Article

Not peer-reviewed version

Similarity Measure of Spatiotemporal Event Setting Sequences: Method Development and A Case Study on Monitoring Coastal Fecal Coliform Pollution Events

Fuyu Xu * and Kate Beard *

Posted Date: 21 March 2023

doi: 10.20944/preprints202303.0366.v1

Keywords: spatiotemporal setting sequences; similarity measure; event sequences; matrix representation; static variables; dynamic variables; basin characteristics; Jaccard index; relative importance analysis; clustering analysis

Article

Similarity Measure of Spatiotemporal Event Setting Sequences: Method Development and A Case Study on Monitoring Coastal Fecal Coliform Pollution Events

Fuyu Xu and Kate Beard

School of Computing and Information Science, University of Maine, USA; fuyu.xu@maine.edu (F.X.); kate.beard@maine.edu; Tel.: 1-207-581-2147 (K.B.)

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 incorporates 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, these similarity measures have many potential real-world applications, and offer researchers a powerful tool for understanding different factors and their dynamics corresponding to occurrences of spatiotemporal event sequences.

Keywords: spatiotemporal setting sequences; similarity measure; event sequences; matrix representation; static variables; dynamic variables; basin characteristics; Jaccard index; relative importance analysis; clustering analysis

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 [1–7]. 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 occurrences of events 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 [1], 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 sequence.

Measures of similarity among event sequence settings have several potential ap-plications 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 be used to identify patterns in the spatial-temporal settings of disease outbreaks or other public health concerns. Such in-formation 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 be used to 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 be used to 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 be used to 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.

We use the term 'setting' as defined by Worboys and Hornsby [8], and distinguish it from spatial context, which has been widely studied. We first review research work on spatial context for a better understanding of spatial or spatiotemporal settings proposed in this paper. Context has been defined in many ways, most often with location as the most important emphasis, namely spatial context. Context has been described as the "location and the identity of nearby people and objects." [9]. Context, as has been used in computer science includes 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 [10–12].

In geography, spatial-temporal contexts refer to the physical and social conditions that exist in a particular place and time [13–16]. These contexts can include factors such as the natural environment, climate, culture, economic conditions, and population characteristics. Spatial-temporal contexts can also 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. We distinguish spatial-temporal settings as referring to the specific location and time frame in which an event or phenomenon occurs. A spatial-temporal setting can 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 general, spatial-temporal contexts describe the general or broader context in which an event or phenomenon occurs, while spatial-temporal settings describe the specific location and time frame in which the event or phenomenon occurs.

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 [17]. In travel behavior research, spatial context was shown to be strongly related to household travel patterns at an international scale [18]. A person's health-related problems can be strongly affected by different types of spatial context, such as environmental exposures [19,20], social environment (characteristics of communities and neighborhoods) [20,21], and ease of access to health services [22]. 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 [23]. 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 [24,25].

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% [26]. 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 [27].

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 [28,29]. Their work focuses on quantifying the impact of changing contexts on similarity measures thus recognizing potential influence of context on similarity measure 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 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 observed 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 are evaluated through a case study. The developed similarity measure serves as an index that combines a set of quantitative and qualitative factors.

2. Materials and Methods

2.1. Model for Event Sequence Settings

A key consideration in the specification of a setting is how to define its bounds both spatially and temporally. For the event sequence similarity measure described by Xu and Beard [1], they 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. What constitutes a spatial setting will thus 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 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.

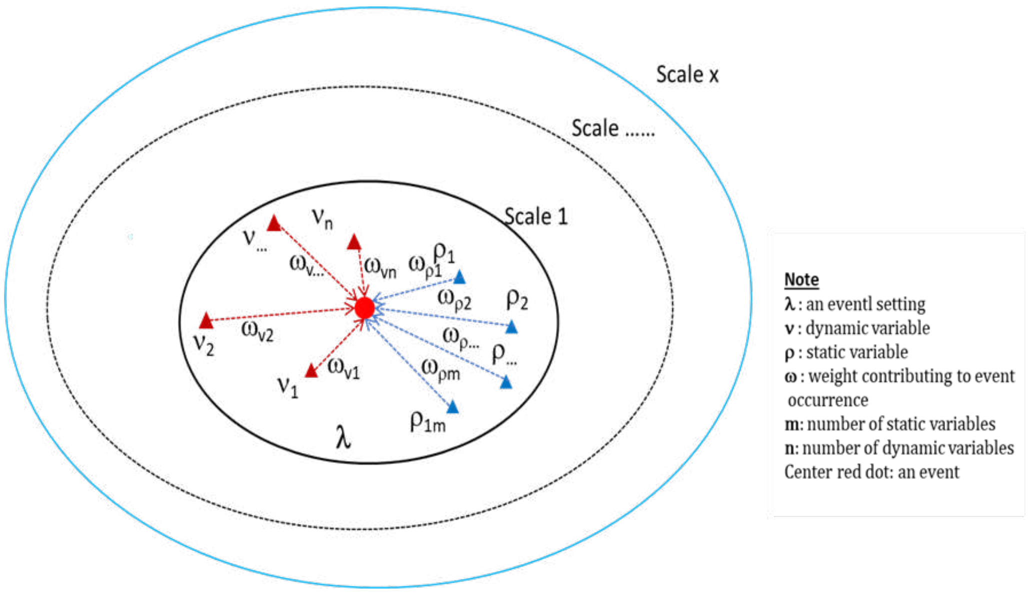

Figure 1. Schematic representation of an event situated setting considering different spatial scales for the setting. Influencing factors with different weights are shown only at Scale 1. More, fewer, or different sets of factors may apply at another scale.

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 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 [1], we specify corresponding sequences of spatiotemporal event settings. Figure 2 graphically illustrates these conceptual sequences of spatiotemporal event settings with n dynamic and m static representative variables.

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 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 × t matrix as schematically illustrated in Figure 3.

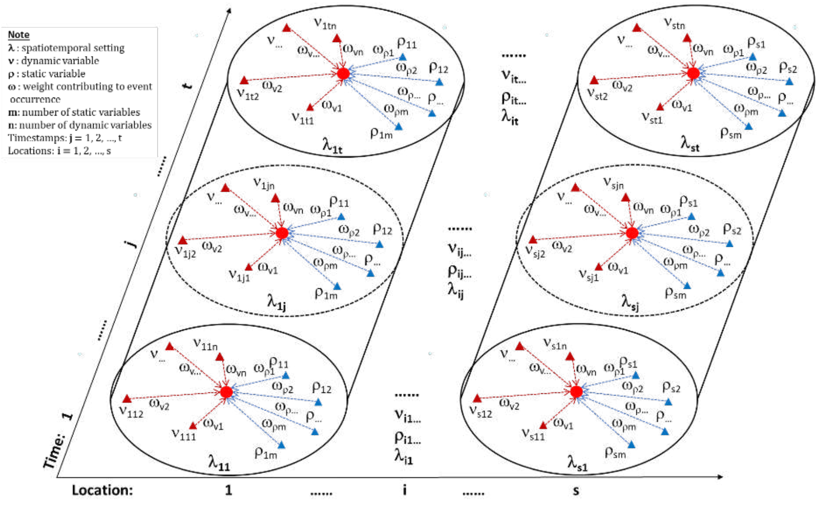

Figure 2. Schematic illustration of sequences of spatial-temporal settings with t time points and s locations.

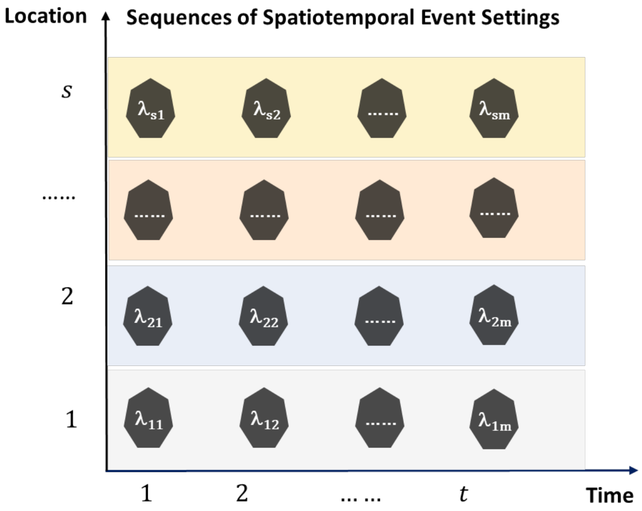

Figure 3. Schematic matrix representation of sequences of spatial-temporal settings with t time points and s locations. λ_{st} - a setting at location s and time t.

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

$$\lambda : \nu_1, \nu_2, \ldots\ldots, \nu_n; \rho_1, \rho_2, \ldots\ldots, \rho_m$$

Such that, Figure 3 can be expanded to Figure 4 to become the variables-based matrix representation of the sequences of spatiotemporal event settings.

$$\begin{array}{cc}
 & \textbf{Temporal points} \\
\rotatebox{90}{\textbf{Spatial locations}} & \begin{array}{c} \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \cdots \\ \mathbf{s} \end{array}
\begin{pmatrix}
\overbrace{\nu_{111}, \nu_{112}, \ldots\ldots, \nu_{11n}, \rho_{11}, \rho_{12}, \ldots\ldots, \rho_{1m}}^{1,}, \overbrace{\nu_{121}, \nu_{122}, \ldots\ldots, \nu_{12n}, \rho_{11}, \rho_{12}, \ldots\ldots, \rho_{1m}}^{2,}, \overset{\ldots\ldots,}{\cdots\cdots,} \overbrace{\nu_{1t1}, \nu_{1t2}, \ldots\ldots, \nu_{1tn}, \rho_{11}, \rho_{12}, \ldots\ldots, \rho_{1m}}^{t} \\
\nu_{211}, \nu_{212}, \ldots\ldots, \nu_{21n}, \rho_{21}, \rho_{22}, \ldots\ldots, \rho_{2m}, \nu_{221}, \nu_{222}, \ldots\ldots, \nu_{22n}, \rho_{21}, \rho_{22}, \ldots\ldots, \rho_{2m}, \cdots\cdots, \nu_{2t1}, \nu_{2t2}, \ldots\ldots, \nu_{2tn}, \rho_{21}, \rho_{22}, \ldots\ldots, \rho_{2m} \\
\nu_{311}, \nu_{312}, \ldots\ldots, \nu_{31n}, \rho_{31}, \rho_{32}, \ldots\ldots, \rho_{3m}, \nu_{321}, \nu_{322}, \ldots\ldots, \nu_{32n}, \rho_{31}, \rho_{32}, \ldots\ldots, \rho_{3m}, \cdots\cdots, \nu_{3t1}, \nu_{3t2}, \ldots\ldots, \nu_{3tn}, \rho_{31}, \rho_{32}, \ldots\ldots, \rho_{3m} \\
\cdots \\
\nu_{s11}, \nu_{s12}, \ldots\ldots, \nu_{s1n}, \rho_{s1}, \rho_{s2}, \ldots\ldots, \rho_{sm}, \nu_{s21}, \nu_{s22}, \ldots\ldots, \nu_{s2n}, \rho_{s1}, \rho_{s2}, \ldots\ldots, \rho_{sm}, \cdots\cdots, \nu_{st1}, \nu_{st2}, \ldots\ldots, \nu_{stn}, \rho_{s1}, \rho_{s2}, \ldots\ldots, \rho_{sm}
\end{pmatrix}
\end{array}$$

Figure 4. Matrix representation of sequences of spatiotemporal event settings with s locations and t time points.

2.3. Similarity Measures of Spatial Settings

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 [1]. 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 [30]. To adapt the Jaccard index for measuring the similarity between spatial settings associated to 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) for the dynamic variables or elements that their values 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):

Equation (1):

$$sim_k(l_1, l_2) = \frac{Sd_{k12}+Ss_{12}}{|l_{d1}\cup l_{d2}|+|l_{s1}\cup l_{s2}|} = \frac{Sd_{k12}+Ss_{12}}{N_d+N_s} \tag{1}$$

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 Equation (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:

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

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

$$sim_k(l_1, l_2) = \frac{c_{kd12}\sum_{i=1}^{Ckd12}\omega_i\left(1-Abs\left(lev(d_{k1i})-lev(d_{k2i})\right)\right)}{N_d} + \frac{c_{s12}\sum_{j=1}^{Cs12}\omega_j\left(1-Abs\left(lev(s_{1j})-lev(s_{2j})\right)\right)}{N_s} \tag{3}$$

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

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

$$sim_k(l_1, l_2) = \frac{\sum_{i=1}^{Ckd12}\left(1-\frac{Abs(lev(d_{k1i})-lev(d_{k2i}))}{n_i-1}\right)+\sum_{j=1}^{Cs12}\left(1-\frac{Abs(lev(s_{1i})-lev(s_{2i}))}{m_j-1}\right)}{N_d+N_s} \tag{4}$$

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

$$sim_k(l_1, l_2) = \frac{c_{kd12}\sum_{i=1}^{Ckd12}\omega_i\left(1-\frac{Abs(lev(d_{k1i})-lev(d_{k2i}))}{n_i-1}\right)}{N_d} + \frac{c_{s12}\sum_{j=1}^{Cs12}\omega_j\left(1-\frac{Abs(lev(s_{1i})-lev(s_{2j}))}{m_j-1}\right)}{N_s} \tag{5}$$

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:

$$lev(d_{k1i}) = \frac{d_{k1i}}{d_{k1i}+d_{k2i}} \text{ and } lev(d_{k2i}) = \frac{d_{k2i}}{d_{k1i}+d_{k2i}} \quad (6)$$

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:

$$lev(s_{1i}) = \frac{s_{1i}}{s_{1i}+s_{2i}} \text{ and } lev(s_{2i}) = \frac{s_{2i}}{s_{1i}+s_{2i}} \quad (7)$$

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):

$$\begin{aligned} sim_{global}(S_1, S_2) &= \frac{\sum_{k=1}^{T} sim_k(l_1, l_2)}{T} \\ &= \frac{\sum_{k=1}^{T}(Sd_{k12}+Ss_{12})}{T(N_d+N_s\ )} \\ &= \frac{\sum_{k=1}^{T} Sd_{k12}}{T(N_d+N_s)} + \frac{\sum_{i=1}^{T} Ss_{12}}{T(N_d+N_s)} \\ &= \frac{\sum_{k=1}^{T} Sd_{k12}}{T(N_d+N_s)} + \frac{Ss_{12}}{N_d+N_s} \end{aligned} \quad (8)$$

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 when 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:

$$sim_{global}(S_1, S_2) = \frac{\sum_{k=1}^{T}\sum_{i=1}^{Ckd12}\left(1-Abs\left(lev(ld_{k1i})-lev(ld_{k2i})\right)\right)}{T(N_d+N_s)} + \frac{\sum_{j=1}^{Cs12}\left(1-Abs\left(lev(ls_{1j})-lev(ls_{2j})\right)\right)}{N_d+N_s} \quad (9)$$

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

$$sim_{global}(S_1, S_2) = \frac{c_{kd12}\sum_{k=1}^{T}\sum_{i=1}^{Ckd12} \omega_i\left(1-Abs\left(lev(ld_{k1i})-lev(ld_{k2i})\right)\right)}{N_d} + \frac{T*c_{s12}\sum_{j=1}^{Cs12} \omega_j\left(1-Abs\left(lev(s_{1j})-lev(s_{2j})\right)\right)}{N_s} \quad (10)$$

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

$$sim_{global}(S_1,S_2) = \frac{\sum_{k=1}^{T}\sum_{i=1}^{Ckd12}\left(1-\frac{Abs(lev(d_{k1i})-lev(d_{k2i}))}{n_i-1}\right)}{T(N_d+N_s)} + \frac{\sum_{j=1}^{Cs12}\left(1-\frac{Abs(lev(s_{1i})-lev(s_{2i}))}{m_j-1}\right)}{N_d+N_s} \qquad (11)$$

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

$$sim_{global}(S_1,S_2) = \frac{c_{kd12}\sum_{k=1}^{T}\sum_{i=1}^{Ckd12}\omega_i\left(1-\frac{Abs(lev(d_{k1i})-lev(d_{k2i}))}{n_i-1}\right)}{N_d} + \frac{T*c_{s12}\sum_{j=1}^{Cs12}\omega_j\left(1-\frac{Abs(lev(s_{1i})-lev(s_{2i}))}{m_j-1}\right)}{N_s} \qquad (12)$$

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 or 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 5 shows the steps for implementing similarity assessment between event settings or sequences of event settings in a specific domain.

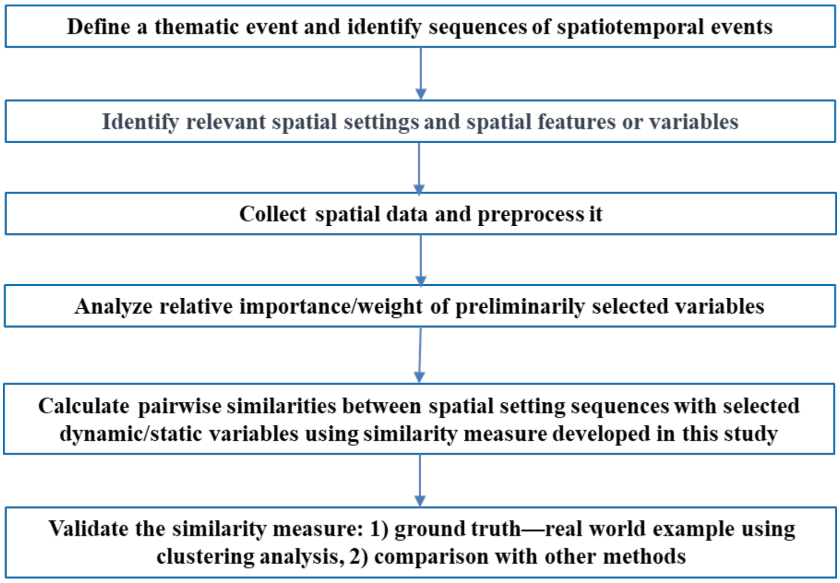
Define a thematic event and identify sequences of spatiotemporal events
Identify relevant spatial settings and spatial features or variables
Collect spatial data and preprocess it
Analyze relative importance/weight of preliminarily selected variables
Calculate pairwise similarities between spatial setting sequences with selected dynamic/static variables using similarity measure developed in this study
Validate the similarity measure: 1) ground truth—real world example using clustering analysis, 2) comparison with other methods

Figure 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 wind direction, wind speed, sites of local manufactures, major pollution sources, concentration of major pollutants, transportation density, etc. 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 the thematic events. Preprocessing or preparation of the collected data includes normal distribution check, normalization, standardization of measurement units, and binarization of categorical data, etc.

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) [31–33] and partial least squares regression (PLSR) [34].

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 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.

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 [35]. 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 [36,37]. 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.

3.1. Experimental Site and Design

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 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 [38]. 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 1.

doi:10.20944/preprints202303.0366.v1

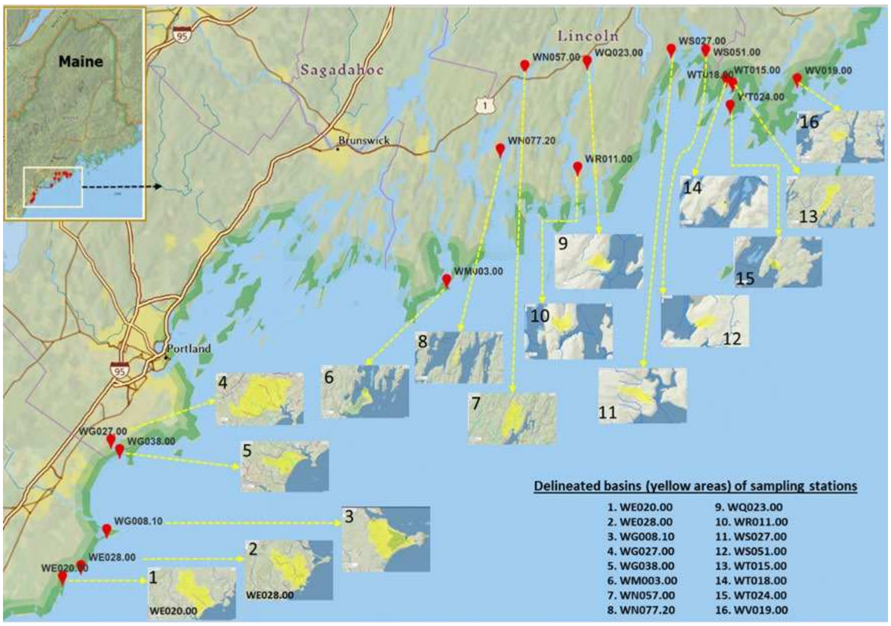

Figure 6. Selected monitoring stations/locations on the Maine coast for depicting spatiotemporal settings of fecal pollution event sequences.

Table 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

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/) and download all associated basin characteristics data. For the static variables described in Table 1, the data were extracted as shown in Table S1. We obtained marine environment related variables and fecal coliform measurements from Maine DMR (Table S2).

3.1.3. Methods

We use partial least squares regression (PLSR) analysis [32,33] to obtain the relative importance of all variables against the fecal coliform scores. We use the similarity measure developed in this study to achieve the similarity matrix of spatial setting sequences and use the method developed in [1] 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 did a cluster analysis [39].

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

The results of the partial least squares regression analysis on 39 variables reveal that some variables are more important than others in predicting the fecal coliform levels (Table 2 and Figure 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.

Table 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		

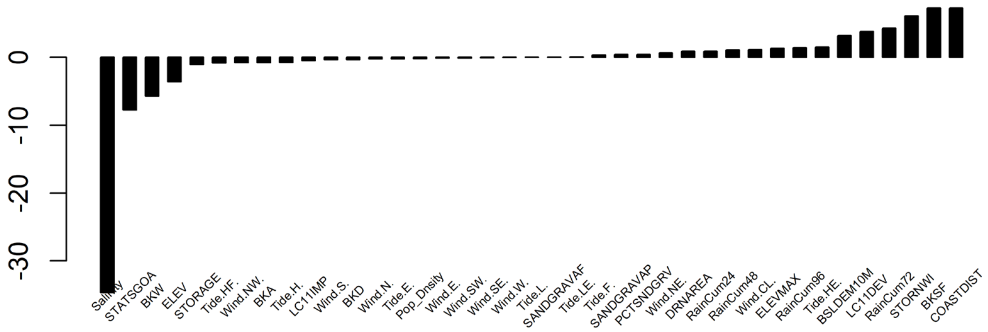

Figure 7. Bar chart of relative importance of 39 selected static and dynamic explanatory variables for fecal coliform bacterial measurements.

To reduce the number of variables for calculating similarity in the formula developed in this study we select the variables with higher weights. In this case study we select 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 3 and Figure 8, which can be used as relative weights for calculating similarities between spatial setting sequences when considering contribution from these individual variables.

Table 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

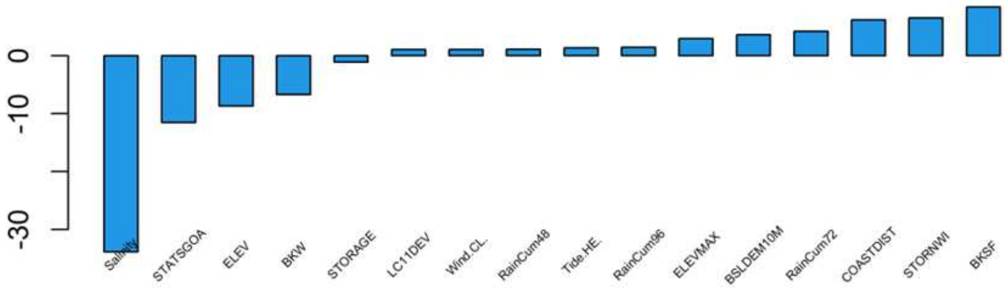

Figure 8. Bar chart of relative importance of 16 selected static and dynamic against fecal coliform bacterial measurements.

3.1. 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 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 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.

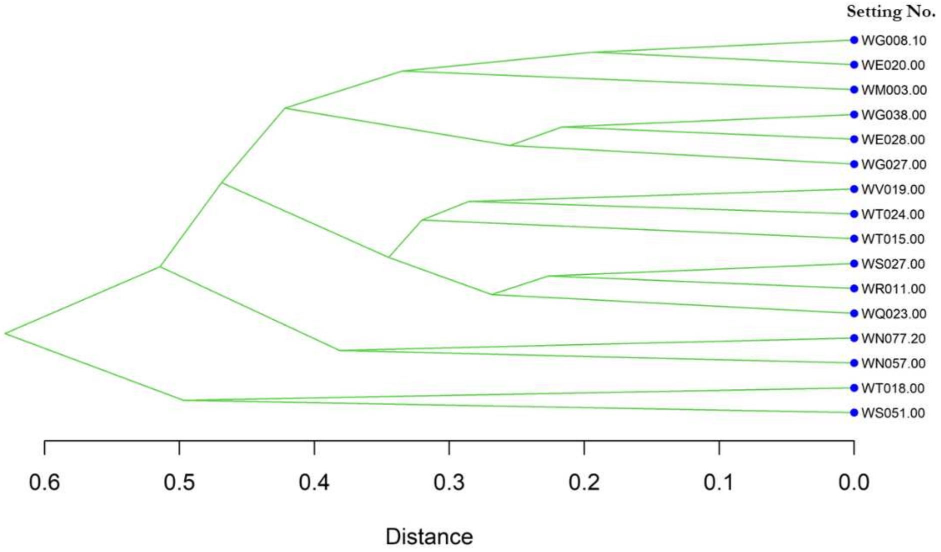

Figure 9. Clusters of 16 spatial setting sequences labeled with monitoring stations.

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 10. Three major clusters are clearly identified.

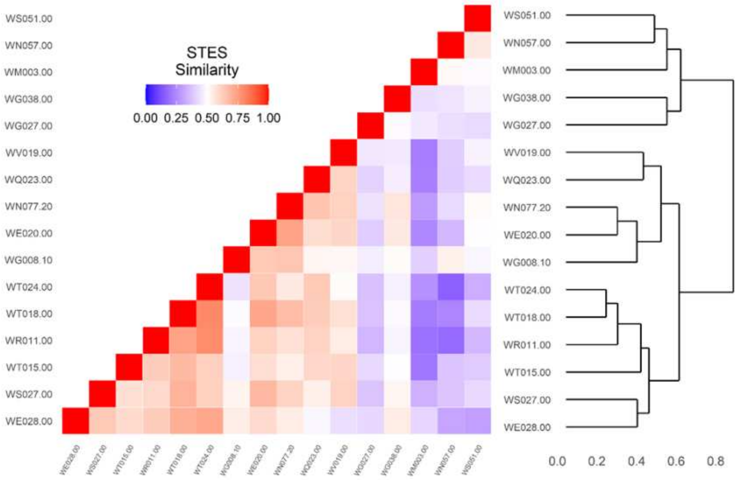

Figure 10. Similarity-based heat map and distance based hierarchical clustering between 16 monitoring stations for fecal pollution event sequences.

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 clustering results above from both setting sequences and event sequences side by side to build the cross-comparison graph (Figure 11). 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.

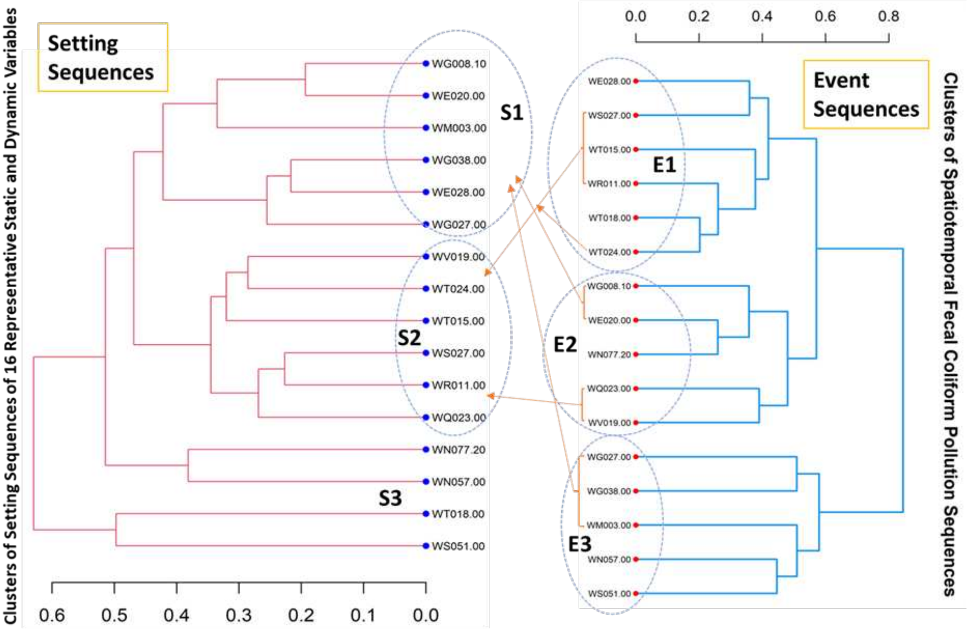

Figure 11. Cross analysis between clusters of setting sequences and clusters of event sequences.

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 a useful tool 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.

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.

5. Conclusions

In conclusion, modeling spatiotemporal event sequences requires a 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 develop 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 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.

Supplementary Materials: Table S1: Static variables of basin characteristics associated with 16 monitoring stations. Table S2: Dynamic variables and fecal coliform scores in 16 monitoring stations.

Funding: This research received no external funding.

Author Contributions: Conceptualization, Fuyu Xu and Kate Beard; methodology, Fuyu Xu and Kate Beard; software, Fuyu Xu and Kate Beard; validation, Fuyu Xu and Kate Beard; investigation, Fuyu Xu and Kate Beard; writing—original draft preparation, Fuyu Xu; writing—review & editing, Kate Beard; supervision, Fuyu Xu and Kate Beard. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, F.; Beard, K. A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. *ISPRS International Journal of Geo-Information* **2021**, *10*, 594.
2. Lupiani, E.; Sauer, C.; Roth-Berghofer, T.; Juarez, J.M.; Palma, J. Implementation of similarity measures for event sequences in myCBR. In Proceedings of the Proceedings of the 18th UKCBR Workshop, 2013.
3. Guralnik, V.; Srivastava, J. Event detection from time series data. In Proceedings of the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999; pp. 33-42.
4. Moen, P. Attribute, event sequence, and event type similarity notions for data mining. *PhD thesis, University of Helsinki* **2000**.
5. Mannila, H.; Ronkainen, P. Similarity of event sequences. In Proceedings of the Temporal Representation and Reasoning, 1997.(TIME'97), Proceedings., Fourth International Workshop on, 1997; pp. 136-139.

6. Obweger, H.; Suntinger, M.; Schiefer, J.; Raidl, G. Similarity searching in sequences of complex events. In Proceedings of the Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on, 2010; pp. 631-640.
7. Wongsuphasawat, K.; Plaisant, C.; Taieb-Maimon, M.; Shneiderman, B. Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting with computers* **2012**, *24*, 55-68.
8. Worboys, M.; Hornsby, K. From objects to events: GEM, the geospatial event model. In Proceedings of the International Conference on Geographic Information Science, 2004; pp. 327-343.
9. Jiang, B.; Yao, X. Location based services and GIS in perspective. In *Location based services and telecartography*; Springer: 2007; pp. 27-45.
10. Dey, A.K. Understanding and using context. *Personal and ubiquitous computing* **2001**, *5*, 4-7.
11. Brézillon, P.; Gonzalez, A.J. *Context in computing: a cross-disciplinary approach for modeling the real world*; Springer: 2014.
12. Loke, S. *Context-aware pervasive systems: architectures for a new breed of applications*; CRC Press: 2006.
13. Zolnik, E.J. Context in human geography: a multilevel approach to study human–environment interactions. *The Professional Geographer* **2009**, *61*, 336-349.
14. Sunley, P. Context in economic geography: the relevance of pragmatism. *Progress in Human Geography* **1996**, *20*, 338-355.
15. Weber, J.; Kwan, M.-P. Evaluating the effects of geographic contexts on individual accessibility: a multilevel Approach1. *Urban Geography* **2003**, *24*, 647-671.
16. Gong, H.; Hassink, R. Context sensitivity and economic-geographic (re) theorising. *Cambridge Journal of Regions, Economy and Society* **2020**, *13*, 475-490.
17. Delbosc, A.; Currie, G. The spatial context of transport disadvantage, social exclusion and well-being. *Journal of Transport Geography* **2011**, *19*, 1130-1137.
18. Timmermans, H.; van der Waerden, P.; Alves, M.; Polak, J.; Ellis, S.; Harvey, A.S.; Kurose, S.; Zandee, R. Spatial context and the complexity of daily travel patterns: an international comparison. *Journal of Transport Geography* **2003**, *11*, 37-46.
19. Cutter, S.L. Vulnerability to environmental hazards. *Progress in human geography* **1996**, *20*, 529-539.
20. Roux, A.V.D.; Mair, C. Neighborhoods and health. *Annals of the New York Academy of Sciences* **2010**, *1186*, 125-145.
21. Sampson, R.J. The neighborhood context of well-being. *Perspectives in biology and medicine* **2003**, *46*, S53-S64.
22. Yang, D.-H.; Goerge, R.; Mullner, R. Comparing GIS-based methods of measuring spatial accessibility to health services. *Journal of medical systems* **2006**, *30*, 23-32.
23. Wolpert, J. The decision process in spatial context. *Annals of the Association of American Geographers* **1964**, *54*, 537-558.
24. Gripenberg, S.; Roslin, T. Up or down in space? Uniting the bottom-up versus top-down paradigm and spatial ecology. *Oikos* **2007**, *116*, 181-188.
25. Tilman, D.; Kareiva, P. *Spatial ecology: the role of space in population dynamics and interspecific interactions (MPB-30)*; Princeton University Press: 2018; Volume 30.
26. Singhal, A.; Luo, J.; Zhu, W. Probabilistic spatial context models for scene content understanding. In Proceedings of the Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, 2003; pp. I-I.
27. Heitz, G.; Koller, D. Learning spatial context: Using stuff to find things. In Proceedings of the European conference on computer vision, 2008; pp. 30-43.
28. Keßler, C. Similarity measurement in context. In Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context, 2007; pp. 277-290.
29. Keßler, C.; Raubal, M.; Janowicz, K. The effect of context on semantic similarity measurement. In Proceedings of the OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", 2007; pp. 1274-1284.
30. Choi, S.-S.; Cha, S.-H.; Tappert, C.C. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics* **2010**, *8*, 43-48.
31. Chao, Y.-C.E.; Zhao, Y.; Kupper, L.L.; Nylander-French, L.A. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. *Journal of occupational and environmental hygiene* **2008**, *5*, 519-529.
32. Tonidandel, S.; LeBreton, J.M. RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology* **2015**, *30*, 207-216.
33. Tonidandel, S.; LeBreton, J.M.; Johnson, J.W. Determining the statistical significance of relative weights. *Psychological methods* **2009**, *14*, 387.
34. Ali, F.; Rasoolimanesh, S.M.; Sarstedt, M.; Ringle, C.M.; Ryu, K. An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. *International Journal of Contemporary Hospitality Management* **2018**.

35. Noble, R.T.; Moore, D.F.; Leecaster, M.K.; McGee, C.D.; Weisberg, S.B. Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing. *Water research* **2003**, *37*, 1637-1643.
36. Dong, J.; Wang, G.; Yan, H.; Xu, J.; Zhang, X. A survey of smart water quality monitoring system. *Environmental Science and Pollution Research* **2015**, *22*, 4893-4906.
37. Prasad, A.; Mamun, K.A.; Islam, F.; Haqva, H. Smart water quality monitoring system. In Proceedings of the 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2015; pp. 1-6.
38. Hughes, K.A. Influence of seasonal environmental variables on the distribution of presumptive fecal coliforms around an Antarctic research station. *Applied and Environmental Microbiology* **2003**, *69*, 4884-4891.
39. Kettenring, J.R. The practice of cluster analysis. *Journal of classification* **2006**, *23*, 3-30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.