-lev(d_{k2i})\right)}{n_i-1}\right)+\sum_{j=1}^{Cs12}\left(1-\frac{Abs\left(lev(s_{1i})-lev(s_{2i})\right)}{m_j-1}\right)}{N_d+N_s} \quad (15)$$

If considering weights or relative importance of individual elements/variables:

$$sim_k(l_1, l_2) = \frac{c_{kd12} \sum_{i=1}^{Ckd12} \omega_i \left(1 - \frac{Abs\left(lev(d_{k1i}) - lev(d_{k2i})\right)}{n_i - 1}\right)}{N_d} + \frac{c_{s12} \sum_{j=1}^{Cs12} \omega_j \left(1 - \frac{Abs\left(lev(s_{1i}) - lev(s_{2j})\right)}{m_j - 1}\right)}{N_s} \quad (16)$$

where

c_{kd12}—the number of common dynamic variables between two settings at timestamp k,

c_{s12} – the number of common static variables between two settings,

ω_i, ω_j—weights or relative importance of dynamic and static independent variables to response variable,

n_i, m_j—ordinal levels of dynamic variable *i* and static variable *j*, respectively,

$lev(d_{k1i}), lev(d_{k2i})$—the relative levels or magnitudes of two corresponding co-occurring elements in two dynamic subsets l_{d1} and l_{d2} at timestamp k, respectively:

$$lev(d_{k1i}) = \frac{d_{k1i}}{d_{k1i}+d_{k2i}} \text{ and } lev(d_{k2i}) = \frac{d_{k2i}}{d_{k1i}+d_{k2i}} \quad (17)$$

ω_i, ω_j— weights or relative importance of dynamic and static independent variables to response variable,

$lev(s_{1i}), lev(s_{2i})$— the relative levels or magnitudes of two corresponding co-occurring elements in two static subsets l_{s1} and l_{s2}, respectively:

$$lev(s_{1i}) = \frac{s_{1i}}{s_{1i}+s_{2i}} \text{ and } lev(s_{2i}) = \frac{s_{2i}}{s_{1i}+s_{2i}} \quad ($$

2.3.2. Pairwise similarity between sequences of spatial settings

Sequences of a spatial setting refer to the different configurations of the setting or a physical space that occur over time due to the changes of the dynamic variables while

static variables are assumed stable during the study timeframe of interest. We can extend the modified Jaccard index-like pairwise similarity measure between individual settings, to calculate the pairwise similarity between sequences of spatial settings if the data from different locations are collected in equal time intervals or in the same order. Assuming we have determined the granularity of time intervals or certain sequential order and the total number of timestamps, T, the similarity between two sequences of spatial settings from two locations (S1 and S2) can be expressed as Equation (8):

$$\begin{aligned} sim_{global}(S_1, S_2) &= \frac{\sum_{k=1}^{T} sim_k(l_1, l_2)}{T} \\ &= \frac{\sum_{k=1}^{T}(Sd_{k12} + Ss_{12})}{T(N_d + N_s)} \\ &= \frac{\sum_{k=1}^{T} Sd_{k12}}{T(N_d + N_s)} + \frac{\sum_{i=1}^{T} Ss_{12}}{T(N_d + N_s)} \\ &= \frac{\sum_{k=1}^{T} Sd_{k12}}{T(N_d + N_s)} + \frac{Ss_{12}}{N_d + N_s} \end{aligned} \tag{19}$$

In dealing with the sequences of spatial settings, we also need to consider the data types and the weights or relative importance of explanatory variables to response variables (events or event sequences of interests). So, we also have four situations in which calculating the similarity between these setting sequences from different locations.

1) Variable type: interval, ratio, binary and categorical; not considering the weights of individual variables:

$$\begin{aligned} & sim_{global}(S_1, S_2) \\ & = \frac{\sum_{k=1}^{T} \sum_{i=1}^{Ckd12} \left(1 - Abs\left(lev(ld_{k1i}) - lev(ld_{k2i})\right)\right)}{T(N_d + N_s)} \\ & + \frac{\sum_{j=1}^{Cs12} \left(1 - Abs\left(lev(ls_{1j}) - lev(ls_{2j})\right)\right)}{N_d + N_s} \end{aligned} \tag{20}$$

2) Variable type: interval, ratio, binary and categorical; considering the weights of individual variables:

$$sim_{global}(S_1, S_2) = \frac{c_{kd12} \sum_{k=1}^{T} \sum_{i=1}^{Ckd12} \omega_i \left(1 - Abs\left(lev(ld_{k1i}) - lev(ld_{k2i})\right)\right)}{N_d} + \frac{T * c_{s12} \sum_{j=1}^{Cs12} \omega_j \left(1 - Abs\left(lev(s_{1j}) - lev(s_{2j})\right)\right)}{N_s} \tag{21}$$

3) Variable type: ordinal; not considering the weights of individual variables:

$$sim_{global}(S_1, S_2) = \frac{\sum_{k=1}^{T} \sum_{i=1}^{Ckd12} \left(1 - \frac{Abs\left(lev(d_{k1i}) - lev(d_{k2i})\right)}{n_i - 1}\right)}{T(N_d + N_s)} + \frac{\sum_{j=1}^{Cs12} \left(1 - \frac{Abs\left(lev(s_{1i}) - lev(s_{2i})\right)}{m_j - 1}\right)}{N_d + N_s} \tag{22}$$

4) Variable type: ordinal; considering the weights of individual variables:

$$sim_{global}(S_1, S_2) = \frac{c_{kd12} \sum_{k=1}^{T} \sum_{i=1}^{Ckd12} \omega_i \left(1 - \frac{Abs\left(lev(d_{k1i}) - lev(d_{k2i})\right)}{n_i - 1}\right)}{N_d} + \frac{T * c_{s12} \sum_{j=1}^{Cs12} \omega_j \left(1 - \frac{Abs\left(lev(s_{1i}) - lev(s_{2i})\right)}{m_j - 1}\right)}{N_s} \tag{23}$$

4.2.4. Setting Similarity Analysis Workflow

To estimate similarity levels between event settings, a critical step is to effectively select and quantify the major attributes representing these settings where events or event sequences occur. As introduced above, the variables can be static, dynamic, or both, potentially covering a wide range of environmental variables. The selection of variables in developing similarity measures will be domain dependent and should be statistically discriminant. In a water quality monitoring application, for example, the static spatial setting variables of interest could include land cover, topography, and soils, and dynamic variables could be weather related. Figure 4.5 shows the steps for implementing similarity assessment between event settings or sequences of event settings in a specific domain.

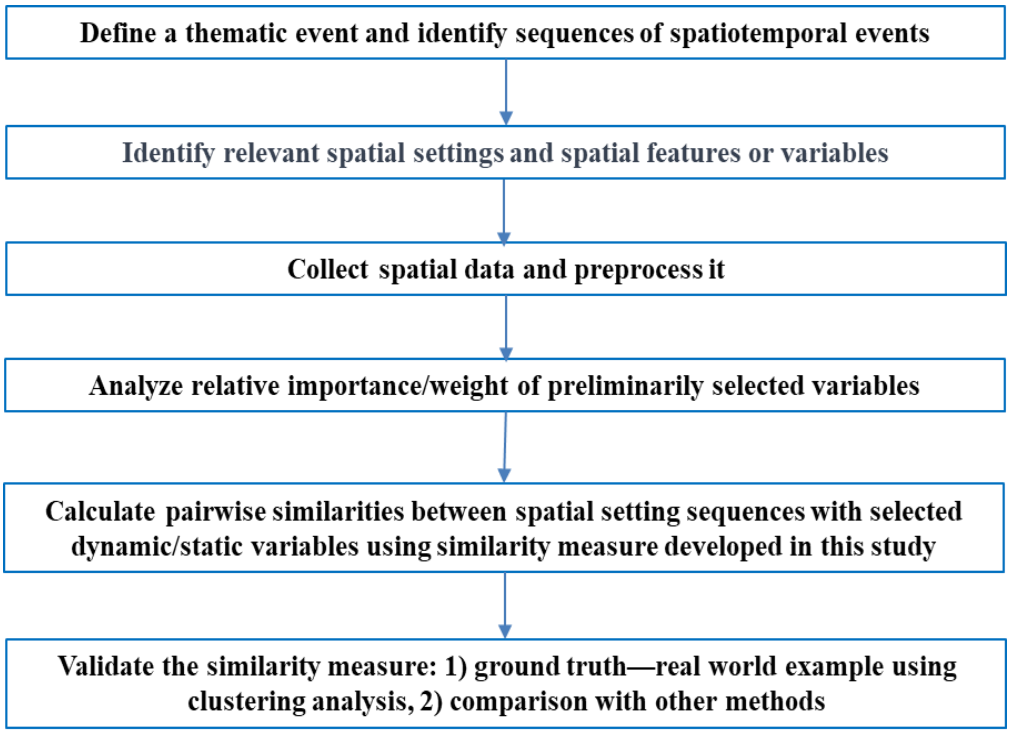


Figure 4.5. Spatial-temporal setting similarity analysis flowchart.

Define a thematic event and identify sequences of spatiotemporal events: assume that we focused on an event or event sequences related research in a specific domain and identified a series of sequences of spatiotemporal events and completed similarity analysis between these sequences.

Identify relevant spatial settings and spatial features or variables: select potential dynamic and static variables representing spatial settings, which are deemed relevant to event occurrences based on domain knowledge. In studying air pollution events, for example, we could include data on such variables as wind direction, wind speed, sites of local manufactures, major pollution sources, concentration of major pollutants, and

transportation density. A correlation matrix for these initial selected variables can be used to eliminate redundant information.

Collect spatial data and preprocess it: collect sufficient data on pre-selected static and dynamic variables intuitively correlated to occurrences of thematic events. Preprocessing or preparation of the collected data mainly includes normal distribution check, normalization, standardization of measurement units, and binarization of categorical data.

Analyze relative importance/weight of preliminarily selected variables: to improve the computation speed and accurate representation of similarity measures we should identify those variables most relevant to the events of interest and reduce the number. To determine which variables are most important to the thematic events and for the similarity measures, we can conduct relative weight analysis (RWA) (Chao et al., 2008; Tonidandel and LeBreton, 2015; Tonidandel et al., 2009) and partial least squares regression (PLSR) (Ali et al., 2018).

Calculate pairwise similarities between spatial setting sequences: Once the most relevant features or variables are identified, we can use the similarity measures developed in this study to compute the pairwise similarity between spatial settings and sequences of spatial-temporal settings and form the similarity matrix.

Validate the similarity measure: with the similarity matrix of spatial setting sequences, we can further conduct clustering analysis to group event sequences associated with locations or stations, and then conduct the comparison analysis with clusters of event sequences as ground truth. The other approach is to compare the results with other methods.

4.3. Case study: Setting Similarity of Coastal Monitoring Stations for Fecal Pollution

To demonstrate the use of our proposed method above, we determined the pairwise similarities of 16 monitoring stations along the Maine coast with the selected setting attributes for costal fecal pollution event sequences. The Maine Department of Marine Resources (DMR) manages the shellfish growing areas in coastal Maine based on the fecal pollution situations observed from more than 2000 monitoring stations. Fecal coliform is a type of bacteria that is found in the intestines and feces of warm-blooded animals, including humans. It is used as an indicator of fecal contamination of water (Noble et al., 2003). Monitoring fecal coliform levels in coastal waters is important because it can help identify sources of contamination and provide an early warning of contamination, enabling faster responses. Maine DMR typically collects water samples at these monitoring stations (>2000) at regular intervals and analyzes them for fecal coliform levels. Grouping monitoring stations as similar spatial settings of fecal pollution events can provide several benefits and advantages. First, it can provide useful information for early detection of pollution events at similar stations (Dong et al., 2015; Prasad et al., 2015). Second, cluster analysis of monitoring stations across a wider area can help to identify trends and patterns in fecal coliform levels and pollution events, which can inform efforts to improve water quality. Third, followed by the previous two benefits, it will help to optimize resource allocation and prioritize monitoring efforts based on areas of higher pollution risk, which can help to reduce costs and increase efficiency in monitoring and management activities. Fourth, it can help to make more informed decisions about pollution control measures, such as beach closures or water treatment. Lastly, it also helps to increase public awareness of

coastal water quality issues and the need for responsible use and management of marine resources.

4.3.1. Experimental Site and Design

4.3.1.1. Site and Variables

In this case study, we selected 16 monitoring stations along the Maine coast, with the following DMR assigned location IDs: WE020.00, WE028.00, WG008.10, WG027.00, WG038.00, WM003.00, WN057.00, WN077.20, WQ023.00, WR011.00, WS027.00, WS051.00, WT015.00, WT018.00, WT024.00, WV019.00, as shown on the map (Figure 4.6). Multiple factors related to fecal coliform concentration around these monitoring stations contribute to characterizing the corresponding spatial settings for fecal pollution events. Some studies have shown that shoreline, basin hydrology, and marine environment affect the retention, survival, and distribution of fecal coliform (Hughes, 2003). Based on data availability, we selected a combination of basin characteristics as static variables and some marine environmental factors as dynamic variables. Their abbreviations and description are shown in Table 4.1.

4.3.1.2. Data Collection

We used the geolocations of the 16 selected monitoring stations to delineate the corresponding basins with StreamStats v4.13.0 (https://streamstats.usgs.gov/ss/, the access date: 25 February 2023) and download all associated basin characteristics data. For the static variables described in Table 4.1, the data were extracted as shown in Table B.1. We obtained marine environment related variables and fecal coliform measurements from Maine DMR (Table B.2)).

4.3.1.3. Methods

We used partial least squares regression (PLSR) analysis (Tonidandel and LeBreton, 2015; Tonidandel et al., 2009) to obtain the relative importance of all variables against the fecal coliform scores. We used the similarity measure developed in this study to achieve the similarity matrix of spatial setting sequences, and used the method developed in (Xu and Beard, 2021) to obtain the similarity matrix of the corresponding fecal pollution event sequences with the same locked timestamps. After converting the similarity matrices of both setting and event sequences to the distance matrices, we performed a cluster analysis (Kettenring, 2006).

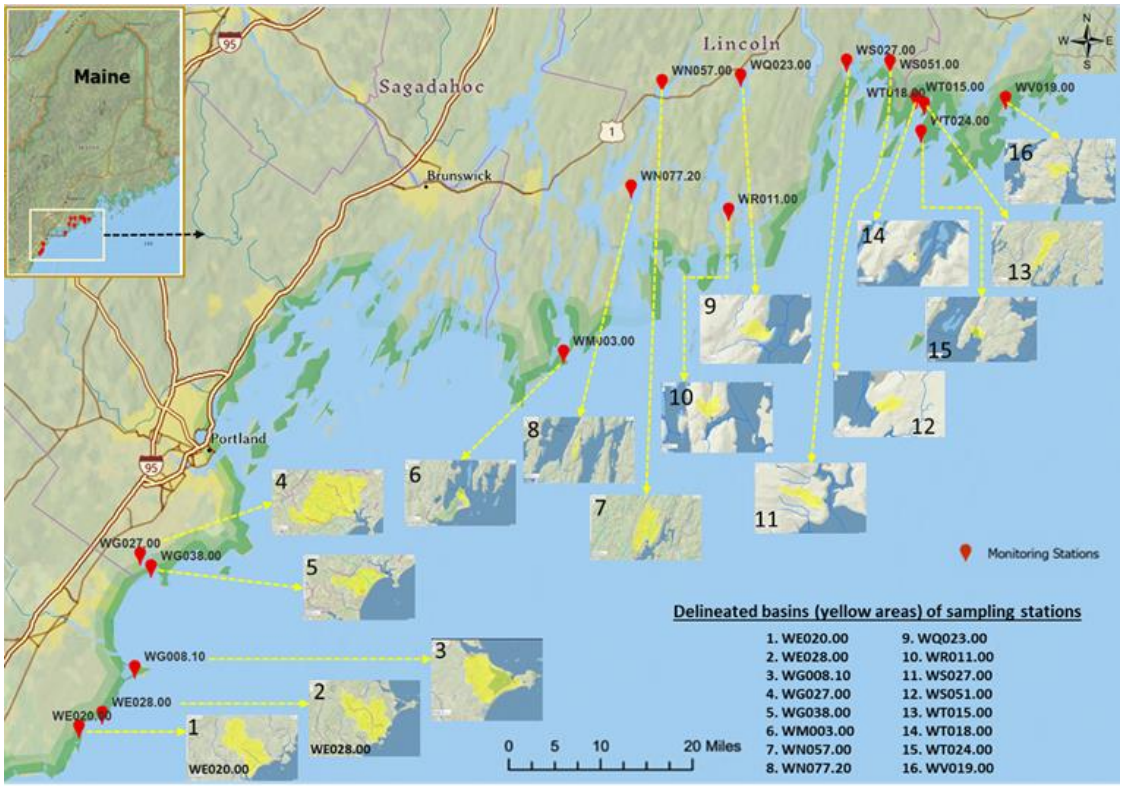


Figure 4.6. Selected monitoring stations/locations on the Maine coast for depicting spatiotemporal settings of fecal pollution event sequences.

Table 4.1. Description and abbreviation of selected basin characteristics and dynamic parameters

Abbreviation/code	Description	Unit
Static Variables	**(Basin Characteristics)**	
BSLDEM10M	Mean basin slope computed from 10 m DEM	percent
COASTDIST	Shortest distance from the coastline to the basin centroid	miles
DRNAREA	Area that drains to a point on a stream	square miles
ELEV	Mean Basin Elevation	feet
ELEVMAX	Maximum basin elevation	feet
LC11DEV	Percentage of developed (urban) land from NLCD 2011 classes 21-24	percent
LC11IMP	Average percentage of impervious area determined from NLCD 2011 impervious dataset	percent
PCTSNDGRV	Percentage of land surface underlain by sand and gravel deposits	percent
SANDGRAVAF	Fraction of land surface underlain by sand and gravel aquifers	dimensionless
SANDGRAVAP	Percentage of land surface underlain by sand and gravel aquifers	percent
STATSGOA	Percentage of area of Hydrologic Soil Type A from STATSGO	percent
STORAGE	Percentage of area of storage (lakes ponds reservoirs wetlands)	percent
STORNWI	Percentage of storage (combined water bodies and wetlands) from the National Wetlands Inventory	percent
BKSF	Bank-full Streamflow	ft^3/s
BKW	Bank-full Width	ft
BKD	Bank-full Depth	ft
BKA	Bank-full Area	ft^2
Pop_Dnsity	Population Density	persons/mi^2
Dynamic Variables		
Tide	Tide stages: H, L, F, E, HF, HE, LF, LE	3 Hours
Salinity	Ocean water salinity	
Wind	Wind direction: E, S, W, N, NW, NE, SW, SE	Direction
RainCum24	Cumulative precipitation in 24 hours	inch
RainCum48	Cumulative precipitation in 48 hours	inch
RainCum72	Cumulative precipitation in 72 hours	inch
RainCum96	Cumulative precipitation in 96 hours	inch

Table 4.2. Relative weights of 39 selected variables with signs.

Negative Variables	Relative Importance	Positive Variables	Relative Importance
Salinity	−34.696	COASTDIST	7.252
STATSGOA	−7.763	BKSF	7.217
BKW	−5.725	STORNWI	6.092
ELEV	−3.630	RainCum72	4.256
STORAGE	−1.069	LC11DEV	3.789
Tide.HF.	−0.817	BSLDEM10M	3.200
Wind.NW.	−0.790	Tide.HE.	1.455
BKA	−0.778	RainCum96	1.389
Tide.H.	−0.771	ELEVMAX	1.298
LC11IMP	−0.472	Wind.CL.	1.121
Wind.S.	−0.373	RainCum48	1.082
BKD	−0.372	RainCum24	0.878
Wind.N.	−0.247	DRNAREA	0.871
Tide.E.	−0.218	Wind.NE.	0.654
Pop_Dnsity	−0.186	PCTSNDGRV	0.399
Wind.E.	−0.109	SANDGRAVAP	0.393
Wind.SW.	−0.106	Tide.F.	0.325
Wind.SE.	−0.095	Tide.LE.	0.042
Wind.W.	−0.056	SANDGRAVAF	0.004
Tide.L.	−0.015		

4.3.2. Relative Weights and Selection of Representative Variables for Spatial Settings

The results of the partial least squares regression analysis on 39 variables revealed that some variables are more important than others in predicting the fecal coliform levels

(Table 4.2 and Figure 4.7). The signs associated with each variable provide insight into the direction of their impact on the fecal coliform levels. Salinity has the highest relative importance and the strongest negative influence on the fecal coliform. On the other hand, shortest distance from the coastline to the basin centroid (COASTDIST), bank-full streamflow (BKSF), and percentage of storage (combined water bodies and wetlands) from the National Wetlands Inventory (STORNWI) have the highest positive influence on the fecal coliform levels.

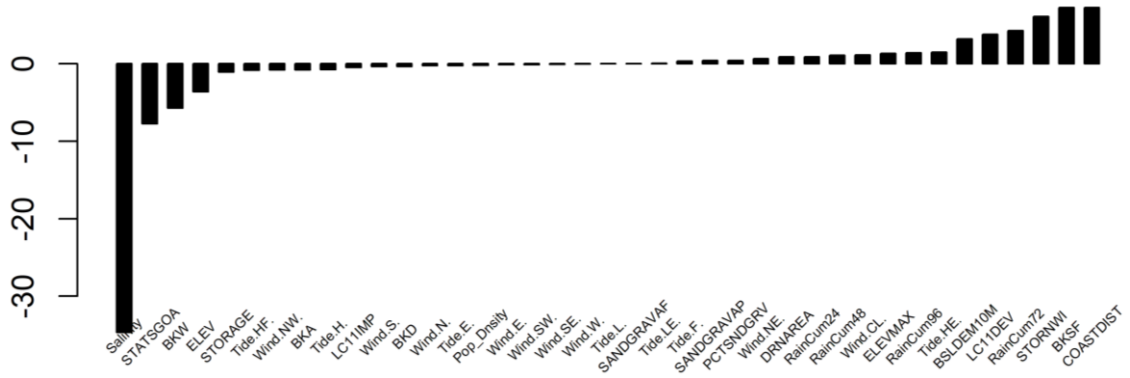


Figure 4.7. Bar chart of relative importance of 39 selected static and dynamic explanatory variables for fecal coliform bacterial measurements.

To reduce the number of variables for calculating similarity in the formula developed in this study, we selected the variables with higher weights. In this case study, we selected those variables with absolute values of relative importance greater than 1. We then re-ran PLSR with these selected variables against corresponding fecal coliform levels. The relative importance of these variables from the second round PLSR is shown in Table 4.3 and Figure 4.8, which can be used as relative weights for calculating similarities between spatial setting sequences when considering contribution from these individual variables.

Table 4.3. Relative weights of 16 selected variables with signs.

Negative Variables	Relative Importance	Positive Variables	Relative Importance
Salinity	−33.900	BKSF	8.500
STATSGOA	−11.500	STORNWI	6.500
ELEV	−8.700	COASTDIST	6.200
BKW	−6.700	RainCum72	4.200
STORAGE	−1.100	BSLDEM10M	3.700
		ELEVMAX	3.000
		Tide.HE.	1.400
		RainCum96	1.400
		LC11DEV	1.100
		Wind.CL.	1.100
		RainCum48	1.100

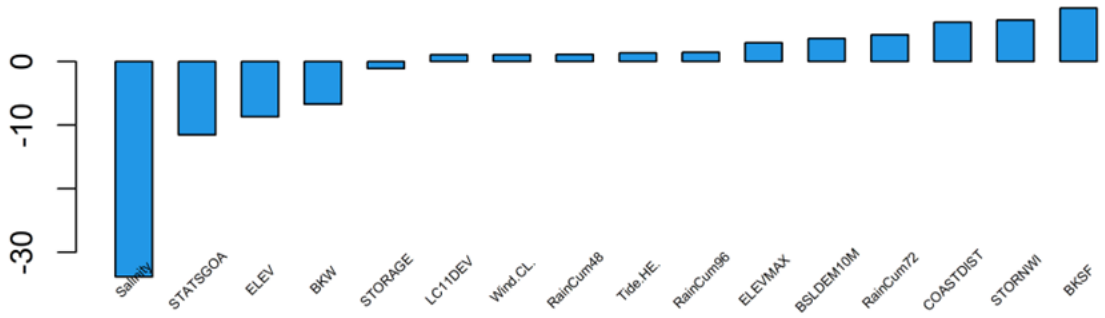


Figure 4.8. Bar chart of relative importance of 16 selected static and dynamic variables against fecal coliform bacterial measurements.

4.3.3. Clustering Analysis of Spatial Setting Sequences and Fecal Pollution Event Sequences

We computed all pairwise similarities between spatial setting sequences using the method of this study using the 16 selected variables in the previous section for 16 rain-

storm-involved timestamps. The clustering analysis of spatial setting sequences labeled with monitoring stations yields interesting insights into the underlying patterns and structures of the data of these selected static and dynamic variables (Figure 4.9). The result indicates that there are 3~4 distinct clusters within the data, with each cluster representing a unique pattern of spatial setting sequences with similar characteristics. Figure 4.9 shows some geographically proximate spatial setting sequences in the same or connected clusters, but not all due to the diverse contributions of different static and dynamic variables. These clusters provide valuable information about the types of spatial setting sequences, which we next relate to clusters of fecal pollution event sequences.

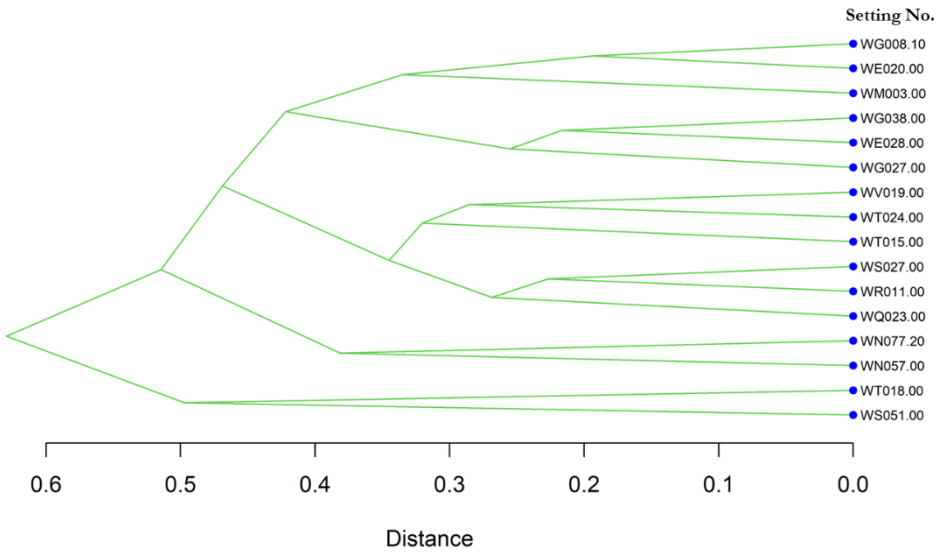


Figure 4.9. Clusters of 16 spatial setting sequences labeled with monitoring stations.

We generated a similarity matrix between fecal pollution event sequences also labeled with monitoring stations and the corresponding setting sequences at the same time frame (16 days). With the conversion to the distance matrix, we implemented the clustering

analysis and the similarity heatmap, and the clustering result is shown in Figure 4.10. Three major clusters are clearly identified.

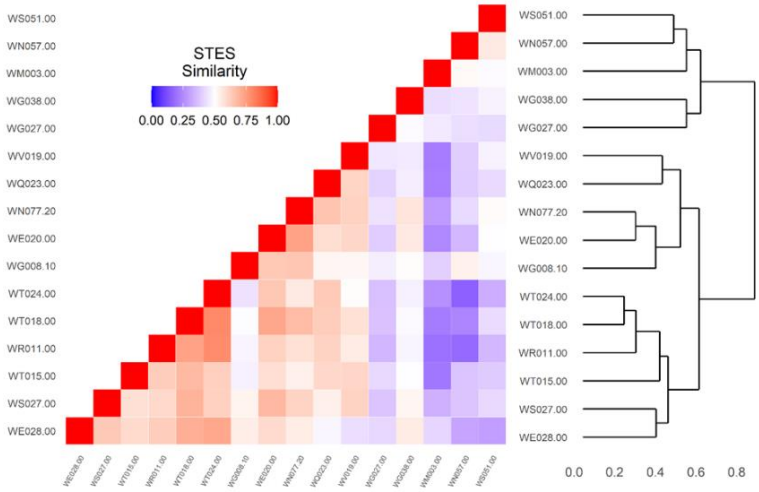


Figure 4.10. Similarity-based heat map and distance based hierarchical clustering between 16 monitoring stations for fecal pollution event sequences.

4.3.4. Cross Analysis between Clusters of Setting Sequences and Clusters of Event Sequences

Cross-analysis between clusters of spatial settings and clusters of events sequences can provide insights into the causes and effects of pollution events in coastal waters. We put the clustering results above from both setting sequences and event sequences side by side to build the cross-comparison and cross mapping graphs (Figure 4.11 and 4.12). By examining components of the major clusters of setting sequences and pollution event sequences, we find cases of at least two stations within one major cluster among the event sequence clusters that were also grouped in the same major cluster of setting sequence clusters. We found 11 out of 16 monitoring stations showing this pattern. Specifically, WS027.00, WT015.00, WT024.00, and WR011.00 in event sequence Cluster E1 are also in setting sequence Cluster S2; WG008.10 and WE020.00 in Cluster E2 are also in Cluster

S1; WQ023.00 and WV019.00 in Cluster E2 are also in Cluster S2; and WG027.00, WG038.00, and WM003.00 in Cluster E3 are also in Cluster S1. This cross-analysis between clusters of spatial settings and event sequences can help to improve our understanding of the complex interactions between environmental factors and basin characteristics and identify drivers for fecal coliform pollution events in coastal marine water.

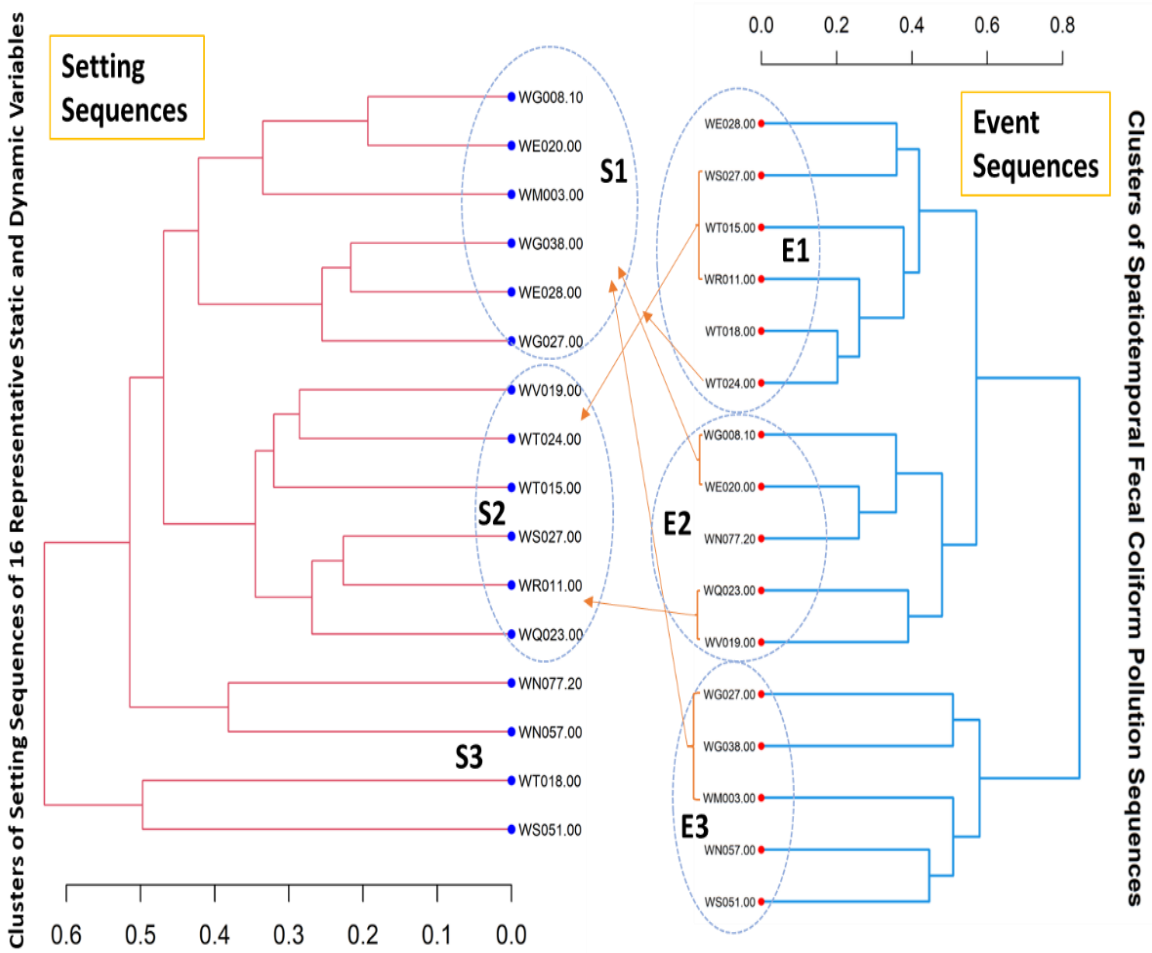


Figure 4.11. Cross analysis between clusters of setting sequences and clusters of event sequences.

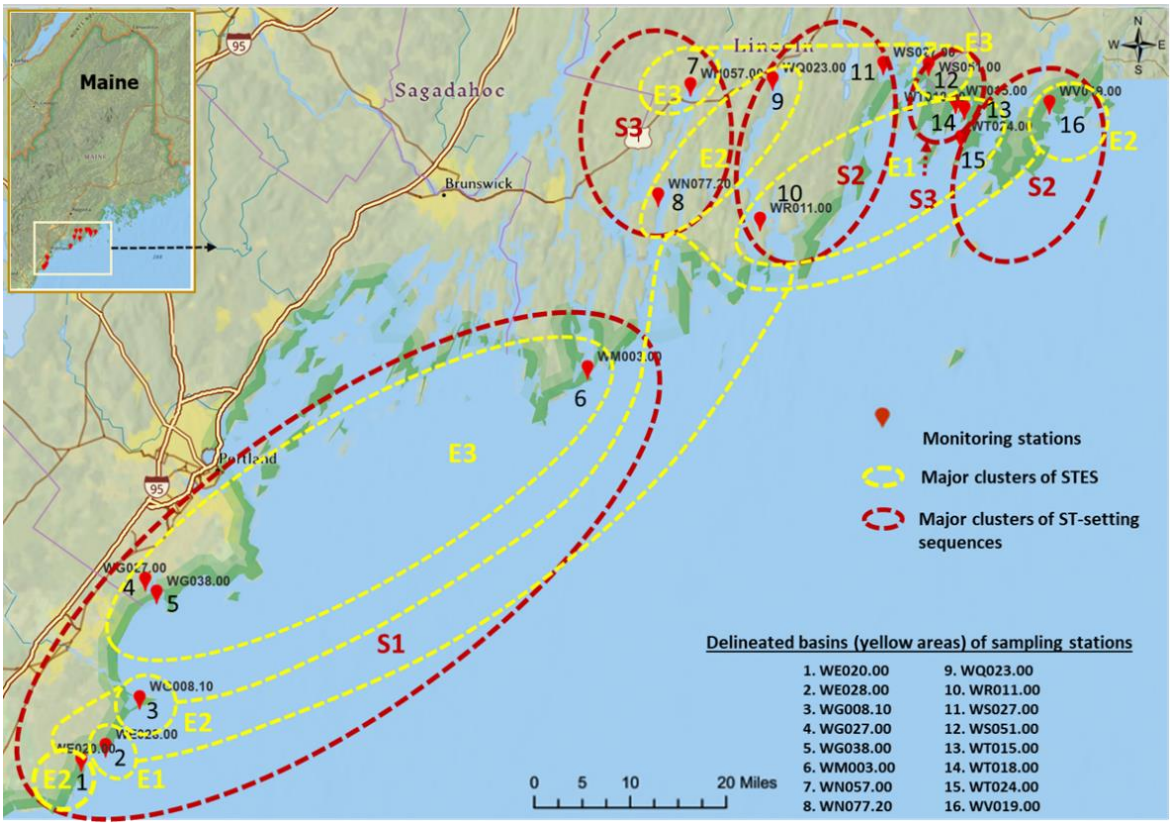


Figure 4.12. Cross mapping between clusters of setting sequences and clusters of event sequences.

4.4. Discussion

We developed similarity measures through modeling spatial setting sequences. The model uses a matrix representation of spatiotemporal event settings and considers both static and dynamic variables. To measure the similarity between spatial settings, the Jaccard index is modified based on the variables' magnitude and the time interval at which dynamic variables are measured. Pairwise similarity between individual spatial settings is crucial for developing similarity measures between sequences of spatiotemporal settings based on specific criteria. The pairwise similarity measure can help to identify patterns and predict future outcomes of corresponding event sequences.

The model's matrix representation of sequences of spatiotemporal settings can be used to represent a set of sensor locations or monitoring stations where event sequences are observed. The matrix representation has the flexibility to include *n* dynamic and *m* static variables that represent all event settings at one spatial scale. The modified Jaccard index measures the similarity between individual spatial settings and forms the basis for similarity measures between sequences of spatiotemporal settings. The modified Jaccard similarity between two spatial setting sequences considers the relative ratios of common features/variables. These measures provide information on the differences or similarity of spatial settings, which in turn contribute to the analysis of event sequences arising from these settings.

Through the case study, we demonstrated how to model the spatial-temporal setting sequences and provide a useful framework for understanding and characterizing spatial setting sequences corresponding to event sequences. The model's focus on defining the bounds of a setting and considering both static and dynamic variables allow for a comprehensive understanding of associated event sequences. The pairwise similarity measure helps identify patterns in event settings or setting sequences to comprehensively understand better the occurrences of events and event sequences. The similarity measures developed in this paper, and the framework incorporating static and dynamic variables to represent settings, will provide useful tools for a range of applications, from environmental settings to predictive modeling.

One potential application of similarity measures for event sequence settings is in the field of disaster management. By analyzing the spatial-temporal settings of past disasters, emergency responders can better predict the likelihood and potential impact of

future disasters and allocate resources more effectively. For example, if a particular region is prone to frequent flooding, similarity measures can be used to identify patterns in the spatial-temporal settings of past floods and help emergency responders anticipate and prepare for future floods in that region. (Castellarin et al., 2001) studied the relationship between hydrological similarity measures and regional flooding frequency. The authors studied the similarity measures between catchments in the distribution of rainfall extremes and the extent of the impervious portion of the catchment. Similarly, our setting sequence similarity measures could be used to compare the frequency distribution of rainfall extremes and the extent of imperviousness across catchments. This could help identify catchments that have similar characteristics in terms of their rainfall and land use and may also be suitable for pooling together in regional flooding frequency analysis. (Kamarinas et al., 2016) found that land cover/land use change (LCLUC) and sediment runoff affected by forestry practices and livestock grazing is temporally related to the water quality. We can potentially extend our similarity measures to this study to compare the temporal patterns of land disturbance and water quality variables (total suspended solids (TSS), turbidity, and visual clarity). This could help identify whether changes in land disturbance are related to changes in water quality over time, and whether there are any spatial relationships between land disturbance and water quality. In addition, as mentioned in this research, there exist nonlinear changes in land disturbance and sediment runoff; our novel approach on similarity between setting sequences can be easily plugged in to study these nonlinear ecosystem dynamics.

Overall, the use of similarity measures for event setting sequences has a wide range of potential applications in various fields, including disaster management, urban planning,

transportation planning, and cultural heritage management. By analyzing the spatiotemporal context of events and their surrounding environmental factors, researchers and practitioners can gain a deeper understanding of the underlying mechanisms that drive those corresponding events and event sequences and use that knowledge to make more informed decisions about the management and planning of future events and activities.

4.5. Conclusions

In conclusion, modeling spatiotemporal event sequences requires careful consideration of spatial and temporal scales to define the bounds of the setting. The dynamic aspects of the setting should also be accounted for by conceptualizing the setting as a sequence. A matrix representation of sequences of spatiotemporal event settings can be developed for each setting with both dynamic and static variables. Pairwise similarity between individual settings and sequences of spatial settings can be calculated based on modifications of the Jaccard index, using a set of spatial features that represent each spatial setting.

With a careful consideration of spatial and temporal scales to define the bounds of the setting, we developed a modeling approach that incorporates dynamic variables or features in addition to static variables. Using a matrix-based representation of spatiotemporal setting sequences, we developed new similarity measures that include quantitative levels of individual elements within the sequence and comparison with locked timestamps or order. These similarity measures allow for the use of all variable data types in the equations. Overall, these similarity measures, along with the matrix-based

representation of spatiotemporal event setting sequences incorporating both static and dynamic variables, provide a novel method in support of event sequence analysis.

Future research could investigate the potential of using the proposed similarity measures to analyze the dynamics of complex systems, such as ecological or economic systems, where events and their settings or contexts can be critical factors in understanding the system behavior. By examining the similarity of event settings, researchers could gain insight into how different factors interact with each other over time and across different spatial scales, which could inform better decision-making in a wide range of fields, from urban planning to disaster management.

Chapter References

Ali, F., Rasoolimanesh, S.M., Sarstedt, M., Ringle, C.M. and Ryu, K. 2018. An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. International Journal of Contemporary Hospitality Management.

Brézillon, P. and Gonzalez, A.J. (2014) Context in computing: a cross-disciplinary approach for modeling the real world, Springer.

Castellarin, A., Burn, D. and Brath, A. 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology 241(3-4), 270-285.

Chao, Y.-C.E., Zhao, Y., Kupper, L.L. and Nylander-French, L.A. 2008. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. Journal of occupational and environmental hygiene 5(8), 519-529.

Choi, S.-S., Cha, S.-H. and Tappert, C.C. 2010. A survey of binary similarity and distance measures. Journal of systemics, cybernetics and informatics 8(1), 43-48.

Cutter, S.L. 1996. Vulnerability to environmental hazards. Progress in human geography 20(4), 529-539.

Delbosc, A. and Currie, G. 2011. The spatial context of transport disadvantage, social exclusion and well-being. Journal of Transport Geography 19(6), 1130-1137.

Dey, A.K. 2001. Understanding and using context. Personal and ubiquitous computing 5(1), 4-7.

Dong, J., Wang, G., Yan, H., Xu, J. and Zhang, X. 2015. A survey of smart water quality monitoring system. Environmental Science and Pollution Research 22, 4893-4906.

Gong, H. and Hassink, R. 2020. Context sensitivity and economic-geographic (re) theorising. Cambridge Journal of Regions, Economy and Society 13(3), 475-490.

Gripenberg, S. and Roslin, T. 2007. Up or down in space? Uniting the bottom-up versus top-down paradigm and spatial ecology. Oikos 116(2), 181-188.

Guralnik, V. and Srivastava, J. 1999 Event detection from time series data, pp. 33-42, ACM.

Heitz, G. and Koller, D. 2008 Learning spatial context: Using stuff to find things, pp. 30-43, Springer.

Hughes, K.A. 2003. Influence of seasonal environmental variables on the distribution of presumptive fecal coliforms around an Antarctic research station. Applied and Environmental Microbiology 69(8), 4884-4891.

Jiang, B. and Yao, X. (2007) Location based services and telecartography, pp. 27-45, Springer.

Kamarinas, I., Julian, J.P., Hughes, A.O., Owsley, B.C. and De Beurs, K.M. 2016. Nonlinear changes in land cover and sediment runoff in a New Zealand catchment dominated by plantation forestry and livestock grazing. Water 8(10), 436.

Keßler, C. 2007 Similarity measurement in context, pp. 277-290, Springer.

Keßler, C., Raubal, M. and Janowicz, K. 2007 The effect of context on semantic similarity measurement, pp. 1274-1284, Springer.

Kettenring, J.R. 2006. The practice of cluster analysis. Journal of classification 23(1), 3-30.

Levin, S.A. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology 73(6), 1943-1967.

Loke, S. (2006) Context-aware pervasive systems: architectures for a new breed of applications, CRC Press.

Lupiani, E., Sauer, C., Roth-Berghofer, T., Juarez, J.M. and Palma, J. 2013 Implementation of similarity measures for event sequences in myCBR.

Malpas, J. 2012. Putting space in place: Philosophical topography and relational geography. Environment and planning D: society and space 30(2), 226-242.

Mannila, H. and Ronkainen, P. 1997 Similarity of event sequences, pp. 136-139, IEEE.

Marston, S.A., Jones III, J.P. and Woodward, K. 2005. Human geography without scale. Transactions of the institute of British geographers 30(4), 416-432.

Moen, P. 2000. Attribute, event sequence, and event type similarity notions for data mining. PhD thesis, University of Helsinki.

Moore, A. 2008. Rethinking scale as a geographical category: from analysis to practice. Progress in human geography 32(2), 203-225.

Noble, R.T., Moore, D.F., Leecaster, M.K., McGee, C.D. and Weisberg, S.B. 2003. Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing. Water research 37(7), 1637-1643.

Obweger, H., Suntinger, M., Schiefer, J. and Raidl, G. 2010 Similarity searching in sequences of complex events, pp. 631-640, IEEE.

Paasi, A. 2004. Place and region: looking through the prism of scale. Progress in human geography 28(4), 536-546.

Prasad, A., Mamun, K.A., Islam, F. and Haqva, H. 2015 Smart water quality monitoring system, pp. 1-6, IEEE.

Roux, A.V.D. and Mair, C. 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186(1), 125-145.

Sampson, R.J. 2003. The neighborhood context of well-being. Perspectives in biology and medicine 46(3), S53-S64.

Schatzki, T.R. (2002) The site of the social: A philosophical account of the constitution of social life and change, Penn State University Press.

Simandan, D. 2019. Revisiting positionality and the thesis of situated knowledge. Dialogues in human geography 9(2), 129-149.

Simandan, D. 2020. Being surprised and surprising ourselves: a geography of personal and social change. Progress in Human Geography 44(1), 99-118.

Singhal, A., Luo, J. and Zhu, W. 2003 Probabilistic spatial context models for scene content understanding, pp. I-I, IEEE.

Sunley, P. 1996. Context in economic geography: the relevance of pragmatism. Progress in Human Geography 20(3), 338-355.

Tilman, D. and Kareiva, P. (2018) Spatial ecology: the role of space in population dynamics and interspecific interactions (MPB-30), Princeton University Press.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A.S., Kurose, S. and Zandee, R. 2003. Spatial context and the complexity of daily travel patterns: an international comparison. Journal of Transport Geography 11(1), 37-46.

Tonidandel, S. and LeBreton, J.M. 2015. RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. Journal of Business and Psychology 30, 207-216.

Tonidandel, S., LeBreton, J.M. and Johnson, J.W. 2009. Determining the statistical significance of relative weights. Psychological methods 14(4), 387.

Weber, J. and Kwan, M.-P. 2003. Evaluating the effects of geographic contexts on individual accessibility: a multilevel Approach1. Urban Geography 24(8), 647-671.

Wolpert, J. 1964. The decision process in spatial context. Annals of the Association of American Geographers 54(4), 537-558.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. and Shneiderman, B. 2012. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with computers 24(2), 55-68.

Woodward, K., Jones III, J.P. and Marston, S.A. 2012. The politics of autonomous space. Progress in Human Geography 36(2), 204-224.

Worboys, M. and Hornsby, K. 2004 From objects to events: GEM, the geospatial event model, pp. 327-343, Springer.

Xu, F. and Beard, K. 2021. A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. ISPRS International Journal of Geo-Information 10(9), 594.

Yang, D.-H., Goerge, R. and Mullner, R. 2006. Comparing GIS-based methods of measuring spatial accessibility to health services. Journal of medical systems 30(1), 23-32.

Zolnik, E.J. 2009. Context in human geography: a multilevel approach to study human–environment interactions. The Professional Geographer 61(3), 336-349.

CHAPTER 5

SCALABILITY AND EXTENDED APPLICATION OF SPATIOTEMPORAL EVENT SEQUENCE SIMILARITY MEASURES: COMPARISON WITH PROSPECTIVE SPACE-TIME SCAN STATISTICS IN CLUSTERING

The content in this chapter is the reformatted version of the published research paper:

Xu, F. and Beard, K. 2021. A comparison of prospective space-time scan statistics and spatiotemporal event sequence based clustering for COVID-19 surveillance. PloS One 16(6), e0252990. https://doi.org/10.1371/journal.pone.0252990

Chapter Abstract

The outbreak of the COVID-19 disease was first reported in Wuhan, China, in December 2019. Cases in the United States began appearing in late January. On March 11, the World Health Organization (WHO) declared a pandemic. By mid-March COVID-19 cases were spreading across the US with several hotspots appearing by April. Health officials point to the importance of surveillance of COVID-19 to better inform decision makers at various levels and efficiently manage distribution of human and technical resources to areas of need. The prospective space-time scan statistic has been used to help identify emerging COVID-19 disease clusters, but results from this approach can encounter strategic limitations imposed by constraints of the scanning window. This paper presents a different approach to COVID-19 surveillance based on a spatiotemporal event sequence (STES) similarity. In this STES based approach, adapted for this pandemic context we compute the similarity of evolving daily COVID-19 incidence rates by county and then cluster these sequences to identify counties with similarly trending COVID-19 case loads. We analyze four study periods and compare the sequence similarity-based clusters to

prospective space-time scan statistic-based clusters. The sequence similarity-based clusters provide an alternate surveillance perspective by identifying locations that may not be spatially proximate but share a similar disease progression pattern. Results of the two approaches taken together can aid in tracking the progression of the pandemic to aid local or regional public health responses and policy actions taken to control or moderate the disease spread.

5.1. Introduction

The first reported case of Coronavirus disease 2019 (COVID-19) appeared in the US in Washington State in January 2020. Cases then began to appear around the country, creating an outbreak more severe than that experienced in the city of Wuhan, China, where the initial outbreak occurred (Huang et al., 2020), as well as in many European countries (Danon et al.; Saglietto et al., 2020). By mid-March 2020 the outbreak had spread to many states and by late April over one million confirmed cases had been reported in the US.

To anticipate and detect outbreaks, the World Health Organization (WHO), many national and local health departments, academic or other non-profit organizations continuously collected information about occurrences of COVID-19. Incidence cases were cumulatively added to different online repositories (Alamo et al.; Latif et al., 2020; Moorthy et al., 2020). Quick detection of emerging geographical clusters or space-time clusters of COVID-19 can aid public health agencies in prioritizing spatial locations for allocation of different kinds of medical resources including testing kits and applying efficient and publicly acceptable interventions. Versions of space-time scan statistics have

been widely used to identify significant clusters of various diseases (Khan et al., 2017; Kulldorff, 1997; 1999; 2001; Kulldorff et al., 2005a) as well as in the current COVID-19 crisis (Desjardins et al., 2020; Qi et al., 2020). Space-time scan statistics use circular or elliptical scanning windows of a series of sizes in combination with varying time intervals to systematically scan a study area to detect clusters of disease cases. The Poisson based space-time scan statistic evaluates each scan window for numbers of cases and tests for locations exceeding the number of expected cases under a Poisson distribution.

The prospective Poisson space-time scan statistic has been successfully used for space-time surveillance of different epidemic diseases. As Kulldorff et al. proposed (Kulldorff, 2001; Kulldorff et al., 2005a), this method focuses on detecting emerging clusters that start at any time during the study period and remain identifiable at the current time (i.e., active or alive), which is the major difference compared to the retrospective space-time scan statistic. Jones et al. used this method to detect twelve "live" or emerging statistically significant (p-value ≤ 0.05) clusters of shigellosis in the city of Chicago (Jones et al., 2006), the results of which helped local health departments to prioritize the assignment and investigation of shigellosis cases. The prospective Poisson space-time scan statistic has also been utilized to identify emerging clusters in other diseases such as thyroid cancer among men in New Mexico (1973-1992) (Kulldorff, 2001), syndromic surveillance (Yih et al., 2010), measles (Yin et al., 2007), and dengue fever (Duczmal et al., 2011). More recently, it has been used to detect "active" clusters of COVID-19 confirmed cases in the United States (Desjardins et al., 2020; Hohl et al., 2020).

While the prospective space-time scan statistic is a good option for detecting emerging space-time clusters of infectious diseases, there remain some limitations. The

effectiveness of the circular scan window decreases as the shape of emerging clusters becomes more irregular. Detected clusters may contain locations without confirmed cases or with low relative risk due to the artifact of the scanning process (Desjardins et al., 2020; Kulldorff et al., 2005a; Li et al., 2019), although this limitation can be minimized by reporting the individual relative risk for the included locations in each cluster . For the Poisson model, the results depend on accurate data on the population at risk, which may be hard to obtain. Furthermore, the prospective space-time scan statistic as an exploratory method, should be followed with other surveillance measures and more detailed investigation of transmission dynamics and pathogenic mechanics of COVID-19 to better understand detected emerging clusters (Desjardins et al., 2020).

While the prospective space-time scan statistic has demonstrated value for COVID-19 surveillance, the objective of this study was to demonstrate a different but complementary view of COVID-19 outbreak patterns. The space time scan statistic detects hotspots but does not inform about locations that may be spatially disparate yet may be exhibiting highly similar patterns in disease case count evolution. To capture this dynamic, we employed an event sequence similarity metric on the sequences of daily COVID incidence rates by county. This event sequence similarity metric was then used to cluster counties exhibiting similarly evolving COVID -19 case histories. The resulting identification of locations exhibiting similar evolutionary patterns in the disease provides another aid for public health responses and understanding of disease dynamics. In the remainder of this paper, we describe this event sequence similarity metric as applied to COVID-19 daily incidence rates and compare it with results of the prospective Poisson space-time scan statistic. We use four time periods to illustrate progression of COVID-19

outbreaks through the lens of prospective space-time scan statistic generated clusters and event sequence similarity clusters. The two approaches provide different but complementary aids to COVID-19 surveillance. One tells us of emerging spatial hotspots, the other tells us of collections of locations that for some reasons have statistically similar evolving COVID-19 incidence patterns.

5.2. Materials and Methods

5.2.1. Data Acquisition and Processing

We accessed COVID-19 raw daily global collection data from the GitHub repository (https://github.com/CSSEGISandData/COVID-19) created and maintained by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) (Dong et al., 2020). The specific time series dataset for this research contains FIPS codes, state names, geolocations, and confirmed cumulative cases, starting from January 22, 2020 through selected ending dates. JH CCSE continues to semi-automatically or automatically update their site daily (https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/).

County level population data for the USA were obtained from the national US Census with estimates for 2019. The ESRI ™ shapefiles of US states and counties used for Geographic Information System (GIS) mapping were downloaded from the TIGER geography portal (US Census Bureau) (https://www.census.gov/cgi-bin/geo/shapefiles/index.php).

We focused the analysis on the 48 contiguous states and Washington D. C.. The dataset was cleaned by filtering out the records without "FIPS" codes and names of counties, and with "FIPS" > 8000 (assigned with "Out of AL", "Out of AK", …, "Out of WY"). We combined the cleaned COVID-19 dataset with the U.S. census data at the county level through the "FIPS" codes and double checked the correctness of the spatial information (Latitude and Longitude). Because the COVID-19 dataset only contains cumulative case counts, we obtained the daily confirmed cases by subtracting the previous day's number from the current day's reported cumulative cases. The daily incidence rate for each county was obtained as daily confirmed cases divided by county population and multiplied by 10,000. We chose the data from the first wave of the COVID-19 pandemic in the US in 2020 for this study. The entire duration of the first wave is further divided into four analysis periods considering the incubation time for the disease mostly ranging from 1-14 days with the average of 5 days (He et al., 2020) and the slow case increment at the beginning time in January and February, 2020. The four analysis periods each start from January 22 and cover roughly 2-4 week separations corresponding to an early period 1) March 13, and spiking periods 2) March 31, 3) April 19 and 4) May 20.

5.2.2. Prospective Poisson space-time scan statistic

We used the prospective Poisson space–time scan statistic as implemented in SaTScan (http://www.satscan.org/) to detect clusters of COVID-19 cases that remained active at the end of each study period. The space–time scan statistic (STSS) is briefly introduced here, and more details can be obtained from (Desjardins et al., 2020; Kulldorff, 2001; Kulldorff et al., 2005a; Kulldorff et al., 2007). With spatial scan statistics we can identify the locations of clusters of cases. A cluster can be defined as a set of points or regions, at a

user defined granularity, with either high or low rates of incidence. For this study, the focus was high rates of COVID-19 incidence. Conceptually the STSS uses a cylinder as the scanning window, where the circular base of the cylinder captures the spatial dimension while the height represents a temporal interval. To identify space-time clusters at the county level, the center of the circular base is co-located with the centroid of each county. As the scan progresses, the radius of the circular base and the height of the cylinder changes from lower bounds to spatial and temporal upper limits. Similar to (Desjardins et al., 2020) we set the maximum scanning window base to include up to 10 percent of the total population to avoid the potential of extremely large clusters (ie. covering a quarter of the country) especially as may occur at the beginning stage of the epidemic, and the upper temporal bound to 50% of the entire study period. As each cylinder moves over the study area, it covers a different set of cases for different time intervals, which can be considered as potential emerging space-time cluster candidates. We set the cluster's duration to a minimum of 2 days and required at least 5 incidents or confirmed cases of COVID-19 as described in (Desjardins et al., 2020).

The age structure of a population will influence the incidence of disease, and deaths from COVID-19 are several times higher in older age groups as noted by others [12]. However, we were unable to access age and sex data at this time for cases in this study, so we could not adjust for age and sex. Assuming that COVID-19 incidence follows a Poisson distribution according to the county population, e.g. the assumed population at risk (Kulldorff, 2001), the likelihood ratio test statistic and the relative risk for each scan cylinder was calculated based on the description in (Desjardins et al., 2020; Kulldorff, 1997; 1999; 2001). The cylinder with the maximum likelihood ratio identifies the location

with the most likely elevated risk for COVID-19. We used Standard Monte Carlo simulations (999) in the SaTScan setting to calculate the statistical significance of detected clusters with a p-value equal or less than 0.05 being considered statistically significant. SaTScan computes the relative risk (RR) for each cluster and individual counties. The RR for a county within a cluster can be calculated as in [18]:

$$RR_{cty} = \frac{c/e}{(C-c)(C-e)}$$

Where, c is the total number of cases in a county, C is the total number of observed cases in the conterminous US, and e is the expected number of cases in a county calculated as $e = p_{cty} * \frac{C}{P}$ (p_{cty} is the population in a county, P is the total population). We used ESRI ArcGIS 10.6 (www.esri.com) GIS software to create cartographic representations for these detected emerging clusters at the county level.

5.2.3. Event sequence similarity-based cluster analysis

Our event sequence similarity approach focuses on the temporal evolution of events occurring at fixed locations. In this study, an event corresponds to the COVID-19 daily incidence rate for a county and a COVID-19 event sequence for a county is the sequence of daily incidence rates covering a specific study period. We compute the similarity of these county level COVID-19 event sequences using a time ordered Jaccard measure (Ayub et al., 2018, pp. 1-6; Jaccard, 1901; Sun et al., 2017). Briefly, this measure uses all co-occurrence time points between two event sequences es_1 and es_2, and calculates the similarity between two events at the co-occurrence timestamp based on their level of

measurement. The similarity between two counties' COVID-19 event sequences is calculated as below:

$$sim_{county}(es_1, es_2) = \frac{\sum_{j=1}^{C}(1 - Abs(lev(es_{1j}) - lev(es_{2j})))}{|es_1 \cup es_2|}$$

where,

$sim_{county}(es_1, es_2)$ – Similarity between county level event sequences es_1 and es_2,

es_{1j}, es_{2j} – the event values for two corresponding co-occurring events in es_1 and es_2 at timestamp j.

$lev(es_{1j}), lev(es_{2j})$ – the relative event levels of two corresponding co-occurring events in es_1 and es_2 at timestamp j, respectively:

$lev(es_{1j}) = \frac{es_{1j}}{es_{1j}+es_{2j}}$ and $lev(es_{2j}) = \frac{es_{2j}}{es_{1j}+es_{2j}}$

C – the total number of co-occurring timestamps,

$Abs(lev(es_{1j}) - lev(es_{2j}))$ – absolute value of difference between relative event levels of two corresponding co-occurring events in es_1 and es_2 at timestamp j,

$|es_1 \cup es_2|$ – Cardinality of the union of two event sequences es_1 and es_2.

We then used the computed COVID-19 event sequence similarity measures between counties as the metric for hierarchical clustering (Ros and Guillaume, 2019). All similarity computations and clustering tasks were implemented in R. The hierarchical

clustering was performed using the hclust R function with the linkage method of Ward.D2. The optimal number of clusters was evaluated using the elbow method (Gustriansyah et al., 2020; Syakur et al., 2018; Zambelli, 2016). This method supports selection of the number of clusters at which the total within-cluster sum of square (WSS) no longer improves. In a plot of number of clusters versus WSS, the optimal cluster number is visually associated with the point at which the WSS value flattens.

5.2.4. Comparison of Prospective Space time Scan and Event Sequence Similarity-based clusters

To support comparison of the two methods we used the counties identified in the prospective Space time scan statistics as having relative risk > 1 as the counties for analysis with the sequence similarity metric. All other counties not included in this set were labeled as OC meaning outside clusters. We include them in Figures, 5.3, 5.6, and 5.9 in the graphs of incidences curves for each study period to show their temporal incidence pattern as a baseline.

5.3. Results

5.3.1. Space-time clusters and sequence similarity-based clusters at county level: Study period 1 (1/22-3/13/2020)

In this early period, COVID-19 was just appearing in the US with the first case reported in Snohomish County Washington on January 19. For this period, the prospective space-time scan statistic identified 11 statistically significant (p-value < 0.05) clusters shown graphically in Figure 5.1 and summarized in Table 5.1. These clusters, aside from

one in California and two in New York, are generally quite large and counties within them with RR > 1 are few and generally spatially dispersed. Because of the generally large size of these clusters, identifying the spatial specificity of an outbreak is limited.

Table 5.1. Attributes of prospective space-time clusters (hotspots) for COVID-19 from 1/23-3/13/2020 at the county level.

Cluster	Start Date	End Date	Duration (Days)	Radius (Km)	Observed Cases	Expected Cases	Relative Risk (RR)	p-value	Population at Risk	#County (total)	#County (RR>1)
1	3/10	3/13	4	806.37	389	38	12.28	<0.001	888,297	238	14
2	3/7	3/13	7	0.00	139	15	10	<0.001	189,707	1	1
3	3/10	3/13	4	551.69	66	18	3.83	<0.001	167,447	404	16
4	3/9	3/13	5	364.08	42	10	4.29	<0.001	87,766	262	16
5	3/12	3/13	2	32.48	102	47	2.21	<0.001	1,267,395	9	6
6	3/12	3/13	2	91.08	10	0	29.12	<0.001	7,438	35	3
7	3/5	3/13	9	49.70	93	42	2.25	<0.001	790,544	3	3
8	3/9	3/13	5	178.04	9	0	26.67	<0.001	2,607	94	3
9	3/10	3/13	4	224.18	12	1	14.16	<0.001	15,926	104	3
10	3/10	3/13	4	253.24	12	1	10.51	<0.001	8,832	64	3
11	3/7	3/13	7	264.34	88	47	1.91	<0.001	824,139	36	12

Note: Space-time clusters were identified using the spatial scan statistic with a Poisson model

Based on the elbow evaluation method, 8 event sequence similarity-based clusters were defined for this period (Figure 5.2). Figure 5.3 shows the map representation of these clusters along with their temporal profiles. Members of Cluster 3 that include counties in Washington State, California and New York show the earliest onset and the fastest case accumulation. Members of Cluster 5 show an early onset that initially tracks Cluster 3 but

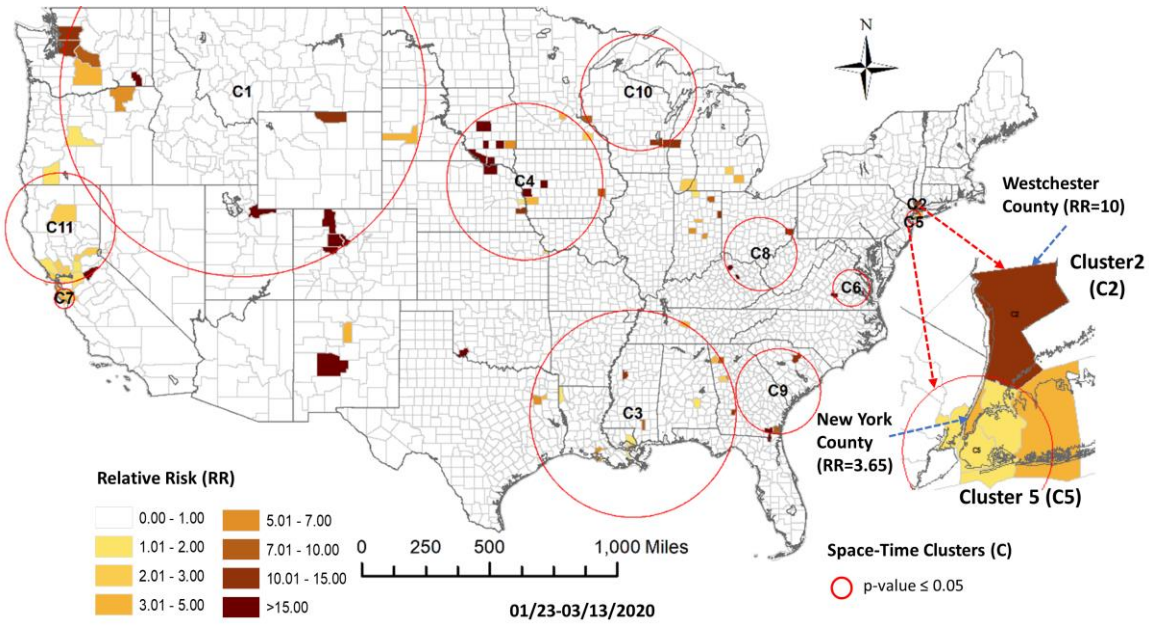


Figure 5.1. COVID-19 space-time scan hotspots in the United States at the county level from 1/22/-3/13/2020.

then abruptly flattens and then decreases in early March. Members of this cluster include 3 counties in California and one in Minnesota. Cluster 2 members show a delayed occurrence in cases but an extremely fast case accumulation over a few days. The 8 members of this cluster are generally in isolated rural settings in Colorado, Oklahoma, Wyoming, South Dakota, Wisconsin, Louisiana and Indiana. Members of Cluster 6 showed initiation of cases at approximately the same time as Cluster 2 but levelled off quickly at a lower incidence rate. The cluster containing counties in New York suggests initial points of entry and situations conducive to rapid acceleration of cases such as high density or tight knit communities. A pairwise comparison of cluster numbers for the 1st study period from

these two approaches can be found in Table C.1.

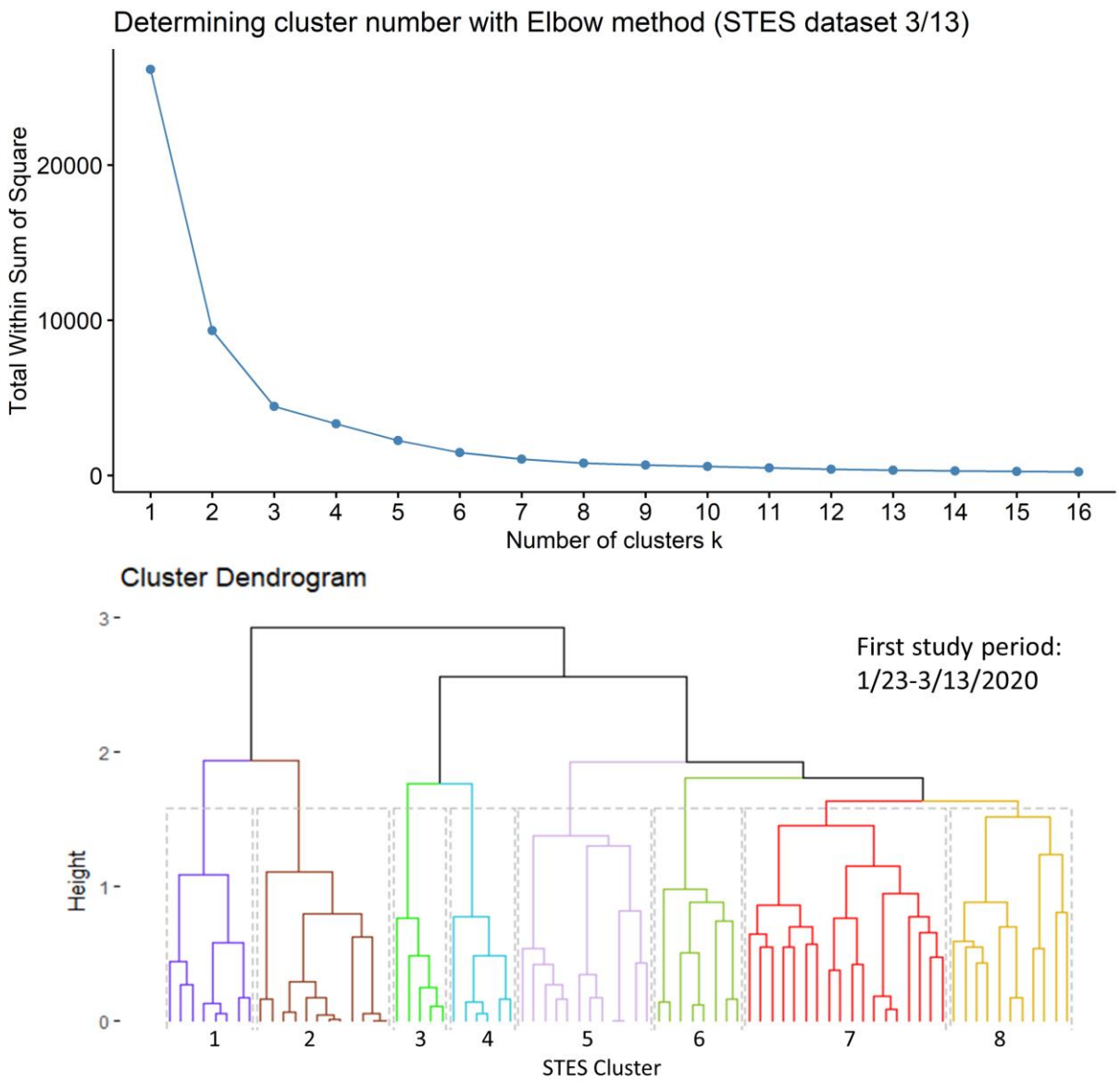


Figure 5.2. Elbow method evaluation and hierarchical clustering results for the 1st period. Notice that the numberings and colors of STES clusters match with those of corresponding clusters on the map and the temporal trend graph in Figure 5.3.

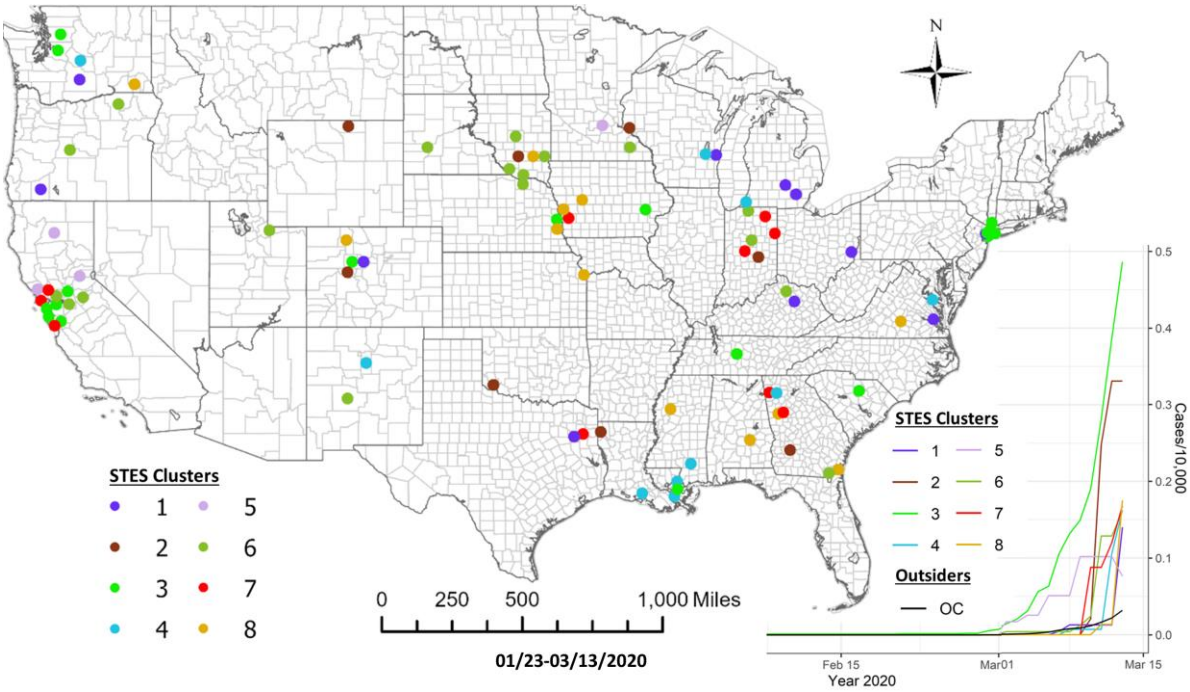


Figure 5.3. Sequence similarity-based COVID-19 clusters along with average temporal trends at the county level through 3/13/2020. This map includes the counties with higher relative risk (RR>1) contained in all the clusters detected by scan statistics in Figure 5.1. The average temporal trends of cumulative cases for STES clusters 1-8 on the map appear at the bottom right. Notice that the colors of STES clusters match with correspondingly colored dots on the map and with the colors of the STES cluster curves on the graph. OC includes all counties not included in the clusters.

5.3.2. Space-time clusters and sequence similarity-based clusters at county level: Study period 2 (1/22-3/31/2020)

Results from the prospective space-time scan statistics analysis for the second study period (through March 31) identified twenty-four space-time clusters of COVID-19 as statistically significant (Figure 5.4 and Table 5.2). This period shows a growing emergence of spatial clusters across the US, but generally more consolidated clusters as the number of cases grow. The space-time clusters are smaller than in the first period and several detected clusters contain a single county (cluster radius = 0). This period shows a shift toward more clusters appearing in the interior US relative to the coasts.

Table 5.2. Attributes of prospective space-time clusters (hotspots) for COVID-19 from 1/23-3/31/2020 at the county level.

Cluster	Start Date	End Date	Duration (Days)	Radius (Km)	Observed Cases	Expected Cases	Relative Risk (RR)	p-value	Population at Risk	#County (total)	#County (RR>1)
1	3/22	3/31	13	89.28	82,928	10,049	14.35	<0.001	6,395,723	22	22
2	3/22	3/31	10	43.08	5,887	1,526	3.95	<0.001	1,074,213	3	3
3	3/20	3/31	12	73.70	3,152	487	6.57	<0.001	292,363	8	8
4	3/27	3/31	5	0.00	3,078	1,012	3.08	<0.001	2,201,911	1	1
5	3/24	3/31	8	73.96	680	68	9.97	<0.001	39,490	20	18
6	3/26	3/31	6	60.42	2,587	1,102	2.37	<0.001	1,370,768	2	2
7	3/24	3/31	8	62.27	2,041	846	2.43	<0.001	1,345,457	4	4
8	3/19	3/31	13	95.88	190	11	17.17	<0.001	5,083	4	3
9	3/30	3/31	2	307.75	1,528	729	2.11	<0.001	1,822,585	262	82
10	3/16	3/31	16	82.42	313	54	5.78	<0.001	28,677	5	5
11	3/20	3/31	12	146.72	214	38	5.6	<0.001	20,460	9	4
12	3/29	3/31	3	325.81	4,574	3,543	1.3	<0.001	6,684,959	257	75
13	3/27	3/31	5	210.38	787	448	1.76	<0.001	647,610	43	10
14	3/30	3/31	2	0.00	1,190	789	1.51	<0.001	3,855,599	1	1
15	3/25	3/31	7	50.46	206	72	2.88	<0.001	57,714	5	2
16	3/23	3/31	9	49.14	84	14	5.86	<0.001	5,999	5	4
17	3/30	3/31	2	240.79	344	179	1.92	<0.001	528,991	11	3
18	3/29	3/31	3	0.00	27	2	11.75	<0.001	1,412	1	1
19	3/14	3/31	18	36.13	105	44	2.4	<0.001	20,986	2	2
20	3/22	3/31	10	42.64	35	8	4.27	<0.001	3,227	4	4
21	3/30	3/31	2	0.00	244	152	1.61	<0.001	991,866	1	1
22	3/24	3/31	8	54.38	22	4	5.76	<0.001	1,899	8	5
23	3/27	3/31	5	139.67	101	50	2.02	<0.001	49,538	2	2
24	3/11	3/31	21	188.69	48	17	2.85	<0.001	6,210	45	16

Note: Space-time clusters were identified using the spatial scan statistic with a Poisson model

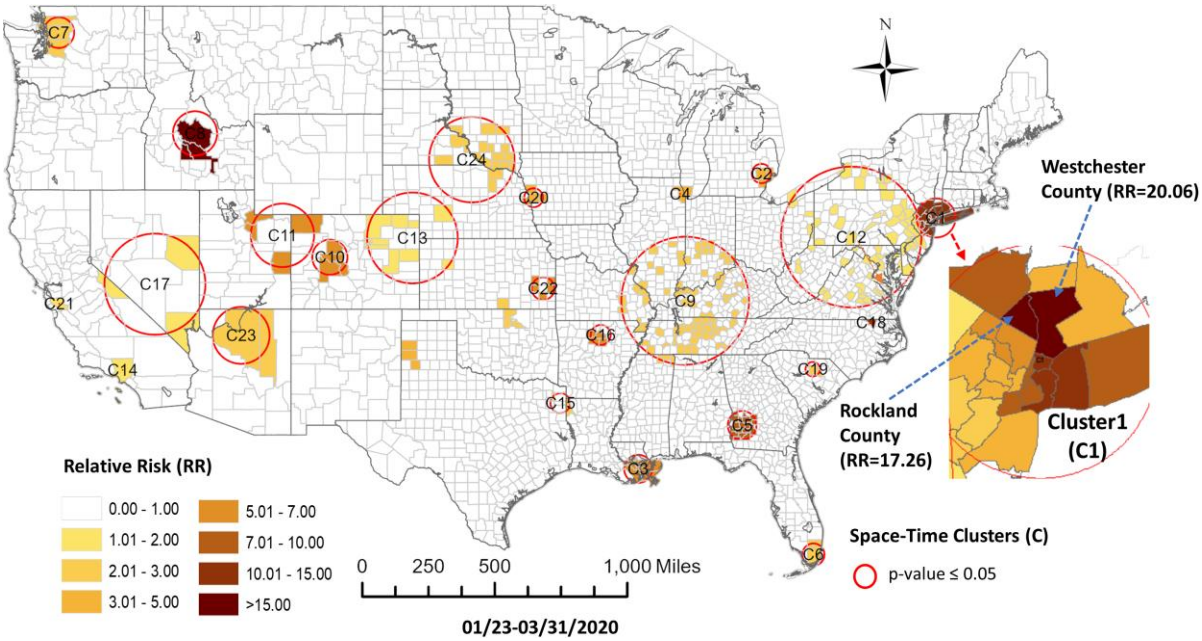


Figure 5.4. COVID-19 space-time scan statistic detected hotspots in the United States at county level through 3/31/2020.

For this second study period the sequence similarity clustering resulted in 8 clusters based on the elbow method evaluation (Figure 5.5). Figure 5.6 shows the map of these clusters and their temporal signatures. For this period, only three clusters deviate from the outside cluster (OC) set pattern. Cluster 7 shows the most rapid increase in cases. Members of this cluster include Miami, San Jose, Los Angeles area counties, Chicago, Detroit, New Orleans and New York metropolitan counties. Members of Cluster 8 show a slower and less rapid increase in cases. Some of these members appear in a group across New Jersey and Pennsylvania, around Baltimore, Denver and Seattle. Cluster 4 follows a similar trajectory with some concentrations around New Orleans, Columbus Georgia, and Indianapolis. Members of this cluster also appear in more isolated rural settings in Arizona, Oklahoma and South Dakota. A pairwise comparison of cluster numbers for the 2nd study period from these two approaches can be found in Table C.2.

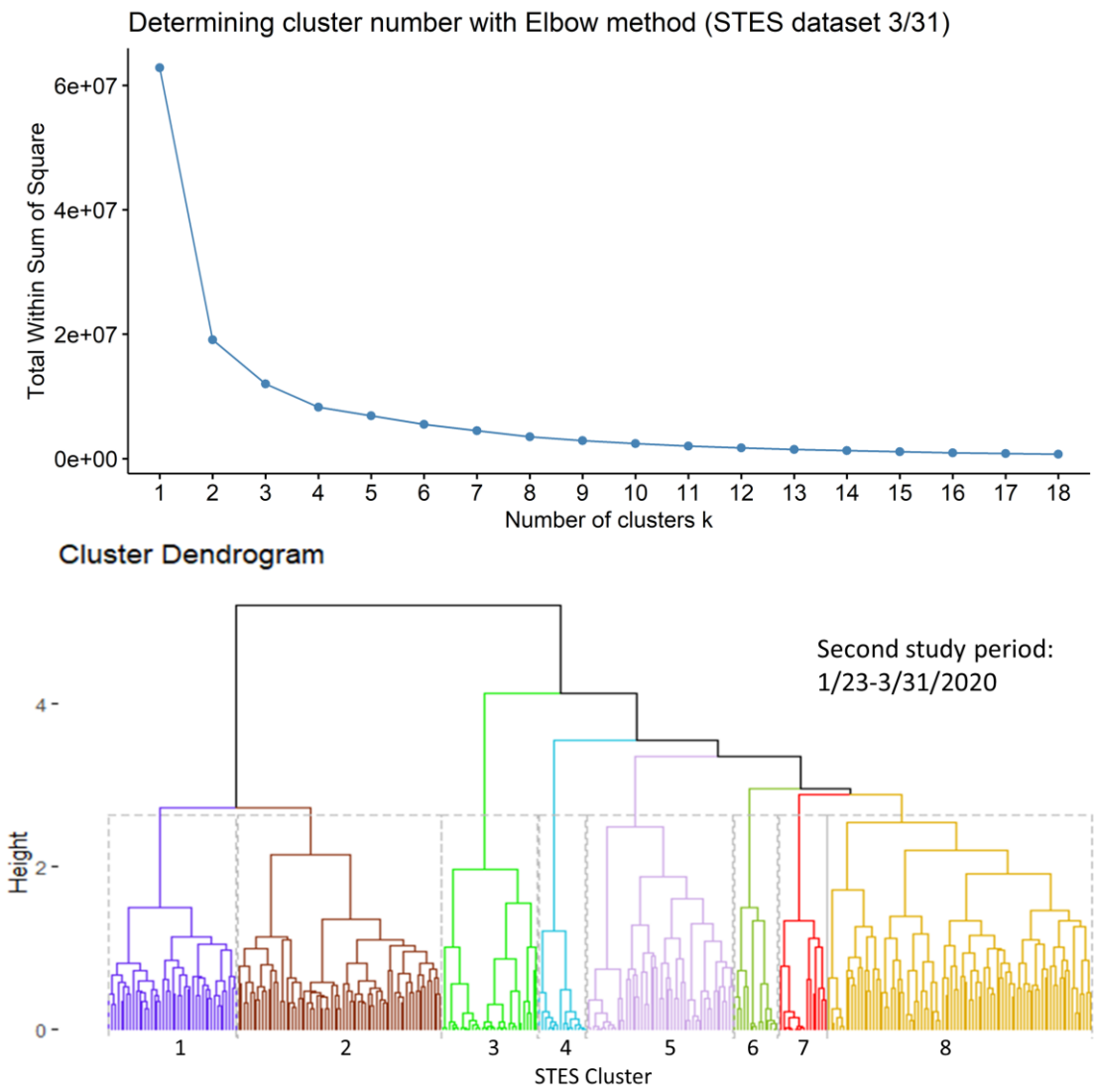


Figure 5.5. Elbow method evaluation and hierarchical clustering results for the 2nd period. Notice that the numberings and colors of STES clusters match with those of corresponding clusters on the map and the temporal trend graph in Figure 5.6.

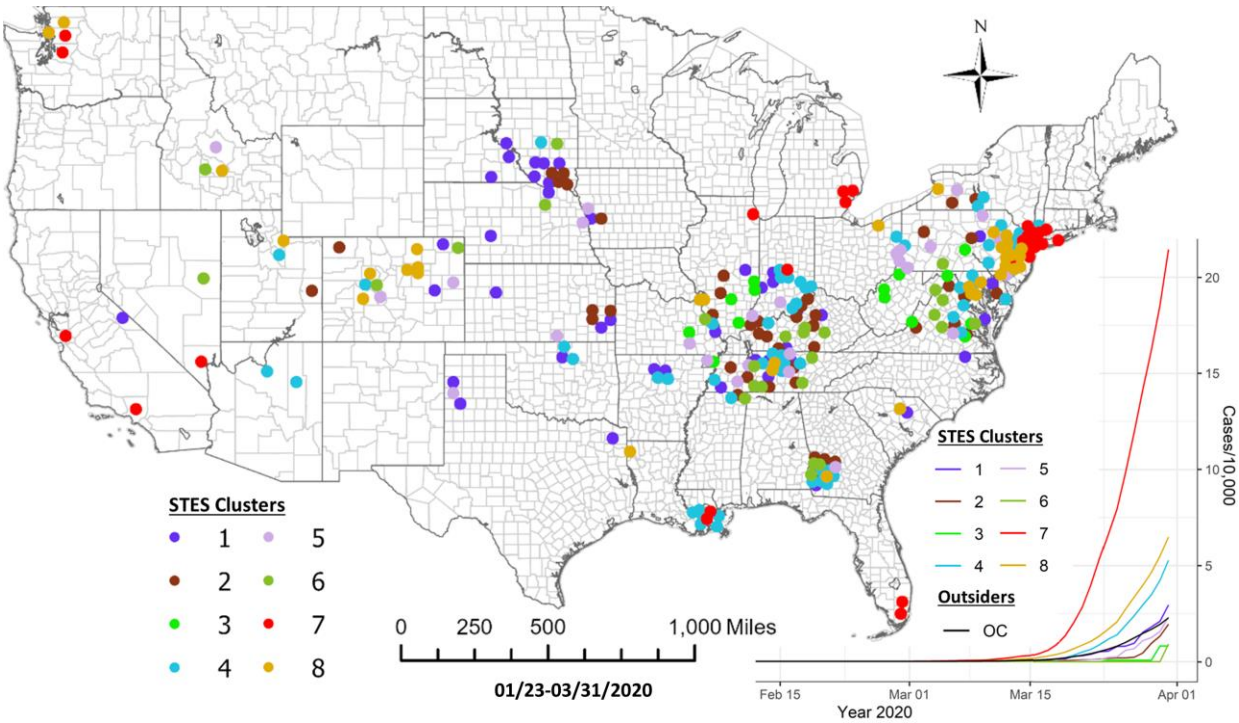


Figure 5.6. **Sequence similarity-based COVID-19 clusters along with average temporal trends at county level during 1/22/2020-3/31/2020.** This map includes the counties with higher relative risk (RR>1) contained in all the clusters detected by scan statistics in Figure 5.3. The average temporal trends of cumulative cases for STES clusters 1-8 on the map appear at the bottom right. Notice that the colors of STES clusters match with correspondingly colored dots on the map and with the colors of the STES cluster curves on the graph. OC includes all counties not included in the clusters.

5.3.3. Space-time clusters and sequence similarity-based clusters at county level: Study period 3 (1/22-4/19/2020)

For the third study period, the prospective space-time cluster statistic detected 47 statistically significant clusters ($p \leq 0.05$) as shown in Figure 5.7. Associated cluster characteristics are shown in Table 5.3. In this period more clusters are emerging in the southern US, with additional new pockets in Montana and a cluster covering Nebraska and South Dakota. Metropolitan New York remains an active cluster and a more condensed

Mid-Atlantic coast cluster has emerged. We see additional consolidation in the size of clusters with 25 appearing as a single county.

Table 5.3. Attributes of prospective space-time clusters (hotspots) for COVID-19 from 1/23-4/19/2020 at the county level.

Cluster	Start Date	End Date	Duration (Days)	Radius (Km)	Observed Cases	Expected Cases	Relative Risk (RR)	p-value	Population at Risk	#County (total)	#County (RR>1)
1	3/21	4/19	30	112.67	317,283	50,808	10.07	<0.001	10,183,190	29	29
2	3/25	4/19	26	73.70	13,048	2,223	5.96	<0.001	468,407	8	8
3	3/27	4/19	24	43.08	22,215	7,189	3.15	<0.001	1,680,202	3	3
4	4/16	4/19	4	0.00	1,670	20	83.28	<0.001	19,232	1	1
5	4/4	4/19	16	0.00	15,161	6,360	2.41	<0.001	2,838,481	1	1
6	3/31	4/19	20	77.77	2,949	441	6.72	<0.001	93,100	22	22
7	4/6	4/19	14	298.19	40,502	27,081	1.52	<0.001	9,421,799	226	93
8	4/10	4/19	10	263.00	1,767	341	5.2	<0.001	137,317	85	26
9	3/30	4/19	21	0.00	8,162	4,404	1.86	<0.001	1,173,224	1	1
10	3/26	4/19	25	0.00	435	36	12.25	<0.001	7,586	1	1
11	4/17	4/19	3	0.00	360	29	12.63	<0.001	30,783	1	1
12	4/1	4/19	19	59.89	1,270	464	2.74	<0.001	116,600	6	5
13	4/9	4/19	11	162.39	832	271	3.07	<0.001	112,063	5	5
14	3/20	4/19	31	84.21	760	281	2.71	<0.001	52,008	6	6
15	3/31	4/19	20	218.29	10,400	8,205	1.27	<0.001	1,932,165	152	77
16	4/5	4/19	15	169.63	400	104	3.84	<0.001	22,025	36	20
17	4/9	4/19	11	42.71	309	67	4.58	<0.001	24,501	3	3
18	4/14	4/19	6	36.59	428	142	3.02	<0.001	97,393	6	6
19	4/13	4/19	7	41.53	100	6	16.58	<0.001	2,434	2	1
20	4/9	4/19	11	144.34	2,800	1,943	1.44	<0.001	999,773	20	14
21	4/14	4/19	6	0.00	109	10	10.73	<0.001	5,683	1	1
22	3/20	4/19	31	0.00	299	88	3.41	<0.001	16,762	1	1

Table 5.3. Continued

23	4/2	4/19	18	0.00	643	349	1.85	<0.001	94,077	1	1
24	4/7	4/19	13	70.67	348	179	1.95	<0.001	41,649	17	14
25	4/15	4/19	5	0.00	123	37	3.35	<0.001	29,216	1	1
26	4/17	4/19	3	192.58	142	51	2.8	<0.001	50,741	11	6
27	4/18	4/19	2	37.48	298	152	1.96	<0.001	386,360	2	2
28	4/3	4/19	17	92.71	301	156	1.93	<0.001	41,584	5	3
29	4/11	4/19	9	0.00	173	77	2.25	<0.001	41,981	1	1
30	4/11	4/19	9	0.00	83	24	3.48	<0.001	14,638	1	1
31	4/15	4/19	5	0.00	41	7	6.16	<0.001	3,595	1	1
32	4/15	4/19	5	72.81	57	13	4.29	<0.001	10,680	8	5
33	4/14	4/19	6	0.00	1,019	763	1.34	<0.001	926,455	1	1
34	4/13	4/19	7	0.00	583	410	1.42	<0.001	336,507	1	1
35	3/28	4/19	23	50.34	32	5	6.04	<0.001	888	2	2
36	4/2	4/19	18	68.61	253	149	1.7	<0.001	28,897	10	9
37	4/12	4/19	8	0.00	59	20	2.96	<0.001	8,797	1	1
38	4/18	4/19	2	0.00	272	174	1.56	<0.001	1,139,191	1	1
39	4/17	4/19	3	0.00	37	10	3.74	<0.001	27,699	1	1
40	3/29	4/19	22	0.00	105	52	2.02	<0.001	9,587	1	1
41	4/18	4/19	2	0.00	20	3	6.4	<0.001	7,819	1	1
42	3/23	4/19	28	44.85	112	58	1.94	<0.001	9,320	5	5
43	4/11	4/19	9	0.00	93	46	2.02	<0.001	17,771	1	1
44	4/18	4/19	2	0.00	14	2	8.17	0.002	3,531	1	1
45	4/14	4/19	6	0.00	22	5	4.71	0.003	2,749	1	1
46	4/18	4/19	2	0.00	53	21	2.49	0.003	31,371	1	1
47	3/24	4/19	27	0.00	102	55	1.85	0.006	10,847	1	1

Note: Space-time clusters were identified using the spatial scan statistic with a Poisson model

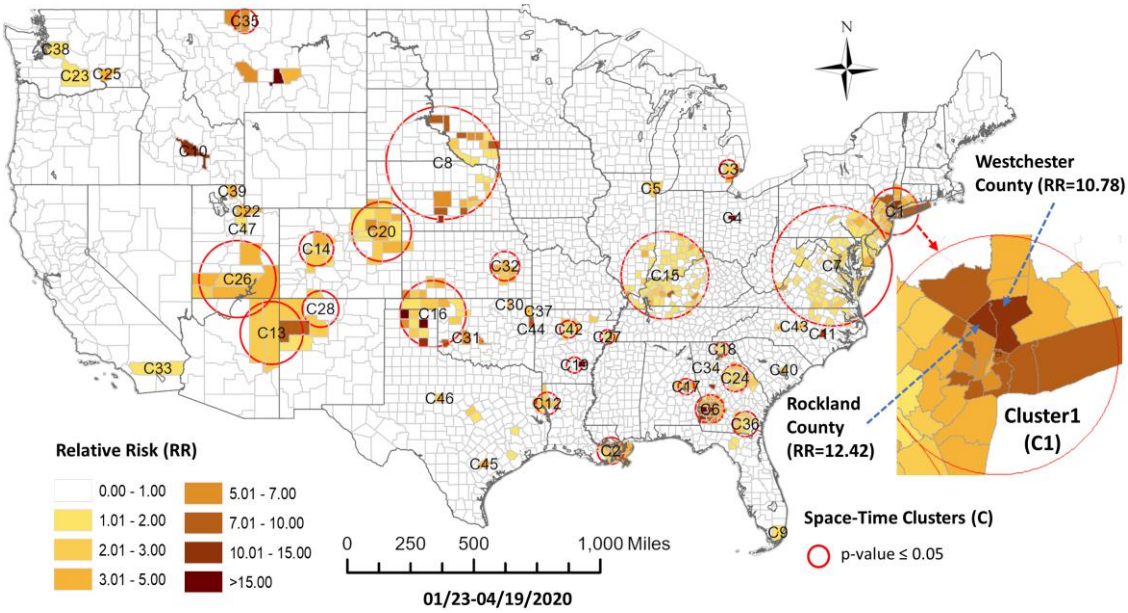


Figure 5.7. COVID-19 space-time scan statistic detected hotspots in the United States at county level through 4/19/2020.

For the third study period, ten sequence similarity-based clusters were selected using the elbow method (Figure 5.8). Figure 5.9 shows the map of these clusters and their temporal profiles. Cluster 8 shows a distinct early and more rapid accumulation of cases. Many members of this cluster were members of Cluster 7 in the previous study period. These members include Chicago, Detroit metropolitan area, Miami, Philadelphia, and metropolitan New York counties. Some significant missing members in Cluster 8 from the previous period Cluster 7 are San Jose, Los Angeles and Las Vegas. Cluster 9 shows a group with the next most rapidly developing number of cases. Within this group, some members appear concentrated around metropolitan New York, Philadelphia, Baltimore and Washington DC, and Denver. Cluster 10, as the third most rapidly merging cluster for this period, has members in a halo like pattern around metropolitan New York, Philadelphia and New Orleans. Other members, however, appear in more isolated rural settings in New

Mexico, Utah, and Washington State. This group includes the Hopi, Zuni, Navajo and Yakima national reservations. Two other clusters to note in this group are Cluster 7 and Cluster 2 which show later initiation times in terms of case accumulation but appear to be accelerating at the end of the study period. Many of these members show a concentration in southern Indiana and western Kentucky respectively, with another grouping of Cluster 7 members appearing in southwestern Georgia on the border with Alabama. A complete pairwise comparison of cluster numbers for the 3rd study period from these two approaches can be found in Table C.3.

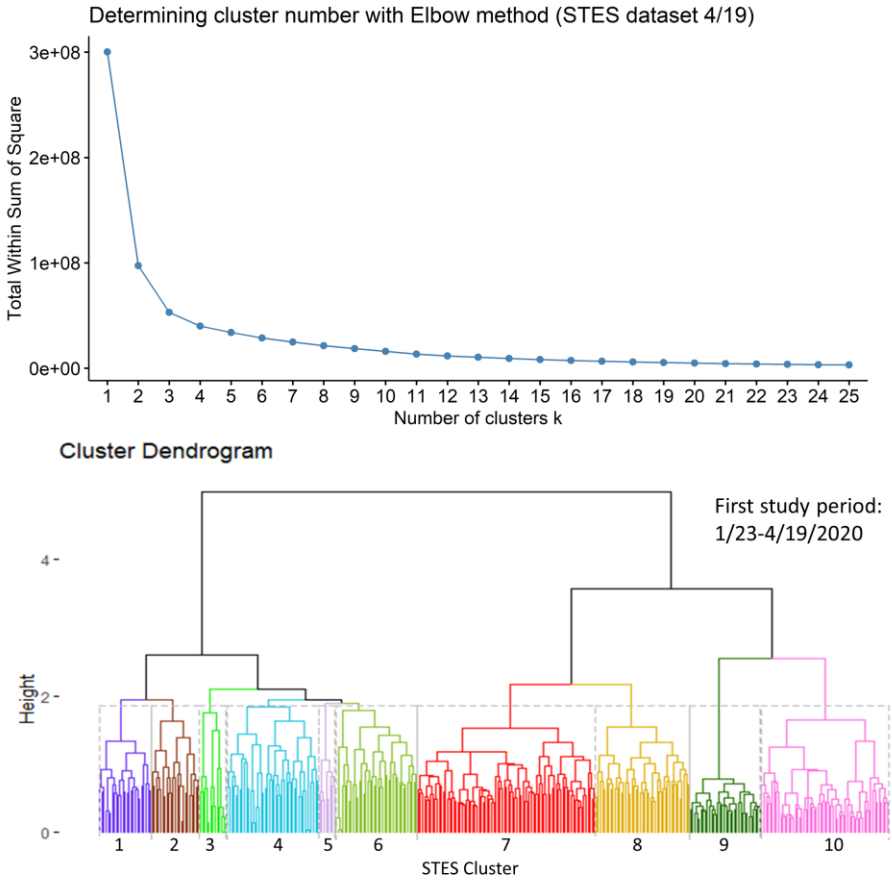


Figure 5.8. Elbow method evaluation and hierarchical clustering results for the 3rd period. Notice that the numberings and colors of STES clusters match with those of corresponding clusters on the map and the temporal trend graph in Figure 5.9.

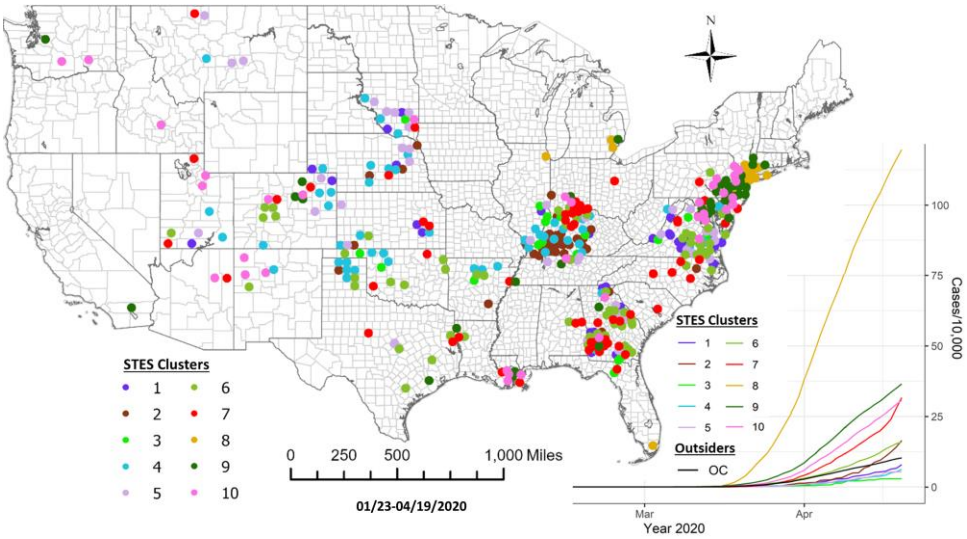


Figure 5.9. Sequence similarity-based COVID-19 emerging clusters along with average temporal trends at county level during 1/22/-4/19/2020. This map includes the counties with higher relative risk (RR>) contained in all the clusters detected by scan statistics in Figure 5.5. The average temporal trends of cumulative cases for STES clusters 1-10 on the map appear at the bottom right. Notice that the colors of STES clusters match with correspondingly colored dots on the map and with the colors of the STES cluster curves on the graph. OC includes all counties not included in the clusters.

5.3.4. Space-time clusters and sequence similarity-based clusters at county level: Study period 4 (1/22-5/20/2020)

For the fourth study period ending on May 20, 2020, the prospective space-time scan statistic identified 87 statistically significant clusters. Table 5.4 provides the characteristics of these 87 active space-time clusters at the end of May 20, 2020. From Figure 5.10 we can observe that in this period clusters continued to emerge in southern states and more clusters emerge in the mountain west. The previous cluster covering Nebraska and South Dakota has expanded into Iowa, North Dakota and Minneapolis. The

metropolitan New York cluster has consolidated and the prior period mid-Atlantic cluster has consolidated to an emerging cluster around Philadelphia.

Table 5.4. Attributes of prospective space-time clusters (hotspots) for COVID-19 from 1/23-5/20/2020 at the county level.

Cluster	Start Date	End Date	Duration (Days)	Radius (Km)	Observed Cases	Expected Cases	Relative Risk (RR)	p-value	Population at Risk	#County (total)	#County (RR>1)
1	3/23	5/20	59	126.60	516,153	128,515	5.51	<0.001	15,225,284	35	35
2	4/7	5/20	44	55.64	77,744	30,138	2.66	<0.001	5,000,478	5	5
3	4/12	5/20	39	332.91	14,779	3,116	4.78	<0.001	411,108	155	109
4	4/17	5/20	34	103.56	41,285	18,966	2.21	<0.001	3,575,889	42	25
5	4/20	5/20	31	215.21	7,183	749	9.63	<0.001	111,251	47	35
6	3/23	5/20	59	73.70	16,614	5,499	3.04	<0.001	625,641	8	8
7	3/26	5/20	56	43.08	34,409	18,624	1.87	<0.001	2,253,493	3	3
8	4/29	5/20	22	0.00	1,336	16	81.34	<0.001	3,508	1	1
9	4/13	5/20	38	0.00	2,487	206	12.07	<0.001	30,632	1	1
10	4/9	5/20	42	191.99	5,571	1,339	4.17	<0.001	184,726	6	6
11	4/15	5/20	36	0.00	1,952	175	11.15	<0.001	25,544	1	1
12	3/24	5/20	58	77.77	4,684	1,282	3.66	<0.001	134,101	22	22
13	4/13	5/20	38	0.00	955	36	26.75	<0.001	4,378	1	1
14	4/15	5/20	36	114.37	3,799	1,339	2.84	<0.001	187,231	21	21
15	4/23	5/20	28	0.00	598	21	28.96	<0.001	3,038	1	1
16	5/12	5/20	9	0.00	344	3	114.45	<0.001	1,002	1	1
17	4/14	5/20	37	36.59	2,623	962	2.73	<0.001	150,923	6	5
18	4/24	5/20	27	42.39	1,579	458	3.45	<0.001	77,989	7	7
19	4/30	5/20	21	0.00	1,436	451	3.18	<0.001	134,923	1	1
20	5/3	5/20	18	0.00	191	4	44.47	<0.001	772	1	1
21	3/23	5/20	59	47.10	519	87	5.99	<0.001	9,665	2	2
22	4/28	5/20	23	45.28	436	77	5.66	<0.001	13,095	3	3
23	5/10	5/20	11	29.09	221	15	14.38	<0.001	3,235	3	3

Table 5.4. Continued

24	5/10	5/20	11	0.00	257	24	10.91	<0.001	7,981	1	1
25	4/30	5/20	21	0.00	354	56	6.27	<0.001	11,332	1	1
26	5/6	5/20	15	0.00	994	383	2.6	<0.001	202,613	1	1
27	4/1	5/20	50	136.56	5,564	3,846	1.45	<0.001	449,669	30	22
28	5/7	5/20	14	0.00	566	155	3.65	<0.001	71,572	1	1
29	5/2	5/20	19	31.84	510	133	3.83	<0.001	20,764	4	4
30	4/19	5/20	32	192.58	305	51	6.02	<0.001	40,867	11	2
31	3/30	5/20	52	0.00	14,842	12,107	1.23	<0.001	1,575,369	1	1
32	4/21	5/20	30	0.00	517	144	3.6	<0.001	25,141	1	1
33	5/11	5/20	10	0.00	248	37	6.71	<0.001	24,329	1	1
34	5/12	5/20	9	45.71	262	47	5.53	<0.001	32,224	3	1
35	4/27	5/20	24	0.00	153	16	9.6	<0.001	2,345	1	1
36	4/29	5/20	22	37.68	576	218	2.65	<0.001	48,225	2	2
37	4/2	5/20	49	42.71	704	312	2.25	<0.001	36,636	3	3
38	5/8	5/20	13	0.00	164	24	6.95	<0.001	5,473	1	1
39	5/19	5/20	2	0.00	2,437	1,721	1.42	<0.001	6,453,712	1	1
40	5/15	5/20	6	0.00	60	3	21	<0.001	841	1	1
41	5/6	5/20	15	29.36	112	17	6.41	<0.001	4,070	2	2
42	5/10	5/20	11	45.67	150	32	4.62	<0.001	8,166	2	2
43	4/6	5/20	45	30.61	309	116	2.67	<0.001	13,014	3	3
44	4/18	5/20	33	0.00	519	257	2.02	<0.001	42,288	1	1
45	5/7	5/20	14	0.00	105	20	5.2	<0.001	5,939	1	1
46	4/25	5/20	26	99.90	124	29	4.23	<0.001	4,072	15	6
47	4/20	5/20	31	30.03	288	124	2.33	<0.001	22,341	3	2
48	3/23	5/20	59	77.39	581	342	1.7	<0.001	39,119	4	2
49	5/13	5/20	8	106.86	270	121	2.24	<0.001	83,127	2	2
50	3/29	5/20	53	0.00	291	139	2.1	<0.001	15,029	1	1
51	4/22	5/20	29	26.90	155	55	2.83	<0.001	8,779	2	2
52	4/7	5/20	44	46.15	317	165	1.92	<0.001	18,980	6	6

Table 5.4. Continued

53	5/2	5/20	19	0.00	103	33	3.16	<0.001	7,699	1	1
54	4/1	5/20	50	53.19	83	22	3.7	<0.001	2,198	3	3
55	4/14	5/20	37	27.39	68	16	4.25	<0.001	1,791	2	2
56	4/23	5/20	28	0.00	156	65	2.4	<0.001	10,718	1	1
57	4/13	5/20	38	21.26	248	128	1.93	<0.001	19,711	2	2
58	4/27	5/20	24	0.00	30	3	10.24	<0.001	323	1	1
59	5/18	5/20	3	0.00	49	9	5.29	<0.001	15,448	1	1
60	4/17	5/20	34	0.00	107	39	2.73	<0.001	8,405	1	1
61	4/18	5/20	33	72.28	534	354	1.51	<0.001	58,978	7	5
62	4/21	5/20	30	140.99	233	125	1.87	<0.001	26,408	6	4
63	4/29	5/20	22	0.00	234	126	1.85	<0.001	30,406	1	1
64	4/22	5/20	29	0.00	115	47	2.43	<0.001	6,538	1	1
65	5/19	5/20	2	0.00	21	2	12.77	<0.001	4,032	1	1
66	5/5	5/20	16	92.43	1,039	796	1.3	<0.001	286,527	2	2
67	4/19	5/20	32	0.00	115	49	2.37	<0.001	10,204	1	1
68	5/8	5/20	13	0.00	192	101	1.9	<0.001	45,852	1	1
69	5/12	5/20	9	0.00	30	4	6.87	<0.001	771	1	1
70	5/17	5/20	4	0.00	123	55	2.23	<0.001	78,471	1	1
71	4/29	5/20	22	0.00	156	79	1.97	<0.001	17,303	1	1
72	3/28	5/20	54	50.34	32	6	5.44	<0.001	656	2	2
73	5/7	5/20	14	80.26	106	46	2.28	<0.001	12,240	5	4
74	4/14	5/20	37	47.62	115	53	2.15	<0.001	7,305	3	3
75	4/9	5/20	42	35.79	123	59	2.09	<0.001	6,343	2	2
76	4/20	5/20	31	0.00	134	68	1.98	<0.001	11,760	1	1
77	4/28	5/20	23	195.74	281	184	1.53	<0.001	48,676	9	4
78	4/16	5/20	35	27.34	243	154	1.57	<0.001	22,877	3	2
79	4/15	5/20	36	0.00	116	59	1.96	<0.001	7,734	1	1
80	4/9	5/20	42	56.31	478	350	1.36	<0.001	49,008	2	1
81	5/18	5/20	3	93.41	130	70	1.86	<0.001	180,113	8	2
82	4/17	5/20	34	0.00	37	11	3.37	<0.001	20,483	1	1

Table 5.4. Continued

83	4/10	5/20	41	30.49	135	76	1.78	<0.001	9,851	2	2
84	5/14	5/20	7	0.00	125	69	1.82	<0.001	43,779	1	1
85	5/12	5/20	9	27.78	87	44	1.97	0.004	16,827	2	1
86	5/3	5/20	18	0.00	20	4	4.6	0.013	670	1	1
87	5/19	5/20	2	80.43	28	8	3.38	0.019	55,557	12	2

Note: Space-time clusters were identified using the spatial scan statistic with a Poisson model

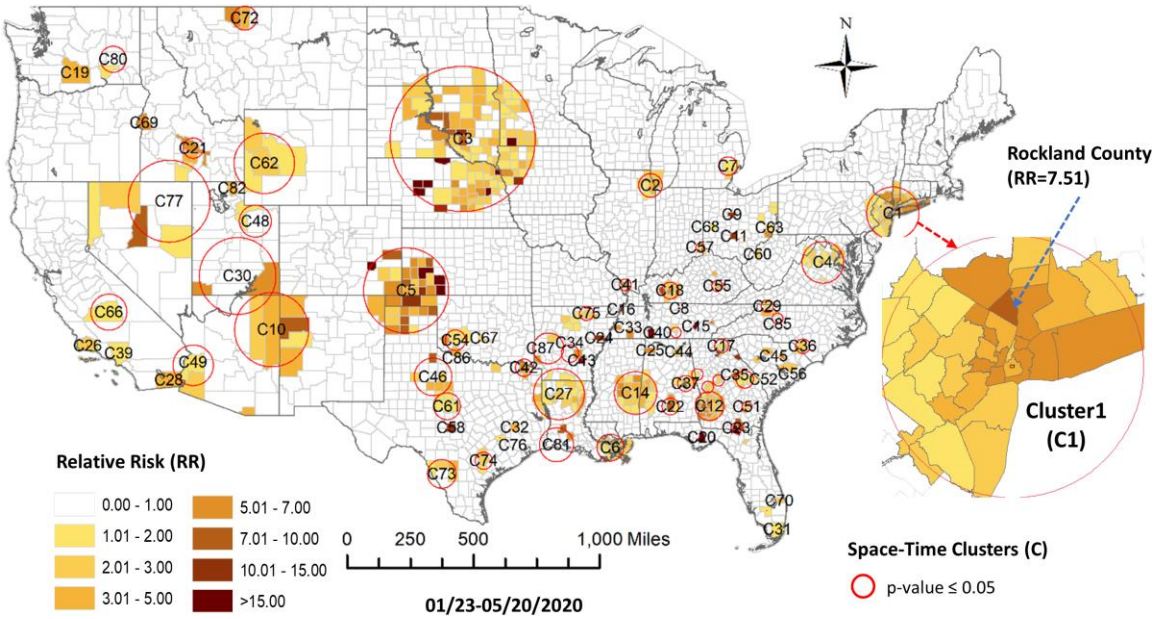


Figure 5.10. **Prospective space-time scan statistic detected clusters of COVID-19 incidents during the study period of 1/22/2020-5/20/2020.**

In this fourth period, using the sequence similarity-based clustering we selected 10 clusters based on the elbow method evaluation (Figure 5.11). Figure 5.12 presents a map of these clusters and their temporal signatures. In this period, Cluster 8 which includes Miami, Chicago, Detroit, Los Angeles, Philadelphia and New York metropolitan counties is the fastest growing in term of cases. Clusters 7 and 9 start out with similar increases in

cases but Cluster 7 members show a levelling off in early May relative to Cluster 9. Cluster 10 shows a delayed start but steady increase starting in early April. Cluster 5 shows a different trajectory in that it shows a much slower start to case accumulation but then exhibits a sharp increase starting in mid-April, increasing more rapidly than Clusters 10 and 7. Cluster 4 initially falls below the outside cluster "OC" group but then shows a sharp jump and more rapid accumulation. More detailed information on pairwise comparison of cluster numbers for the 4th study period from these two approaches can be found in Table C.4.

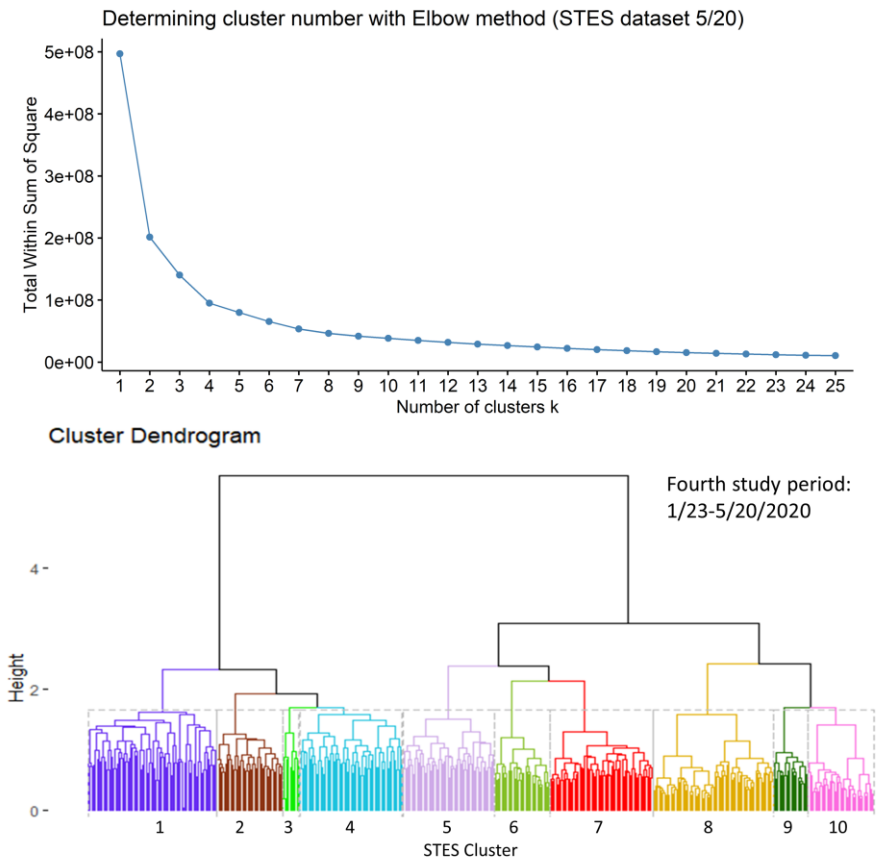


Figure 5.11. Elbow method evaluation and hierarchical clustering results for the 4th period. Notice that the numberings and colors of STES clusters match with those of corresponding clusters on the map and the temporal trend graph in Figure 5.12.

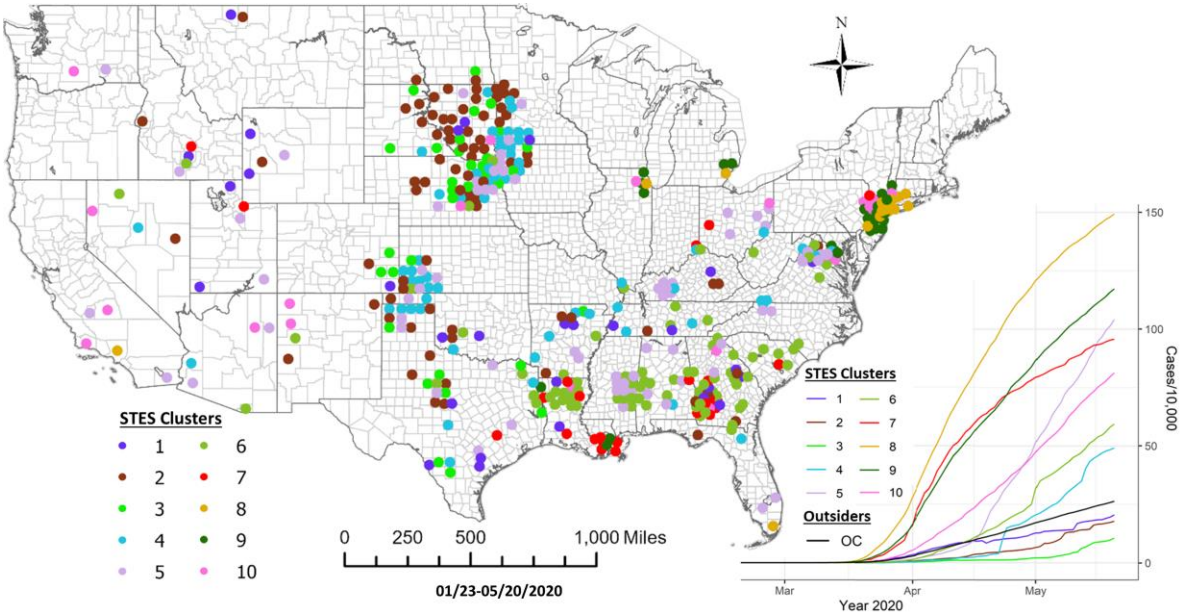


Figure 5.12. **Sequence similarity-based COVID-19 clusters along with average temporal trends at county level during 1/22/-5/20/2020.** This map includes the counties with higher relative risk (RR>1) contained in all the clusters detected by scan statistics in Figure 5.10. The average temporal trends of cumulative cases for STES clusters 1-10 on the map appear at the bottom right. Notice that the colors of STES clusters match with correspondingly colored dots on the map and with the colors of the STES cluster curves on the graph. OC includes all counties not included in the clusters.

5.4. Discussion

For this study we compared two approaches for COVID-19 surveillance. In combination, the two approaches provide complementary views that can offer a more comprehensive picture of surveillance information to further aid public health analysis and monitoring. The space-time scan statistic identifies emerging clusters as locations where the observed number of cases most exceeds the expected number of cases in space-time based on the underlying population. This approach provokes questions of why the disease is emerging at at such a location during a period of time. For disease progression, where the temporal pattern is equally important, similarity in the sequence of daily incidence rates

adds valuable information as it points to locations where the disease is progressing in a similar fashion. This view provokes questions of why these sometimes spatially dispersed locations are behaving in a similar way.

An initial working hypothesis for the STES sequence similarity metric in an environmental monitoring context was that locations that are spatially close are more likely to exhibit similar event sequences. While this is born out in some instances in this pandemic context, we found that in all study periods, similar sequence patterns of COVID-19 cases can be quite spatially separated. This result suggests that spatial proximity is not always a driver of sequence similarity. It has been reported that socio-economic or demographic characteristics could explain the different transmission rates or patterns between communities and locations (Dowd et al., 2020). Because members of these clusters share similar temporal disease progressions, questions arise as to whether they share some similar underlying characteristics such as similar population density, similar populations at risk, similar changes in surveillance programs, or possibly similar intervention strategies at work.

Sequence similarity Cluster 3 in the first study period which covers the first appearance of COVID-19 in the US shows the earliest and fastest accumulating number of cases suggesting initial points of entry. As members of this cluster include Snohomish and King counties in Washington State, several California counties in the San Francisco Bay area, and Bronx, Kings, Queens, Wassau and New York counties in New York state these do align with the known entry points on the east and west coasts. Seemingly unusual members in this cluster are Johnson County Iowa; Kershaw County, South Carolina; Williamson, Tennessee; and Douglas, Nebraska. An interesting question is why this last

subgroup of locations shares a similar profile with the coastal points of entry. Sequence similarity-based Cluster 2 in the first period is another interesting collection which is very spatially dispersed. Most of the members are rural communities that include Sheridan Wyoming, Davison South Dakota, Jackson Oklahoma, Hancock Indiana, Pitkin Colorado, Caddo Louisiana and Pierce Wisconsin. The temporal profile for this group is initially flat until mid-March at which point it shows a very rapid accumulation of cases. Such spatially dispersed cluster members that exhibit similar behaviours are targets for further investigation of potential contextual similarities. Of particular interest from epidemiological and health policy perspectives are spatially dispersed cluster members that exhibit similar flattening or decreasing patterns as these would be interesting to explore to understand if they have similar demographic characteristics or if they shared similar intervention measures.

We note that the sequence similarity clusters suggest some connections which are not conveyed by the scan statistic clusters. For example, in the third study period the scan statistic results indicate several new clusters. An examination of the sequence similarity clusters in this period indicate that several members of Cluster 10 were first nation or tribal reservations. In other words, several of the spatially dispersed reservations across the west showed a similar onset and progression in COVID-19 cases.

Another difference between the two approaches is that the sequence similarity-based clusters starting in the third period begin to show evidence of a spatial diffusion effect. For example, members of Cluster 8 with the earliest and fastest accumulating sequence similarity often appear to be surrounded by or in close spatial association with

the next closest lagging group, Cluster 9. A similar pattern appears between Cluster 8 and Cluster 9 members in the fourth study period.

Recent research has pointed to different continents of origin for the introduction of COVID-19 into the US (Gonzalez-Reiche et al., 2020; Worobey et al., 2020). Genomic epidemiology research supports the belief that isolates from China primarily seeded the original COVID-19 outbreak on the US West Coast and that European isolates seeded the pandemic in New York (and the US East Coast) (Deng et al., 2020). Given some connectivity suggested by the sequence similarity based approach there may exist opportunities for productive combination with phylogenetic tracing and transmission pathway studies (Zhang et al., 2020).

We recognize that both approaches can be impacted by limitations in data collection. Several publications have noted reporting lags although these are most problematic with respect to death reports rather than daily reported case counts (Aliprantis and Tauber, 2020; Angelopoulos et al., 2020; Casella, 2020; Kogan et al., 2021). There is clearly the potential for inaccuracies in data collection covering many different jurisdictions. If for example, reports of new cases are delayed by a day or two from a jurisdiction this could potentially change the similarity in the sequences of county daily case counts. However, given the length of the study periods here we expect lags of one to two days to have minor impact.

Chapter References

Abadi, D., Madden, S. and Lindner, W. (2016) Data Stream Management, pp. 409-428, Springer.

Alamo, T., Reina, D.G., Mammarella, M. and Abella, A. Open data resources for fighting covid-19. arXiv 2004.06111 [Preprint], 2020 [posted 2020 Apr 2013; last revised 2020 May 2011; cited 2020 Sept 2010]. Available from: https://arxiv.org/abs/2004.06111.

Ali, F., Rasoolimanesh, S.M., Sarstedt, M., Ringle, C.M. and Ryu, K. 2018. An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. International Journal of Contemporary Hospitality Management.

Aliprantis, D. and Tauber, K. 2020. Measuring deaths from COVID-19. Economic Commentary 18, 1-7.

André-Jönsson, H. and Badal, D.Z. 1997 Using signature files for querying time-series data, pp. 211-220, Springer.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M. and Poelitz, C. 2010 Extracting events from spatial time series, pp. 48-53, IEEE.

Angelopoulos, A.N., Pathak, R., Varma, R. and Jordan, M.I. 2020. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. Harvard Data Science Review Special Issue 1-COVID-19.

Ayub, M., Ghazanfar, M.A., Maqsood, M. and Saleem, A. 2018, pp. 1-6 A Jaccard base similarity measure to improve performance of CF based recommender systems, pp. 1-6.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2014) Hierarchical modeling and analysis for spatial data, Crc Press.

Bartariya, S. and Rastogi, A. 2016. Security in wireless sensor networks: Attacks and solutions. environment 5(3).

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

Beard, K., Emerson, J., Deese, H.E., Rude, A., Scott, M. and Pettigrew, N.R. 2011. Use of the EventViewer for visualizing and exploring events extracted from Ocean Observing System Data. Marine Technology Society Journal 45(1), 112-124.

Behmann, J., Hendriksen, K., Mцller, U., Bцscher, W. and Plцmer, L. 2016. Support Vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors. Geoinformatica 20(4), 693-714.

Berndt, D.J. and Clifford, J. 1994 Using dynamic time warping to find patterns in time series, pp. 359-370, Seattle, WA.

Bershad, B., Draves, R.P. and Forin, A. 1992 Using microbenchmarks to evaluate system performance, pp. 148-153, IEEE.

Bollobas, B., Das, G., Gunopulos, D. and Mannila, H. 1997 Time-series similarity problems and well-separated geometric sets, pp. 454-456.

Brézillon, P. and Gonzalez, A.J. (2014) Context in computing: a cross-disciplinary approach for modeling the real world, Springer.

Cardell-Oliver, R., Smettem, K., Kranz, M. and Mayer, K. 2004. Field testing a wireless sensor network for reactive environmental monitoring.

Casella, F. 2020. Can the COVID-19 epidemic be controlled on the basis of daily test reports? IEEE Control Systems Letters 5(3), 1079-1084.

Castellarin, A., Burn, D. and Brath, A. 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology 241(3-4), 270-285.

Chao, Y.-C.E., Zhao, Y., Kupper, L.L. and Nylander-French, L.A. 2008. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. Journal of occupational and environmental hygiene 5(8), 519-529.

Choi, S.-S., Cha, S.-H. and Tappert, C.C. 2010. A survey of binary similarity and distance measures. Journal of systemics, cybernetics and informatics 8(1), 43-48.

Chung, N.C., Miasojedow, B., Startek, M. and Gambin, A. 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC bioinformatics 20(15), 1-11.

Cutter, S.L. 1996. Vulnerability to environmental hazards. Progress in human geography 20(4), 529-539.

Danon, L., Brooks-Pollock, E., Bailey, M. and Keeling, M.J. A spatial model of CoVID-19 transmission in England and Wales: early spread and peak timing. medRxiv [Preprint], 2020 medRxiv 20022566 [posted 20022020 Feb 20022514; cited

20022020 Sept 20022510]. Available from: https://www.medrxiv.org/content/20022510.20021101/20022020.20022502.20022512.20022566v20022561.

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 14(7), 1394-1403.

Delbosc, A. and Currie, G. 2011. The spatial context of transport disadvantage, social exclusion and well-being. Journal of Transport Geography 19(6), 1130-1137.

Deng, X., Gu, W., Federman, S., Du Plessis, L., Pybus, O.G., Faria, N.R., Wang, C., Yu, G., Bushnell, B. and Pan, C.-Y. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science 369(6503), 582-587.

Desjardins, M.R., Hohl, A. and Delmelle, E.M. 2020. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Applied Geography 118, 102202.

Dey, A.K. 2001. Understanding and using context. Personal and ubiquitous computing 5(1), 4-7.

Dong, E., Du, H. and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infectious Diseases 20(5), 533-534.

Dong, J., Wang, G., Yan, H., Xu, J. and Zhang, X. 2015. A survey of smart water quality monitoring system. Environmental Science and Pollution Research 22, 4893-4906.

Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y. and Mills, M.C. 2020. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences 117(18), 9696-9698.

Du, F., Shneiderman, B., Plaisant, C., Malik, S. and Perer, A. 2016. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. IEEE transactions on visualization and computer graphics 23(6), 1636-1649.

Duczmal, L.H., Moreira, G.J., Burgarelli, D., Takahashi, R.H., Magalhães, F.C. and Bodevan, E.C. 2011. Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. International Journal of Health Geographics 10(1), 29.

Dutta, K. and Jayapal, M. 2015 Big data analytics for real time systems, pp. 1-13.

Elliot, P., Wakefield, J.C., Best, N.G. and Briggs, D. (2000) Spatial epidemiology: methods and applications, Oxford University Press.

Elliott, P. and Wartenberg, D. 2004. Spatial epidemiology: current approaches and future challenges. Environmental health perspectives 112(9), 998.

Fu, T.-c. 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence 24(1), 164-181.

Gehani, N.H., Jagadish, H.V. and Shmueli, O. 1992. Event specification in an active object-oriented database. ACM sigmod record 21(2), 81-90.

Goldin, D.Q. and Kanellakis, P.C. 1995 On similarity queries for time-series data: constraint specification and implementation, pp. 137-153, Springer.

Gong, H. and Hassink, R. 2020. Context sensitivity and economic-geographic (re) theorising. Cambridge Journal of Regions, Economy and Society 13(3), 475-490.

Gonzalez-Reiche, A.S., Hernandez, M.M., Sullivan, M.J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J. and Khan, Z. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369(6501), 297-301.

Gripenberg, S. and Roslin, T. 2007. Up or down in space? Uniting the bottom‐up versus top‐down paradigm and spatial ecology. Oikos 116(2), 181-188.

Gundersen, O.E. 2012 Toward measuring the similarity of complex event sequences in real-time, pp. 107-121, Springer.

Guralnik, V. and Srivastava, J. 1999 Event detection from time series data, pp. 33-42, ACM.

Gustriansyah, R., Suhandi, N. and Antony, F. 2020. Clustering optimization in RFM analysis based on k-means. Indones. J. Electr. Eng. Comput. Sci 18(1), 470-477.

Hamming, R.W. 1950. Error detecting and error correcting codes. Bell System technical journal 29(2), 147-160.

Hasan, A., Teymourian, K. and Paschke, A. 2015 Probabilistic event pattern discovery, pp. 241-257, Springer.

He, W., Yi, G.Y. and Zhu, Y. 2020. Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for

COVID-19: Meta-analysis and sensitivity analysis. Journal of Medical Virology 92(11), 2543-2550.

Heitz, G. and Koller, D. 2008 Learning spatial context: Using stuff to find things, pp. 30-43, Springer.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. and Caron, E. 2016. A survey of event extraction methods from text for decision support systems. Decision Support Systems 85, 12-22.

Hohl, A., Delmelle, E., Desjardins, M. and Lan, Y. 2020. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. Spatial and Spatio-temporal Epidemiology, 100354.

Hu, H. 2016. Online Near Real-time Mine Disaster Monitoring System Based on Wireless Sensor Networks. International Journal of Online Engineering 12(3).

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J. and Gu, X. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223), 497-506.

Hughes, K.A. 2003. Influence of seasonal environmental variables on the distribution of presumptive fecal coliforms around an Antarctic research station. Applied and Environmental Microbiology 69(8), 4884-4891.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37, 547-579.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. New phytologist 11(2), 37-50.

Jacobs, B.E. and Walczak, C.A. 1983. A generalized query-by-example data manipulation language based on database logic. IEEE Transactions on Software Engineering (1), 40-57.

Jassby, A.D. and Powell, T.M. 1990. Detecting changes in ecological time series. Ecology 71(6), 2044-2052.

Jiang, B. and Yao, X. (2007) Location based services and telecartography, pp. 27-45, Springer.

Jones, R.C., Liberatore, M., Fernandez, J.R. and Gerber, S.I. 2006. Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. Public Health Reports 121(2), 133-139.

Kamarinas, I., Julian, J.P., Hughes, A.O., Owsley, B.C. and De Beurs, K.M. 2016. Nonlinear changes in land cover and sediment runoff in a New Zealand catchment dominated by plantation forestry and livestock grazing. Water 8(10), 436.

Keßler, C. 2007 Similarity measurement in context, pp. 277-290, Springer.

Keßler, C., Raubal, M. and Janowicz, K. 2007 The effect of context on semantic similarity measurement, pp. 1274-1284, Springer.

Kettenring, J.R. 2006. The practice of cluster analysis. Journal of classification 23(1), 3-30.

Khan, D., Rossen, L.M., Hamilton, B.E., He, Y., Wei, R. and Dienes, E. 2017. Hot spots, cluster detection and spatial outlier analysis of teen birth rates in the U.S., 2003-2012. Spatial and Spatio-temporal Epidemiology 21, 67-75.

Kogan, N.E., Clemente, L., Liautaud, P., Kaashoek, J., Link, N.B., Nguyen, A.T., Lu, F.S., Huybers, P., Resch, B., Havas, C., Petutschnig, A., Davis, J., Chinazzi, M., Mustafa, B., Hanage, W.P., Vespignani, A. and Santillana, M. 2021. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. Science Advances 7(10), eabd6989.

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

Kulldorff, M. (1999) Scan statistics and applications, pp. 303-322, Springer.

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

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. and Mostashari, F. 2005a. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2(3), e59.

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

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. and Platt, R. 2007. Multivariate scan statistics for disease surveillance. Statistics in Medicine 26(8), 1824-1833.

Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M.N.K. and Weller, A. 2020. Leveraging data science to combat covid-

19: A comprehensive review. IEEE Transactions on Artificial Intelligence 1(1), 85-103.

Lee, O.-J. and Jung, J.E. 2017. Sequence clustering-based automated rule generation for adaptive complex event processing. Future Generation Computer Systems 66, 100-109.

Leskovec, J., Rajaraman, A. and Ullman, J.D. (2014) Mining of massive datasets, Cambridge university press.

Levenshtein, V.I. 1966 Binary codes capable of correcting deletions, insertions, and reversals, pp. 707-710.

Levin, S.A. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology 73(6), 1943-1967.

Li, M., Shi, X., Li, X., Ma, W., He, J. and Liu, T. 2019. Sensitivity of disease cluster detection to spatial scales: an analysis with the spatial scan statistic method. International Journal of Geographical Information Science 33(11), 2125-2152.

Loke, S. (2006) Context-aware pervasive systems: architectures for a new breed of applications, CRC Press.

Luckham, D. 2016. Event Processing Glossary-Version 2.0, Event Processing Technical Society. http://www. ep-ts. com/component/option, com_docman/task, doc_download/gid, 66/Itemid, 84/.

Luckham, D. and Schulte, R. 2011. EPTS Event Processing Glossary v2. 0. Event Processing Technical Society.

Luckham, D.C. (2001) The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley Longman Publishing Co., Inc.

Luckham, D.C. and Frasca, B. 1998. Complex event processing in distributed systems. Computer Systems Laboratory Technical Report CSL-TR-98-754. Stanford University, Stanford 28.

Lupiani, E., Sauer, C., Roth-Berghofer, T., Juarez, J.M. and Palma, J. 2013 Implementation of similarity measures for event sequences in myCBR.

Luu, V.-T., Forestier, G., Weber, J., Bourgeois, P., Djelil, F. and Muller, P.-A. 2020. A review of alignment based similarity measures for web usage mining. Artificial Intelligence Review 53(3), 1529-1551.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R. and Anderson, J. 2002 Wireless sensor networks for habitat monitoring, pp. 88-97, Acm.

Malpas, J. 2012. Putting space in place: Philosophical topography and relational geography. Environment and planning D: society and space 30(2), 226-242.

Mannila, H. and Moen, P. 1999 Similarity between event types in sequences, pp. 271-280, Springer.

Mannila, H. and Ronkainen, P. 1997 Similarity of event sequences, pp. 136-139, IEEE.

Mannila, H. and Salmenkivi, M. 2001 Finding simple intensity descriptions from event sequence data, pp. 341-346, ACM.

Mannila, H., Toivonen, H. and Verkamo, A.I. 1995 Discovering frequent episodes in sequences extended abstract.

Marston, S.A., Jones III, J.P. and Woodward, K. 2005. Human geography without scale. Transactions of the institute of British geographers 30(4), 416-432.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V. 2007. Fault diagnosis using dynamic trend analysis: A review and recent developments. Engineering Applications of artificial intelligence 20(2), 133-146.

Mirbagheri, S.M. and Hamilton, H.J. 2020 Similarity Matching of Temporal Event-Interval Sequences, pp. 420-425, Springer.

Moen, P. 2000. Attribute, event sequence, and event type similarity notions for data mining. PhD thesis, University of Helsinki.

Moore, A. 2008. Rethinking scale as a geographical category: from analysis to practice. Progress in human geography 32(2), 203-225.

Moorthy, V., Restrepo, A.M.H., Preziosi, M.-P. and Swaminathan, S. 2020. Data sharing for novel coronavirus (COVID-19). Bulletin of the World Health Organization 98(3), 150.

Nittel, S. 2009. A survey of geosensor networks: Advances in dynamic environmental monitoring. Sensors 9(7), 5664-5678.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S. 2013 Using of Jaccard coefficient for keywords similarity.

Noble, R.T., Moore, D.F., Leecaster, M.K., McGee, C.D. and Weisberg, S.B. 2003. Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing. Water research 37(7), 1637-1643.

Obweger, H., Suntinger, M., Schiefer, J. and Raidl, G. 2010 Similarity searching in sequences of complex events, pp. 631-640, IEEE.

Oliveira, L.M. and Rodrigues, J.J. 2011. Wireless Sensor Networks: A Survey on Environmental Monitoring. JCM 6(2), 143-151.

Orozco, C.V., Tonini, M., Conedera, M. and Kanveski, M. 2012. Cluster recognition in spatial-temporal sequences: the case of forest fires. Geoinformatica 16(4), 653-673.

Othman, M.F. and Shazali, K. 2012. Wireless sensor network applications: A study in environment monitoring system. Procedia Engineering 41, 1204-1210.

Paasi, A. 2004. Place and region: looking through the prism of scale. Progress in human geography 28(4), 536-546.

Padhy, P., Martinez, K., Riddoch, A., Ong, H. and Hart, J.K. 2005. Glacial environment monitoring using sensor networks.

Paschke, A. and Boley, H. (2009) Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, pp. 215-252, IGI Global.

Peterson, M.R., Doom, T.E. and Raymer, M.L. 2005 Ga-facilitated knn classifier optimization with varying similarity measures, pp. 2514-2521, IEEE.

Peuquet, D.J. and Duan, N. 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International journal of geographical information systems 9(1), 7-24.

Prasad, A., Mamun, K.A., Islam, F. and Haqva, H. 2015 Smart water quality monitoring system, pp. 1-6, IEEE.

Prasath, V., Alfeilat, H.A.A., Lasassmeh, O. and Hassanat, A. 2017. Distance and similarity measures effect on the performance of K-nearest neighbor classifier-a review. arXiv preprint arXiv:1708.04321.

Prinzie, A. and Van den Poel, D. 2011. Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: an Acquisition Pattern Analysis application. Journal of Intelligent Information Systems 36(3), 283-304.

Qi, H., Xiao, S., Shi, R., Ward, M.P., Chen, Y., Tu, W., Su, Q., Wang, W., Wang, X. and Zhang, Z. 2020. COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. Science of the Total Environment, 138778.

Ros, F. and Guillaume, S. 2019. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Systems with Applications 128, 96-108.

Roux, A.V.D. and Mair, C. 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186(1), 125-145.

Rude, A. and Beard, K. 2012 High-Level Event Detection in Spatially Distributed Time Series, pp. 160-172, Springer Berlin Heidelberg, Berlin, Heidelberg.

Saglietto, A., D'Ascenzo, F., Zoccai, G.B. and De Ferrari, G.M. 2020. COVID-19 in Europe: the Italian lesson. Lancet 395(10230), 1110-1111.

Sampson, R.J. 2003. The neighborhood context of well-being. Perspectives in biology and medicine 46(3), S53-S64.

Schatzki, T.R. (2002) The site of the social: A philosophical account of the constitution of social life and change, Penn State University Press.

Shahar, Y. 1997. A framework for knowledge-based temporal abstraction. Artificial intelligence 90(1-2), 79-133.

Sheil, H., Rana, O. and Reilly, R. 2018. Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks. arXiv preprint arXiv:1807.08207.

Shekhar, S., Evans, M.R., Kang, J.M. and Mohan, P. 2011. Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3), 193-214.

Shurkhovetskyy, G., Andrienko, N., Andrienko, G. and Fuchs, G. 2018 Data abstraction for visualizing large time series, pp. 125-144, Wiley Online Library.

Simandan, D. 2019. Revisiting positionality and the thesis of situated knowledge. Dialogues in human geography 9(2), 129-149.

Simandan, D. 2020. Being surprised and surprising ourselves: a geography of personal and social change. Progress in Human Geography 44(1), 99-118.

Sinclair, C., Pierce, L. and Matzner, S. 1999 An application of machine learning to network intrusion detection, pp. 371-377, IEEE.

Singhal, A., Luo, J. and Zhu, W. 2003 Probabilistic spatial context models for scene content understanding, pp. I-I, IEEE.

Smith, T.F. and Waterman, M.S. 1981. Comparison of biosequences. Advances in applied mathematics 2(4), 482-489.

Stehle, S. and Peuquet, D.J. 2015. Analyzing spatio-temporal patterns and their evolution via sequence alignment. Spatial Cognition & Computation 15(2), 68-85.

Sun, S.-B., Zhang, Z.-H., Dong, X.-L., Zhang, H.-R., Li, T.-J., Zhang, L. and Min, F. 2017. Integrating Triangle and Jaccard similarities for recommendation. PloS One 12(8), e0183570.

Sunley, P. 1996. Context in economic geography: the relevance of pragmatism. Progress in Human Geography 20(3), 338-355.

Syakur, M., Khotimah, B., Rochman, E. and Satoto, B. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering 336(1), 012017.

Tao, C., Wongsuphasawat, K., Clark, K., Plaisant, C., Shneiderman, B. and Chute, C.G. 2012 Towards event sequence representation, reasoning and visualization for EHR data, pp. 801-806, ACM.

Tilman, D. and Kareiva, P. (2018) Spatial ecology: the role of space in population dynamics and interspecific interactions (MPB-30), Princeton University Press.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A.S., Kurose, S. and Zandee, R. 2003. Spatial context and the complexity of daily travel patterns: an international comparison. Journal of Transport Geography 11(1), 37-46.

Tonidandel, S. and LeBreton, J.M. 2015. RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. Journal of Business and Psychology 30, 207-216.

Tonidandel, S., LeBreton, J.M. and Johnson, J.W. 2009. Determining the statistical significance of relative weights. Psychological methods 14(4), 387.

Tonini, M., Tuia, D. and Ratle, F. 2009. Detection of clusters using space–time scan statistics. International journal of wildland fire 18(7), 830-836.

Vanderbilt, K.L., Lin, C.-C., Lu, S.-S., Kassim, A.R., He, H., Guo, X., San Gil, I., Blankman, D. and Porter, J.H. 2015. Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network. Ecosphere 6(10), 1-18.

Vorontsov, I.E., Kulakovskiy, I.V. and Makeev, V.J. 2013. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms for Molecular Biology 8(1), 1-11.

Vrotsou, K. (2010) Everyday mining: Exploring sequences in event-based data, Linköping University Electronic Press.

Vrotsou, K. and Forsell, C. 2011 A qualitative study of similarity measures in event-based data, pp. 170-179, Springer.

Wang, T.-Y., Yang, M.-H. and Wu, J.-Y. 2016a. Distributed Detection of Dynamic Event Regions in Sensor Networks With a Gibbs Field Distribution and Gaussian Corrupted Measurements. IEEE Transactions on Communications 64(9), 3932-3945.

Wang, W., Hu, C., Chen, N., Xiao, C. and Jia, S. 2016b. Spatio-Temporal Risk Assessment Process Modeling for Urban Hazard Events in Sensor Web Environment. ISPRS International Journal of Geo-Information 5(11), 203.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. and Keogh, E. 2013. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26(2), 275-309.

Weber, J. and Kwan, M.-P. 2003. Evaluating the effects of geographic contexts on individual accessibility: a multilevel Approach1. Urban Geography 24(8), 647-671.

Weiss, G.M. and Hirsh, H. 1998 Learning to Predict Rare Events in Event Sequences, pp. 359-363.

Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J. and Welsh, M. 2006. Deploying a wireless sensor network on an active volcano. IEEE internet computing 10(2), 18-25.

Wolpert, J. 1964. The decision process in spatial context. Annals of the Association of American Geographers 54(4), 537-558.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. and Shneiderman, B. 2012. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with computers 24(2), 55-68.

Wongsuphasawat, K. and Shneiderman, B. 2009 Finding comparable temporal categorical records: A similarity measure with an interactive visualization, pp. 27-34, IEEE.

Woodward, K., Jones III, J.P. and Marston, S.A. 2012. The politics of autonomous space. Progress in Human Geography 36(2), 204-224.

Worboys, M. 2005. Event-oriented approaches to geographic phenomena. International Journal of Geographical Information Science 19(1), 1-28.

Worboys, M. and Hornsby, K. 2004 From objects to events: GEM, the geospatial event model, pp. 327-343, Springer.

Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O. and Lemey, P. 2020. The emergence of SARS-CoV-2 in Europe and North America. Science 370(6516), 564-570.

Xu, F. and Beard, K. 2021. A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. ISPRS International Journal of Geo-Information 10(9), 594.

Yang, D.-H., Goerge, R. and Mullner, R. 2006. Comparing GIS-based methods of measuring spatial accessibility to health services. Journal of medical systems 30(1), 23-32.

Yang, J., McAuley, J., Leskovec, J., LePendu, P. and Shah, N. 2014 Finding progression stages in time-evolving event sequences, pp. 783-794, ACM.

Yeh, C.-C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Zimmerman, Z., Silva, D.F., Mueen, A. and Keogh, E. 2018. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Mining and Knowledge Discovery 32(1), 83-123.

Yih, W.K., Deshpande, S., Fuller, C., Heisey-Grove, D., Hsu, J., Kruskal, B.A., Kulldorff, M., Leach, M., Nordin, J. and Patton-Levine, J. 2010. Evaluating real-time syndromic surveillance signals from ambulatory care data in four states. Public Health Reports 125(1), 111-120.

Yin, F., Li, X., Ma, J. and Feng, Z. 2007. The early warning system based on the prospective space-time permutation statistic. Wei Sheng Yan Jiu (in Chinese: Journal of Hygiene Research) 36(4), 455-458.

Yin, J., Hu, D.H. and Yang, Q. 2009 Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields, pp. 1321-1327.

Zambelli, A.E. 2016. A data-driven approach to estimating the number of clusters in hierarchical clustering. F1000Research 5.

Zhang, W., Govindavari, J.P., Davis, B.D., Chen, S.S., Kim, J.T., Song, J., Lopategui, J., Plummer, J.T. and Vail, E. 2020. Analysis of genomic characteristics and transmission routes of patients with confirmed SARS-CoV-2 in Southern California during the early stage of the US COVID-19 pandemic. JAMA Network Open 3(10), e2024191-e2024191.

Zolnik, E.J. 2009. Context in human geography: a multilevel approach to study human–environment interactions. The Professional Geographer 61(3), 336-349.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The studies in this dissertation have highlighted the importance of measuring similarity or differences between data objects, particularly the sequences of spatiotemporal events and associated settings. The importance and application of measuring similarity is evident in various fields such as geography, biology, computer science, linguistics, logic, business analytics, and statistics. The ability of using appropriate similarity measures to compare different event sequences from corresponding event settings and identify patterns can provide insight into underlying processes and potentially inform decision-making. In spatiotemporal analysis, measuring similarity or differences between events, settings, or sequences of events or settings can help enhance the understanding of processes over time and geospatial locations.

6.1. Conclusions

6.1.1. Matrix-Based STES Representation and Similarity Measure

This dissertation has developed a novel framework and matrix-based spatiotemporal event sequence representation that unifies punctual and interval-based representation of events. The dissertation proposed STES similarity measure has been demonstrated to be effective in two real-world applications for analyzing spatiotemporal event sequences extracted from space-time series of water quality monitoring systems and the COVID-19 case progression dataset in USA for 4 months of 2020. The proposed novel

similarity measure for spatiotemporal event sequences (STES) has several advantages over Edit Distance, which is a commonly used similarity measure for event sequence data.

First, the STES similarity measure is based on the modified Jaccard index with temporal order constraints and is designed specifically for spatiotemporal event sequences extracted from space-time series, a common data type in many environmental applications among others with spatiotemporal constraints. In contrast, Edit Distance is a more general similarity measure for sequence data and does not consider spatiotemporal structures in the data.

Second, the STES similarity measure is suitable for STES of both punctual and interval events with the option of considering quantitative levels of individual events and filling the gap for investigating similarities between STES of different types of events.

Third, a unified matrix-based representation of the spatiotemporal event sequence followed by the matrix of pairwise similarities offers more efficient and faster operation and computation compared to the Edit Distance. This matrix-based representation contributes to a flexible toolbox for efficient data mining techniques, such as clustering and classification.

6.1.2. Matrix-Based Representation of Setting Sequences and Similarity Measure

This dissertation has modeled spatial-temporal event settings and developed a new similarity measure for event setting sequences that incorporates dynamic variables alongside static variables. This approach considers spatial and temporal scales to define the bounds of the setting and uses a matrix-based representation and an extended Jaccard index that allows for the use of all variable data types. This approach is successfully applied in a

case study involving setting sequences and pollution event sequences associated with the same monitoring stations. The working hypothesis evaluated by this study is that more similar event sequence settings give rise to more similar event sequences. The case study results suggest evidence in favor of this hypothesis. Cluster analysis of similar spatial-temporal settings or setting sequences, related directly or indirectly to events and event sequences of interest, show alignment with clusters of similar event sequences. As spatial location plays a major role in the concept of settings the case study results also show evidence that STES that are closer in space tend to be more similar.

6.1.3. Scalability and Its Extended Application of STES Similarity Measure

This study successfully implemented the new developed STES similarity measure to detect hotspots or clusters of COVID-19 in a large dataset of tracking COVID-19 cases nationwide at county level. The STES-based approach adapted for this pandemic context computes the similarity of evolving normalized COVID-19 daily cases by county and clusters them to identify counties with similarly evolving COVID-19 case histories. The prospective space-time scan statistic has been used to identify emerging disease clusters, but it can encounter strategic limitations imposed by the spatial constraints of the scanning window. The two approaches identify different patterns in the disease spread, and their results can complement each other and aid in tracking the progression of the pandemic. By comparing two different approaches to identifying emerging disease clusters, researchers can better track the progression of the pandemic and aid in the development of effective public health responses and policy actions.

6.2. Future Work

Future work could focus on expanding the applicability of the proposed similarity measures in this study for spatiotemporal event sequences to a broader range of real-world applications. The framework presented for a novel matrix-based spatiotemporal event sequence representation could be further developed to accommodate even more event data types, with a focus on making the approach more robust to noisy or missing data. This could be achieved through the integration of machine learning techniques, such as deep learning or reinforcement learning, to improve the accuracy and reliability of the similarity measures.

In addition, further research could investigate the potential of using the proposed similarity measures to analyze the dynamics of complex systems, such as ecological or economic systems, where events and their settings or contexts can be critical factors in understanding the system behavior. By examining the similarity of event settings, researchers could gain insight into how different factors interact with each other over time and across different spatial scales, which could inform better decision-making in a wide range of fields, from urban planning to disaster management.

Furthermore, to extend the use of STES similarity measures to analyze lagged events and event sequences, we can incorporate a time lag parameter into the similarity calculation. This would allow us to quantify the similarity between events that are temporally separated by a fixed lag. One approach to incorporating time lags is to create lagged versions of the original event sequence and compute the similarity measures between the lagged sequences. For example, to compute the similarity between event sequences separated by a lag of one-time step, we can create two lagged sequences, one

with all events shifted by one time step, and one with the original events. This approach can be extended to analyze event sequences with longer lags by creating additional lagged sequences and computing the similarity measures between them. The lagged sequences can be created by shifting the events by a fixed number of time steps or by using sliding windows of fixed size. Another approach is to use dynamic time warping (DTW), a commonly used technique for comparing time series with temporal distortions. DTW can be used to align event sequences with different temporal offsets, allowing us to compute similarity measures even when events occur at different times in the two sequences.

Finally, future work could also explore the potential of combining the proposed similarity measures with other analytical tools, such as social network analysis or geographic information systems (GIS), to gain a more comprehensive understanding of the relationships between events, their associated settings or contexts, and the broader social and physical environments in which they occur. This could lead to the development of new methods for detecting and tracking emerging patterns of behavior or disease spread, and for designing more effective interventions to address these issues.

Overall, all studies conducted in this dissertation demonstrate the importance of measuring similarity or differences between event or setting data objects in spatiotemporal analysis. The developed similarity measures and frameworks in this dissertation offer researchers powerful tools for understanding different factors and their dynamics corresponding to occurrences of spatiotemporal event sequences. These similarity measures have many potential real-world applications and can inform decision-making in various fields, including public health, environmental monitoring, and resource allocation. The development of novel similarity measures that can accommodate different event data

types and incorporate dynamic variables alongside static variables can provide valuable insights into underlying processes and enhance our understanding of spatiotemporal phenomena.

REFERENCES

Abadi, D., Madden, S. and Lindner, W. (2016) Data Stream Management, pp. 409-428, Springer.

Alamo, T., Reina, D.G., Mammarella, M. and Abella, A. Open data resources for fighting covid-19. arXiv 2004.06111 [Preprint], 2020 [posted 2020 Apr 2013; last revised 2020 May 2011; cited 2020 Sept 2010]. Available from: https://arxiv.org/abs/2004.06111.

Ali, F., Rasoolimanesh, S.M., Sarstedt, M., Ringle, C.M. and Ryu, K. 2018. An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. International Journal of Contemporary Hospitality Management.

Aliprantis, D. and Tauber, K. 2020. Measuring deaths from COVID-19. Economic Commentary 18, 1-7.

André-Jönsson, H. and Badal, D.Z. 1997 Using signature files for querying time-series data, pp. 211-220, Springer.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M. and Poelitz, C. 2010 Extracting events from spatial time series, pp. 48-53, IEEE.

Angelopoulos, A.N., Pathak, R., Varma, R. and Jordan, M.I. 2020. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. Harvard Data Science Review Special Issue 1-COVID-19.

Ayub, M., Ghazanfar, M.A., Maqsood, M. and Saleem, A. 2018, pp. 1-6 A Jaccard base similarity measure to improve performance of CF based recommender systems, pp. 1-6.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2014) Hierarchical modeling and analysis for spatial data, Crc Press.

Bartariya, S. and Rastogi, A. 2016. Security in wireless sensor networks: Attacks and solutions. environment 5(3).

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

Beard, K., Emerson, J., Deese, H.E., Rude, A., Scott, M. and Pettigrew, N.R. 2011. Use of the EventViewer for visualizing and exploring events extracted from Ocean Observing System Data. Marine Technology Society Journal 45(1), 112-124.

Behmann, J., Hendriksen, K., Mцller, U., Bцscher, W. and Plцmer, L. 2016. Support Vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors. Geoinformatica 20(4), 693-714.

Berndt, D.J. and Clifford, J. 1994 Using dynamic time warping to find patterns in time series, pp. 359-370, Seattle, WA.

Bershad, B., Draves, R.P. and Forin, A. 1992 Using microbenchmarks to evaluate system performance, pp. 148-153, IEEE.

Bollobas, B., Das, G., Gunopulos, D. and Mannila, H. 1997 Time-series similarity problems and well-separated geometric sets, pp. 454-456.

Brézillon, P. and Gonzalez, A.J. (2014) Context in computing: a cross-disciplinary approach for modeling the real world, Springer.

Cardell-Oliver, R., Smettem, K., Kranz, M. and Mayer, K. 2004. Field testing a wireless sensor network for reactive environmental monitoring.

Casella, F. 2020. Can the COVID-19 epidemic be controlled on the basis of daily test reports? IEEE Control Systems Letters 5(3), 1079-1084.

Castellarin, A., Burn, D. and Brath, A. 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology 241(3-4), 270-285.

Chao, Y.-C.E., Zhao, Y., Kupper, L.L. and Nylander-French, L.A. 2008. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. Journal of occupational and environmental hygiene 5(8), 519-529.

Choi, S.-S., Cha, S.-H. and Tappert, C.C. 2010. A survey of binary similarity and distance measures. Journal of systemics, cybernetics and informatics 8(1), 43-48.

Chung, N.C., Miasojedow, B., Startek, M. and Gambin, A. 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC bioinformatics 20(15), 1-11.

Cutter, S.L. 1996. Vulnerability to environmental hazards. Progress in human geography 20(4), 529-539.

Danon, L., Brooks-Pollock, E., Bailey, M. and Keeling, M.J. A spatial model of CoVID-19 transmission in England and Wales: early spread and peak timing. medRxiv [Preprint], 2020 medRxiv 20022566 [posted 20022020 Feb 20022514; cited 20022020 Sept 20022510]. Available from: https://www.medrxiv.org/content/20022510.20021101/20022020.20022502.20022512.20022566v20022561.

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 14(7), 1394-1403.

Delbosc, A. and Currie, G. 2011. The spatial context of transport disadvantage, social exclusion and well-being. Journal of Transport Geography 19(6), 1130-1137.

Deng, X., Gu, W., Federman, S., Du Plessis, L., Pybus, O.G., Faria, N.R., Wang, C., Yu, G., Bushnell, B. and Pan, C.-Y. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science 369(6503), 582-587.

Desjardins, M.R., Hohl, A. and Delmelle, E.M. 2020. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Applied Geography 118, 102202.

Dey, A.K. 2001. Understanding and using context. Personal and ubiquitous computing 5(1), 4-7.

Dong, E., Du, H. and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infectious Diseases 20(5), 533-534.

Dong, J., Wang, G., Yan, H., Xu, J. and Zhang, X. 2015. A survey of smart water quality monitoring system. Environmental Science and Pollution Research 22, 4893-4906.

Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y. and Mills, M.C. 2020. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences 117(18), 9696-9698.

Du, F., Shneiderman, B., Plaisant, C., Malik, S. and Perer, A. 2016. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. IEEE transactions on visualization and computer graphics 23(6), 1636-1649.

Duczmal, L.H., Moreira, G.J., Burgarelli, D., Takahashi, R.H., Magalhães, F.C. and Bodevan, E.C. 2011. Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. International Journal of Health Geographics 10(1), 29.

Dutta, K. and Jayapal, M. 2015 Big data analytics for real time systems, pp. 1-13.

Elliot, P., Wakefield, J.C., Best, N.G. and Briggs, D. (2000) Spatial epidemiology: methods and applications, Oxford University Press.

Elliott, P. and Wartenberg, D. 2004. Spatial epidemiology: current approaches and future challenges. Environmental health perspectives 112(9), 998.

Fu, T.-c. 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence 24(1), 164-181.

Gehani, N.H., Jagadish, H.V. and Shmueli, O. 1992. Event specification in an active object-oriented database. ACM sigmod record 21(2), 81-90.

Goldin, D.Q. and Kanellakis, P.C. 1995 On similarity queries for time-series data: constraint specification and implementation, pp. 137-153, Springer.

Gong, H. and Hassink, R. 2020. Context sensitivity and economic-geographic (re) theorising. Cambridge Journal of Regions, Economy and Society 13(3), 475-490.

Gonzalez-Reiche, A.S., Hernandez, M.M., Sullivan, M.J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J. and Khan, Z. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369(6501), 297-301.

Gripenberg, S. and Roslin, T. 2007. Up or down in space? Uniting the bottom-up versus top-down paradigm and spatial ecology. Oikos 116(2), 181-188.

Gundersen, O.E. 2012 Toward measuring the similarity of complex event sequences in real-time, pp. 107-121, Springer.

Guralnik, V. and Srivastava, J. 1999 Event detection from time series data, pp. 33-42, ACM.

Gustriansyah, R., Suhandi, N. and Antony, F. 2020. Clustering optimization in RFM analysis based on k-means. Indones. J. Electr. Eng. Comput. Sci 18(1), 470-477.

Hamming, R.W. 1950. Error detecting and error correcting codes. Bell System technical journal 29(2), 147-160.

Hasan, A., Teymourian, K. and Paschke, A. 2015 Probabilistic event pattern discovery, pp. 241-257, Springer.

He, W., Yi, G.Y. and Zhu, Y. 2020. Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis. Journal of Medical Virology 92(11), 2543-2550.

Heitz, G. and Koller, D. 2008 Learning spatial context: Using stuff to find things, pp. 30-43, Springer.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. and Caron, E. 2016. A survey of event extraction methods from text for decision support systems. Decision Support Systems 85, 12-22.

Hohl, A., Delmelle, E., Desjardins, M. and Lan, Y. 2020. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. Spatial and Spatio-temporal Epidemiology, 100354.

Hu, H. 2016. Online Near Real-time Mine Disaster Monitoring System Based on Wireless Sensor Networks. International Journal of Online Engineering 12(3).

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J. and Gu, X. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223), 497-506.

Hughes, K.A. 2003. Influence of seasonal environmental variables on the distribution of presumptive fecal coliforms around an Antarctic research station. Applied and Environmental Microbiology 69(8), 4884-4891.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37, 547-579.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. New phytologist 11(2), 37-50.

Jacobs, B.E. and Walczak, C.A. 1983. A generalized query-by-example data manipulation language based on database logic. IEEE Transactions on Software Engineering (1), 40-57.

Jassby, A.D. and Powell, T.M. 1990. Detecting changes in ecological time series. Ecology 71(6), 2044-2052.

Jiang, B. and Yao, X. (2007) Location based services and telecartography, pp. 27-45, Springer.

Jones, R.C., Liberatore, M., Fernandez, J.R. and Gerber, S.I. 2006. Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. Public Health Reports 121(2), 133-139.

Kamarinas, I., Julian, J.P., Hughes, A.O., Owsley, B.C. and De Beurs, K.M. 2016. Nonlinear changes in land cover and sediment runoff in a New Zealand catchment dominated by plantation forestry and livestock grazing. Water 8(10), 436.

Keßler, C. 2007 Similarity measurement in context, pp. 277-290, Springer.

Keßler, C., Raubal, M. and Janowicz, K. 2007 The effect of context on semantic similarity measurement, pp. 1274-1284, Springer.

Kettenring, J.R. 2006. The practice of cluster analysis. Journal of classification 23(1), 3-30.

Khan, D., Rossen, L.M., Hamilton, B.E., He, Y., Wei, R. and Dienes, E. 2017. Hot spots, cluster detection and spatial outlier analysis of teen birth rates in the U.S., 2003-2012. Spatial and Spatio-temporal Epidemiology 21, 67-75.

Kogan, N.E., Clemente, L., Liautaud, P., Kaashoek, J., Link, N.B., Nguyen, A.T., Lu, F.S., Huybers, P., Resch, B., Havas, C., Petutschnig, A., Davis, J., Chinazzi, M., Mustafa, B., Hanage, W.P., Vespignani, A. and Santillana, M. 2021. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. Science Advances 7(10), eabd6989.

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

Kulldorff, M. (1999) Scan statistics and applications, pp. 303-322, Springer.

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

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. and Mostashari, F. 2005a. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2(3), e59.

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

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. and Platt, R. 2007. Multivariate scan statistics for disease surveillance. Statistics in Medicine 26(8), 1824-1833.

Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M.N.K. and Weller, A. 2020. Leveraging data science to combat covid-19: A comprehensive review. IEEE Transactions on Artificial Intelligence 1(1), 85-103.

Lee, O.-J. and Jung, J.E. 2017. Sequence clustering-based automated rule generation for adaptive complex event processing. Future Generation Computer Systems 66, 100-109.

Leskovec, J., Rajaraman, A. and Ullman, J.D. (2014) Mining of massive datasets, Cambridge university press.

Levenshtein, V.I. 1966 Binary codes capable of correcting deletions, insertions, and reversals, pp. 707-710.

Levin, S.A. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology 73(6), 1943-1967.

Li, M., Shi, X., Li, X., Ma, W., He, J. and Liu, T. 2019. Sensitivity of disease cluster detection to spatial scales: an analysis with the spatial scan statistic method. International Journal of Geographical Information Science 33(11), 2125-2152.

Loke, S. (2006) Context-aware pervasive systems: architectures for a new breed of applications, CRC Press.

Luckham, D. 2016. Event Processing Glossary-Version 2.0, Event Processing Technical Society. http://www. ep-ts. com/component/option, com_docman/task, doc_download/gid, 66/Itemid, 84/.

Luckham, D. and Schulte, R. 2011. EPTS Event Processing Glossary v2. 0. Event Processing Technical Society.

Luckham, D.C. (2001) The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley Longman Publishing Co., Inc.

Luckham, D.C. and Frasca, B. 1998. Complex event processing in distributed systems. Computer Systems Laboratory Technical Report CSL-TR-98-754. Stanford University, Stanford 28.

Lupiani, E., Sauer, C., Roth-Berghofer, T., Juarez, J.M. and Palma, J. 2013 Implementation of similarity measures for event sequences in myCBR.

Luu, V.-T., Forestier, G., Weber, J., Bourgeois, P., Djelil, F. and Muller, P.-A. 2020. A review of alignment based similarity measures for web usage mining. Artificial Intelligence Review 53(3), 1529-1551.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R. and Anderson, J. 2002 Wireless sensor networks for habitat monitoring, pp. 88-97, Acm.

Malpas, J. 2012. Putting space in place: Philosophical topography and relational geography. Environment and planning D: society and space 30(2), 226-242.

Mannila, H. and Moen, P. 1999 Similarity between event types in sequences, pp. 271-280, Springer.

Mannila, H. and Ronkainen, P. 1997 Similarity of event sequences, pp. 136-139, IEEE.

Mannila, H. and Salmenkivi, M. 2001 Finding simple intensity descriptions from event sequence data, pp. 341-346, ACM.

Mannila, H., Toivonen, H. and Verkamo, A.I. 1995 Discovering frequent episodes in sequences extended abstract.

Marston, S.A., Jones III, J.P. and Woodward, K. 2005. Human geography without scale. Transactions of the institute of British geographers 30(4), 416-432.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V. 2007. Fault diagnosis using dynamic trend analysis: A review and recent developments. Engineering Applications of artificial intelligence 20(2), 133-146.

Mirbagheri, S.M. and Hamilton, H.J. 2020 Similarity Matching of Temporal Event-Interval Sequences, pp. 420-425, Springer.

Moen, P. 2000. Attribute, event sequence, and event type similarity notions for data mining. PhD thesis, University of Helsinki.

Moore, A. 2008. Rethinking scale as a geographical category: from analysis to practice. Progress in human geography 32(2), 203-225.

Moorthy, V., Restrepo, A.M.H., Preziosi, M.-P. and Swaminathan, S. 2020. Data sharing for novel coronavirus (COVID-19). Bulletin of the World Health Organization 98(3), 150.

Nittel, S. 2009. A survey of geosensor networks: Advances in dynamic environmental monitoring. Sensors 9(7), 5664-5678.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S. 2013 Using of Jaccard coefficient for keywords similarity.

Noble, R.T., Moore, D.F., Leecaster, M.K., McGee, C.D. and Weisberg, S.B. 2003. Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing. Water research 37(7), 1637-1643.

Obweger, H., Suntinger, M., Schiefer, J. and Raidl, G. 2010 Similarity searching in sequences of complex events, pp. 631-640, IEEE.

Oliveira, L.M. and Rodrigues, J.J. 2011. Wireless Sensor Networks: A Survey on Environmental Monitoring. JCM 6(2), 143-151.

Orozco, C.V., Tonini, M., Conedera, M. and Kanveski, M. 2012. Cluster recognition in spatial-temporal sequences: the case of forest fires. Geoinformatica 16(4), 653-673.

Othman, M.F. and Shazali, K. 2012. Wireless sensor network applications: A study in environment monitoring system. Procedia Engineering 41, 1204-1210.

Paasi, A. 2004. Place and region: looking through the prism of scale. Progress in human geography 28(4), 536-546.

Padhy, P., Martinez, K., Riddoch, A., Ong, H. and Hart, J.K. 2005. Glacial environment monitoring using sensor networks.

Paschke, A. and Boley, H. (2009) Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, pp. 215-252, IGI Global.

Peterson, M.R., Doom, T.E. and Raymer, M.L. 2005 Ga-facilitated knn classifier optimization with varying similarity measures, pp. 2514-2521, IEEE.

Peuquet, D.J. and Duan, N. 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International journal of geographical information systems 9(1), 7-24.

Prasad, A., Mamun, K.A., Islam, F. and Haqva, H. 2015 Smart water quality monitoring system, pp. 1-6, IEEE.

Prasath, V., Alfeilat, H.A.A., Lasassmeh, O. and Hassanat, A. 2017. Distance and similarity measures effect on the performance of K-nearest neighbor classifier-a review. arXiv preprint arXiv:1708.04321.

Prinzie, A. and Van den Poel, D. 2011. Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: an Acquisition Pattern Analysis application. Journal of Intelligent Information Systems 36(3), 283-304.

Qi, H., Xiao, S., Shi, R., Ward, M.P., Chen, Y., Tu, W., Su, Q., Wang, W., Wang, X. and Zhang, Z. 2020. COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. Science of the Total Environment, 138778.

Ros, F. and Guillaume, S. 2019. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Systems with Applications 128, 96-108.

Roux, A.V.D. and Mair, C. 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186(1), 125-145.

Rude, A. and Beard, K. 2012 High-Level Event Detection in Spatially Distributed Time Series, pp. 160-172, Springer Berlin Heidelberg, Berlin, Heidelberg.

Saglietto, A., D'Ascenzo, F., Zoccai, G.B. and De Ferrari, G.M. 2020. COVID-19 in Europe: the Italian lesson. Lancet 395(10230), 1110-1111.

Sampson, R.J. 2003. The neighborhood context of well-being. Perspectives in biology and medicine 46(3), S53-S64.

Schatzki, T.R. (2002) The site of the social: A philosophical account of the constitution of social life and change, Penn State University Press.

Shahar, Y. 1997. A framework for knowledge-based temporal abstraction. Artificial intelligence 90(1-2), 79-133.

Sheil, H., Rana, O. and Reilly, R. 2018. Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks. arXiv preprint arXiv:1807.08207.

Shekhar, S., Evans, M.R., Kang, J.M. and Mohan, P. 2011. Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3), 193-214.

Shurkhovetskyy, G., Andrienko, N., Andrienko, G. and Fuchs, G. 2018 Data abstraction for visualizing large time series, pp. 125-144, Wiley Online Library.

Simandan, D. 2019. Revisiting positionality and the thesis of situated knowledge. Dialogues in human geography 9(2), 129-149.

Simandan, D. 2020. Being surprised and surprising ourselves: a geography of personal and social change. Progress in Human Geography 44(1), 99-118.

Sinclair, C., Pierce, L. and Matzner, S. 1999 An application of machine learning to network intrusion detection, pp. 371-377, IEEE.

Singhal, A., Luo, J. and Zhu, W. 2003 Probabilistic spatial context models for scene content understanding, pp. I-I, IEEE.

Smith, T.F. and Waterman, M.S. 1981. Comparison of biosequences. Advances in applied mathematics 2(4), 482-489.

Stehle, S. and Peuquet, D.J. 2015. Analyzing spatio-temporal patterns and their evolution via sequence alignment. Spatial Cognition & Computation 15(2), 68-85.

Sun, S.-B., Zhang, Z.-H., Dong, X.-L., Zhang, H.-R., Li, T.-J., Zhang, L. and Min, F. 2017. Integrating Triangle and Jaccard similarities for recommendation. PloS One 12(8), e0183570.

Sunley, P. 1996. Context in economic geography: the relevance of pragmatism. Progress in Human Geography 20(3), 338-355.

Syakur, M., Khotimah, B., Rochman, E. and Satoto, B. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering 336(1), 012017.

Tao, C., Wongsuphasawat, K., Clark, K., Plaisant, C., Shneiderman, B. and Chute, C.G. 2012 Towards event sequence representation, reasoning and visualization for EHR data, pp. 801-806, ACM.

Tilman, D. and Kareiva, P. (2018) Spatial ecology: the role of space in population dynamics and interspecific interactions (MPB-30), Princeton University Press.

Timmermans, H., van der Waerden, P., Alves, M., Polak, J., Ellis, S., Harvey, A.S., Kurose, S. and Zandee, R. 2003. Spatial context and the complexity of daily travel patterns: an international comparison. Journal of Transport Geography 11(1), 37-46.

Tonidandel, S. and LeBreton, J.M. 2015. RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. Journal of Business and Psychology 30, 207-216.

Tonidandel, S., LeBreton, J.M. and Johnson, J.W. 2009. Determining the statistical significance of relative weights. Psychological methods 14(4), 387.

Tonini, M., Tuia, D. and Ratle, F. 2009. Detection of clusters using space–time scan statistics. International journal of wildland fire 18(7), 830-836.

Vanderbilt, K.L., Lin, C.-C., Lu, S.-S., Kassim, A.R., He, H., Guo, X., San Gil, I., Blankman, D. and Porter, J.H. 2015. Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network. Ecosphere 6(10), 1-18.

Vorontsov, I.E., Kulakovskiy, I.V. and Makeev, V.J. 2013. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms for Molecular Biology 8(1), 1-11.

Vrotsou, K. (2010) Everyday mining: Exploring sequences in event-based data, Linköping University Electronic Press.

Vrotsou, K. and Forsell, C. 2011 A qualitative study of similarity measures in event-based data, pp. 170-179, Springer.

Wang, T.-Y., Yang, M.-H. and Wu, J.-Y. 2016a. Distributed Detection of Dynamic Event Regions in Sensor Networks With a Gibbs Field Distribution and Gaussian Corrupted Measurements. IEEE Transactions on Communications 64(9), 3932-3945.

Wang, W., Hu, C., Chen, N., Xiao, C. and Jia, S. 2016b. Spatio-Temporal Risk Assessment Process Modeling for Urban Hazard Events in Sensor Web Environment. ISPRS International Journal of Geo-Information 5(11), 203.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. and Keogh, E. 2013. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26(2), 275-309.

Weber, J. and Kwan, M.-P. 2003. Evaluating the effects of geographic contexts on individual accessibility: a multilevel Approach1. Urban Geography 24(8), 647-671.

Weiss, G.M. and Hirsh, H. 1998 Learning to Predict Rare Events in Event Sequences, pp. 359-363.

Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J. and Welsh, M. 2006. Deploying a wireless sensor network on an active volcano. IEEE internet computing 10(2), 18-25.

Wolpert, J. 1964. The decision process in spatial context. Annals of the Association of American Geographers 54(4), 537-558.

Wongsuphasawat, K., Plaisant, C., Taieb-Maimon, M. and Shneiderman, B. 2012. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with computers 24(2), 55-68.

Wongsuphasawat, K. and Shneiderman, B. 2009 Finding comparable temporal categorical records: A similarity measure with an interactive visualization, pp. 27-34, IEEE.

Woodward, K., Jones III, J.P. and Marston, S.A. 2012. The politics of autonomous space. Progress in Human Geography 36(2), 204-224.

Worboys, M. 2005. Event-oriented approaches to geographic phenomena. International Journal of Geographical Information Science 19(1), 1-28.

Worboys, M. and Hornsby, K. 2004 From objects to events: GEM, the geospatial event model, pp. 327-343, Springer.

Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O. and Lemey, P. 2020. The emergence of SARS-CoV-2 in Europe and North America. Science 370(6516), 564-570.

Xu, F. and Beard, K. 2021. A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. ISPRS International Journal of Geo-Information 10(9), 594.

Yang, D.-H., Goerge, R. and Mullner, R. 2006. Comparing GIS-based methods of measuring spatial accessibility to health services. Journal of medical systems 30(1), 23-32.

Yang, J., McAuley, J., Leskovec, J., LePendu, P. and Shah, N. 2014 Finding progression stages in time-evolving event sequences, pp. 783-794, ACM.

Yeh, C.-C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Zimmerman, Z., Silva, D.F., Mueen, A. and Keogh, E. 2018. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Mining and Knowledge Discovery 32(1), 83-123.

Yih, W.K., Deshpande, S., Fuller, C., Heisey-Grove, D., Hsu, J., Kruskal, B.A., Kulldorff, M., Leach, M., Nordin, J. and Patton-Levine, J. 2010. Evaluating real-time syndromic surveillance signals from ambulatory care data in four states. Public Health Reports 125(1), 111-120.

Yin, F., Li, X., Ma, J. and Feng, Z. 2007. The early warning system based on the prospective space-time permutation statistic. Wei Sheng Yan Jiu (in Chinese: Journal of Hygiene Research) 36(4), 455-458.

Yin, J., Hu, D.H. and Yang, Q. 2009 Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields, pp. 1321-1327.

Zambelli, A.E. 2016. A data-driven approach to estimating the number of clusters in hierarchical clustering. F1000Research 5.

Zhang, W., Govindavari, J.P., Davis, B.D., Chen, S.S., Kim, J.T., Song, J., Lopategui, J., Plummer, J.T. and Vail, E. 2020. Analysis of genomic characteristics and transmission routes of patients with confirmed SARS-CoV-2 in Southern California during the early stage of the US COVID-19 pandemic. JAMA Network Open 3(10), e2024191-e2024191.

Zolnik, E.J. 2009. Context in human geography: a multilevel approach to study human–environment interactions. The Professional Geographer 61(3), 336-349.

APPENDICES

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

The following are available online at

https://www.mdpi.com/article/10.3390/ijgi10090594/s1

Table A.1. Table S1: Precipitation data of 43 monitoring stations in the Maine coast (2010–2014).

Table A.2. Table S2: Event sequence matrix of 43×192 from eventization based on ≥1″ precipitation.

Table A.3. Table S3: Event sequence matrix of 43×52 from eventization based on ≥2″ precipitation.

Table A.4. Table S4: Similarity matrix of 43 × 43 from the event sequence matrix of **Table A.2**.

Software Availability:

(1) STS.eventize, STS.eventize1, STS.eventize2, and STS.eventize3 (conversion of space-time series to event sequences considering either variation of events or not based on domain context and user's requirements), and

(2) STES.sim1, STES.sim2, STES.sim3, STES.simOr, and STES.simOr2 (Calculation of global and local similarities between event sequences based on sequences of different event types and generation of global and local similarity matrices as user defined local granularity or window size).

Source code: https://frank888.github.io/STES_similarity.html (accessed on 8 September 2021).

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

The following supporting information can be downloaded at:

https://www.mdpi.com/article/10.3390/geographies3020016/s1

Table B.1. Table S1: Static variables of basin characteristics associated with 16 monitoring stations.

Table B.2. Table S2: Dynamic variables and fecal coliform scores in 16 monitoring stations.

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 5

The following are available online at:

https://www.mdpi.com/article/10.3390/ijgi10090594/s1

Table C.1. S1 Table. Comparison of space-time clusters from SaTScan and STES based hierarchical clustering with the dataset from 1/23-3-13/2020. This table is merged through FIPS of US counties, and also includes other selected output parameters from SaTScan such as p-values, LOC_RR (location or county relative risk), CLU_RR (cluster relative risk), LOC_LAT (location latitude), LOC_LONG (location longitude). https://doi.org/10.1371/journal.pone.0252990.s001 (XLSX)

Table C.2. Table. Comparison of space-time clusters from SaTScan and STES based hierarchical clustering with the dataset from 1/23-3-31/2020. This table is merged through FIPS of US counties, and also includes other selected output parameters from SaTScan such as p-values, LOC_RR (location or county relative risk), CLU_RR (cluster relative risk), LOC_LAT (location latitude), LOC_LONG (location longitude). https://doi.org/10.1371/journal.pone.0252990.s002 (XLSX)

Table C.3. S3 Table. Comparison of space-time clusters from SaTScan and STES based hierarchical clustering with the dataset from 1/23-4-19/2020. This table is merged through FIPS of US counties, and also includes other selected output parameters from SaTScan such as p-values, LOC_RR (location or county relative risk),

CLU_RR (cluster relative risk), LOC_LAT (location latitude), LOC_LONG (location longitude). https://doi.org/10.1371/journal.pone.0252990.s003 (XLSX)

Table C.4. S4 Table. Comparison of space-time clusters from SaTScan and STES based hierarchical clustering with the dataset from 1/23-5-20/2020. This table is merged through FIPS of US counties, and also includes other selected output parameters from SaTScan such as p-values, LOC_RR (location or county relative risk), CLU_RR (cluster relative risk), LOC_LAT (location latitude), LOC_LONG (location longitude). https://doi.org/10.1371/journal.pone.0252990.s004 (XLSX)

Table C.5. S5 Table. The minimal data set underlying the results described in this manuscript. https://doi.org/10.1371/journal.pone.0252990.s005 (CSV)

BIOGRAPHY OF THE AUTHOR

Fuyu Xu, *aka* “Frank”, received his Bachelor of Science degree in Agriculture from Fujian Agriculture and Forestry University in China many years ago. After completing his undergraduate degree, he taught plant genetics at the same university for three years before pursuing his Master of Science degree in Ecology with a major in Ecological Genetics at the Chinese Academy of Sciences in Beijing and Shenyang, China. During and after his master's program, he devoted several years to conducting research on tree genetic improvement and plant physiology.

In 1997, Frank joined the University of Maine as a visiting scientist and later enrolled in the graduate school, where he earned his Master of Science in Forest Resources in 2002. This program further expanded his knowledge and expertise in forestry genetics and plant physiology. In 2009, he completed his first Ph.D. in Biotechnology and Modern Genetics from Michigan Technological University. Following his Ph.D., Frank returned to the University of Maine as a postdoctoral researcher, where he worked on rice stress molecular biology and genetics.

Driven by his passion for spatial data analysis and his desire to make a significant contribution to the field of spatial information science and engineering, Frank embarked on a second Ph.D. at the University of Maine. Fuyu Xu is a candidate for the Doctor of Philosophy degree in Spatial Information Science and Engineering from the University of Maine in May 2023.