Quantitative spatio-temporal sequence discovery

Group G10:
Anuj Karpatne, Vijay Borra

Abstract

Quantitative spatio-temporal (ST) sequences represent chains of ST event types in a ST framework. For example, it has been widely observed that deforestation in peat-land regions in Southeast Asia are often followed by forest fires, in close proximity of space and time [9]. Such patterns between deforestation and forest fires are attributed to the physical phenomena of degradation of soil due to logging, which reduces the moisture in the soil, making the underneath peat reserves more susceptible to fire leading to massive emissions of carbon [6]. Given an array of real-valued global variables representing distinct features of the area under study, a spatio-temporal neighborhood relation, an event detection / segmentation scheme and a user-defined similarity measure between two real-valued interval-based events, quantitative ST sequence discovery reports all sequential patterns with interestingness measure above a user-defined threshold along with an account of their time lags. Current literature primarily focuses on boolean event types with instantaneous event instances represented as points in space and time [5]. However, in many application domains including, forestry, the input variables are continuous and are in a raster data format, where the events span arbitrary time intervals. In this project, we attempt to formalize the quantitative ST sequence discovery problem that deals with raster data inputs while accounting for events which span arbitrary time intervals.

1 Introduction

In climate and earth science research, we are provided with a collection of real-valued global variables at each location on the earth, taken at a specified high spatial resolution, and varying with time. Figure 2 provides an overview of the format of the dataset.

Examples of such real-valued variables include vegetation index, precipitation, nighttime light information, temperature, fire information, flood information etc. Each signal is representative of a distinct feature about the domain, and captures the information about changes in a particular feature. There are complex feedback structures between Earth system processes: land cover, climate systems and ecosystems. In a framework which utilizes multiple variables
at varying locations and time, we can harness information about the influence of events occurring in a particular feature space on another variable, in close proximity of space and time.

Understanding such chains of events occurring across multiple signals can help in developing better descriptive models of the occurrence and evolution of events in a global domain. It can help in finding interesting patterns of events occurring across multiple signals, some of which have been observed in existing domain literature, and provide new insights in understanding the global system at large.

2 Motivation

In the peat-land marsh forests of Indonesia, a common trend that has been observed in related domain literature [8] is that deforestation in peat-land forests are often accompanied by fire at those cleared land locations in a span of about 3 years. Such patterns between deforestation and forest fires are attributed to the physical phenomena of degradation of soil due to logging, which reduces the moisture in the soil, making the underneath peat reserves more susceptible to fire. Figure ?? shows an example sequential relationship of loss in vegetation
data at a particular location, attributed to deforestation, eventually leading to the occurrence of a forest fire after a certain time lag, observed as a low difference between the day and night surface temperature difference values. Normally, there is a difference between the day and light surface temperature values for a forest location due to the cooling effect of forests at night. However, in case of a forest fire, the land is burning and the land surface temperature exhibits low difference between its day and night time values. Another example of its application includes characterizing land cover change due to urbanization at a location near and around a city. In urban amplification examples, a trend that is commonly observed is a decrease in the vegetation index (because of land clearing), subsequent increase in the nighttime lights, and increase in the land surface temperature due to construction of sub-urban cities.

The development of quantitative ST sequence patterns is aimed at discovering and evaluating the occurrence of such chains of events occurring in the real-world using multiple datasets representing distinct features about the domain.

Figure 2: Example sequential pattern between drop in vegetation index and low difference between day and night land surface temperature values observed in Indonesia. The figure shows the noisy characteristics of remote sensing data which have to be additionally handled.
3 Problem Definition

The problem can be defined more formally in the following fashion:

3.1 Input

3.1.1 Set of global variables

Given a set of real-valued global variables, let each variable be represented as a continuous time series of an attribute at each location. Let $X_j(i,t)$ denote value of the $j^{th}$ global variable at location $i$ and time $t$, where $1 \leq j \leq J$, $1 \leq i \leq I$ and $0 \leq t \leq T$.

3.1.2 Neighborhood Relationship

Let the spatial neighborhood relationship between 2 points $i_1$ and $i_2$ be denoted by $N(i_1,i_2)$ where $N(i_1,i_2) = 1$ if $i_1$ and $i_2$ satisfy a predefined neighborhood relationship, 0 otherwise. For the simple case, we can consider our neighborhood relationship as the basic 1-hop neighborhood.

3.1.3 Event Detection Scheme

Given a time series for the $j^{th}$ global variable at location $i$, an event detection scheme partitions the time series into separate time intervals, where each time interval represents a distinct underlying process in the given global variable at the specified location. Such distinct processes are termed as events occurring in different global variables at different locations. An example event detection scheme is shown in Figure where the time series is partitioned into distinct events, $e_1$ and $e_2$. Let $E^j$ denote the set of all possible event instances (time intervals) which occur in the $j^{th}$ variable at all locations.

3.1.4 Event Relationship

Given two events $e_1$ and $e_2$ occurring in different global variables and at different locations, $e_1$ and $e_2$ are said to satisfy the follow relationship : $e_1 \rightarrow e_2$, if the time lag between the start time of time interval $e_2$ from the start time of time interval $e_1$ falls is a positive value. An event relationship of the form $X_{j_1} \xrightarrow{R} X_{j_2}$ is said to exist between variables $X_{j_1}$ and $X_{j_2}$, $1 \leq j_1, j_2 \leq k$, if event (time interval) pairs $(e_1, e_2)$ tend to satisfy the follow relationship and adhere to some predefined relationship condition $R$ between two events, where $e_1 \in E^{j_1}$ and $e_2 \in E^{j_2}$. There can be multiple categories of relationships that can be associated between continuous spatio-temporal variables that can be explored and the relationship condition $R$ between event pairs can then be tailored to contribute to the desired event relationship under study.

For simplicity, we only consider the following kinds of relationships between variables as meaningful which we attempt to discover:

- $(X_{j_1} \uparrow \rightarrow X_{j_2} \uparrow)$, if Increase in $X_{j_1}$ $\rightarrow$ Increase in $X_{j_2}$
• \((X^{j_1} \uparrow \rightarrow X^{j_2} \downarrow)\), if \(\text{Increase in } X^{j_1} \rightarrow \text{Decrease in } X^{j_2}\)
• \((X^{j_1} \downarrow \rightarrow X^{j_2} \uparrow)\), if \(\text{Decrease in } X^{j_1} \rightarrow \text{Increase in } X^{j_2}\)
• \((X^{j_1} \downarrow \rightarrow X^{j_2} \downarrow)\), if \(\text{Decrease in } X^{j_1} \rightarrow \text{Decrease in } X^{j_2}\)

They can be further clubbed into the following two major event relationships, which can be studied in the scope of the project.

• \((X^{j_1} \uparrow \rightarrow X^{j_2})\), if \(X^{j_1}\) and \(X^{j_2}\) are positively correlated
• \((X^{j_1} \downarrow \rightarrow X^{j_2})\), if \(X^{j_1}\) and \(X^{j_2}\) are negatively correlated

More detailed and complex event relationships can be studied in future work.

3.1.5 Threshold
A user-defined threshold on the interestingness measure of a pattern has to be specified to prune the total number of possible patterns so as to extract only meaningful patterns from the entire pattern search space. Also, a limit on the maximum length of a sequential pattern that can be discovered can be specified.

3.2 Output
Extracted set of all sequential patterns in global spatio-temporal variables exceeding a user-defined interestingness measure threshold along with an statistical account of the time lags for each extracted sequential pattern. Examples of such sequential patterns include:

• \(X^1 \uparrow \rightarrow X^2\),
• \(X^1 \uparrow \rightarrow X^2 \rightarrow X^3\), and
• \(X^2 \rightarrow X^5 \rightarrow X^4\).

3.3 Objective and Constraints
The objective of the problem is to develop an interestingness measure for sequential patterns in global variables and utilize it for designing an algorithm for mining sequential patterns that is correct, complete, minimizes computational time and is scalable. Also, statistical methods have to applied on the time lag distributions of each sequential pattern to produce meaningful inferences about the dynamics of the extracted sequence. The constraints include the fact that the length of the time series for each global variable is constant and is the same at every location, and that we are only interested in mining sequential patterns with either positively- or negatively-correlated relationships. A correlation value of 0 between events is assumed to carry no information about their relationship.
4 Challenges

The major challenges that need to be addressed are

- **Modeling relationships between variables** - In the case of continuous interval-based attributes, there is a variety of relationships that can be defined between events occurring across multiple variables. For example, an increase in \( X^{j_1} \rightarrow \) increase in \( X^{j_2} \) and so on. This creates a need for coming up with new pattern semantics to associate relationships between events and a new interestingness measure to attribute the sequential relationship of variables in a given pattern. In our case, we study two kinds of relationships - positively-correlated relationship (\( X^{j_1} \uparrow \rightarrow X^{j_2} \)), and negatively-correlated relationship (\( X^{j_1} \downarrow \rightarrow X^{j_2} \)).

- **Exponential Search Space** - With the introduction of new pattern semantics to capture relationships between continuous interval-based variables, we observe that the search space of all possible patterns exponentially increases. For the simple case of positively-correlated and negatively-correlated relationships, we can see that the search space becomes of the order of \( 2^{2^n} \), where \( n \) is the length of the maximal sequential pattern. Hence, searching for each possible sequential pattern is practically infeasible and we would have to develop computational algorithms to prune the search space extracting only the meaningful patterns.

5 Related Work

Given the existing literature for spatio-temporal sequential pattern mining, we can broadly classify this area into four broad categories.

The first class of methods focuses on sequential pattern mining using association rules. Association rules have been extensively for finding frequent itemsets (groups of items) occurring in a transaction specially in market-basket data analysis [1]. Sequential pattern mining using association rules were further developed to capture the temporal associations between transactions occurring in a sequence [2, 10, 11]. However, they depend on the notion of a 'transaction' in the sequential pattern mining process. The 'transactionization' of events that is applied in market-basket data analysis cannot be applied in the context of mining spatio-temporal patterns, owing to the continuous nature of space and time.

The second class of methods addresses the problem of mining trajectory patterns from historical spatio-temporal data. Given the trajectory information of multiple objects moving a spatial domain, they find frequently occurring sequential traveling patterns of objects moving in time [7, 3, 4]. However, their approach differs from our problem as it requires the trajectory information of the same object to be provided apriori, which is not applicable to the case of mining sequential patterns in events occurring in spatio-temporal global variables.
The third class of methods finds sequential patterns in spatio-temporal datasets, where the events are instantaneous and event types are categorical [5]. However, they are event-centric models where the events are of boolean nature, characterized by either their occurrence or absence and hence is not fit for our case.

The fourth class of problems involves interval-based events that can be represented as a set of continuous values occurring across multiple spatio-temporal global variables. This class of problems is the focus of our project.

6 Approach

The proposed approach can be broken down into the following sub-procedures.

6.1 Event Detection

Given a time series $X^j_i$ for the $j^{th}$ global variable at location $i$, an event detection scheme partitions the time series into separate time intervals $e^j_{ip}$, $1 \leq j \leq J$, $1 \leq i \leq I$, and $1 \leq p \leq P$, where each time interval represents a distinct event in $X^j_i$. A number of techniques can be employed for detecting events in such fashion and thus partitioning the time series into separate disjoint time intervals. Some of the methods which can be explored in our context are:

- Segmentation Methods
- Anomaly Detection
- Fourier Transforms
- Regression-based Models
- Spatio-temporal Scan Statistic

The choice of the appropriate event detection scheme can be explored according to the requirement of the problem. For simplicity, we consider a standard segmentation algorithm as input to our problem for detecting segments in a continuous global variable at a given location representing distinct events. Such a method would take into account the seasonality and the noise inherently present in the global variable for performing event detection.

Once the global variables have been partitioned into distinct interval-based events, let us define $E^j$ as the set of all possible event instances (time intervals) which occur in the $j^{th}$ variable at all locations.

6.2 Assigning relationships between events

An array of proximity measures can be considered to quantify the relationship between events $e_1$ and $e_2$ occurring at different locations and in different global variables depending on the prototype of events obtained by event detection.
scheme, the desired pattern semantics and the requirements of the application problem. In our example, where we are interested in the positively-correlated and negatively-correlated relationships between global variables, we can look at the following measures to capture such relationships:

- **Pearson’s Correlation Coefficient**
  The correlation between two time series $e_1(t), 0 \leq t \leq t_1$ and $e_2(t), 0 \leq t \leq t_2$, defined as
  \[
  M(e_1, e_2) = \frac{\text{cov}(\hat{e}_1, \hat{e}_2)}{\sigma_{\hat{e}_1}, \sigma_{\hat{e}_2}}
  \]  
  (1)
  where $\text{cov}$ denotes the covariance between two time series, $\sigma$ is the standard deviation of a single time series. Since the events can span over different lengths of time intervals, $t_1$ and $t_2$, we can use their standardized transformations, $\hat{e}_1$ and $\hat{\epsilon}_1$, by representing them on a standard time interval length ($t_s$). Cross-correlation can also be employed to find correlation between (interval-based) events occurring with a time lag.

- **L1 (Manhattan) Distance**
  \[
  M(e_1, e_2) = ||\hat{e}_1 - \hat{e}_2||_1
  \]  
  (2)
  where $||.||_1$ denotes the $L_1$ norm of a vector while $\hat{e}_1$ and $\hat{e}_2$ are the mean-centered and standardized representations of $e_1$ and $e_2$ over a common time interval length ($t_s$).

- **Rank Correlation Measures**
  To measure the relationship between two time series occurring at a time lag, we can look at their ranked representations and explore their rank correlation coefficients such as:

  **Spearman’s Correlation Coefficient**  It is the correlation measure between the ranked representations of two variables.
  \[
  M(e_1, e_2) = \frac{\text{cov}(\hat{e}_1, \hat{e}_2)}{\sigma_{\hat{e}_1}, \sigma_{\hat{e}_2}}
  \]  
  (3)
  where $\hat{e}_1$ and $\hat{e}_2$ are the ranked representations of time intervals $e_1$ and $e_2$. The Spearman’s Correlation Coefficient captures the non-linear monotonic dependence of $e_1$ and $e_2$ which correlation fails to capture.

  **Kendall Tau Coefficient**  It captures the similarity of the orderings of the time series when ranked by each of their continuous values. It looks at possible pairs ($n_p$ in number) with values from each of the two ranked interval-based events and calls them concordant ($n_c$ in number) if they
agree with each other’s ordering, or discordant \((n_d \text{ in number})\) if they disagree with each other’s ordering. The Kendall Tau Coefficient is then given by

\[
M(e_1, e_2) = \frac{n_c - n_d}{n_p}
\] (4)

**Gamma Test Statistic** It is similar to the Kendall Tau measure and is given by

\[
M(e_1, e_2) = \frac{n_c - n_d}{n_c + n_d}
\] (5)

where \(n_c\) is the total number of concordant pairs and \(n_d\) is the total number of discordant pairs.

- **Discretization Methods**
  
  An alternate approach is to classify each interval using a vocabulary of known event classes such as *increase, decrease, sharp increase, sharp decrease, upward spike, downward spike, and constant*. Relationships between event classes across multiple global variables can then be studied. However, such a method would suffer from lack in generality as the event vocabulary is too specific on the particular use-case and would have to be redefined for a new use-case.

For simplicity, we consider the Pearson’s Correlation between two events as the measure to capture positively-correlated and negatively-correlated relationships between global variables in our case.

### 6.3 Defining Interestingness Measure

We first consider size-2 patterns. Without loss of generality, let the pattern under study be represented as \(X^{j_1} \rightarrow X^{j_2}\), where \(X^{j_1}\) and \(X^{j_2}\) are respectively the \(j_1^{th}\) and \(j_2^{th}\) global variables. Let \(E^{j_1}\) and \(E^{j_2}\) denote the set of all possible event instances which occur in the \(j_1^{th}\) and \(j_2^{th}\) variables respectively at all locations. We are interested in finding relationships of the form \(E^{j_1} \rightarrow E^{j_2}\).

Let us consider all event pairs, \((e_{j_1}^{i_1}, e_{j_2}^{i_2})\) such that \(e_{j_1}^{i_1} \in E^{j_1}\) occurs at a given location and time, and \(e_{j_2}^{i_2} \in E^{j_2}\) occurs in the spatial neighborhood of \(e_{j_1}^{i_1}\) and with a positive time lag. Let the complete set of such event pairs be denoted by \(\text{Pair}(E^{j_1}, E^{j_2}) = \{(e_{j_1}^{i_1}, e_{j_2}^{i_2}) | i = 1, 2 \ldots n\}\) with the set of their corresponding time lags between \(e_{j_1}^{i_1}\) and \(e_{j_2}^{i_2}\) denoted by

\[
T(E^{j_1}, E^{j_2}) = \{\text{TimeLag}(e_{j_1}^{i_1}, e_{j_2}^{i_2}) | i = 1, 2 \ldots n\}
\] (6)

We compute the correlation measure between each such event pair and call the set of corresponding correlation values as

\[
M(E^{j_1}, E^{j_2}) = \{M(e_{j_1}^{i_1}, e_{j_2}^{i_2}) | i = 1, 2 \ldots n\}
\] (7)
We observe that the interestingness measure \( I(P) \) exhibits weak anti-monotonic property which can be exploited to construct the Weighted Participation Index (WPI). The WPI of a size-\( k \) sequence, \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_k} \) is the minimum \( WPM \) of the event set \( E_{j_1} \) and \( E_{j_2} \) in the required relationship.

Further, let \( M^+/-(E_{j_1}, E_{j_2}) \) denote the set of all \( M(e_{j_1}^i, e_{j_2}^i) \) values which are either \( \geq 0 \) or \( \leq 0 \). More formally,

\[
M^+(E_{j_1}, E_{j_2}) = \{ M(e_{j_1}^i, e_{j_2}^i) \mid P(e_{j_1}^i, e_{j_2}^i) \in \text{Pair}(E_{j_1}, E_{j_2}), \text{and } M(e_{j_1}^i, e_{j_2}^i) \geq 0 \} \tag{8}
\]

\[
M^-(E_{j_1}, E_{j_2}) = \{ M(e_{j_1}^i, e_{j_2}^i) \mid P(e_{j_1}^i, e_{j_2}^i) \in \text{Pair}(E_{j_1}, E_{j_2}), \text{and } M(e_{j_1}^i, e_{j_2}^i) \leq 0 \} \tag{9}
\]

We observe that only event pairs in \( M^+(E_{j_1}, E_{j_2}) \) (event pairs with positive correlation values) contribute to the positively-correlated relationship, \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_2} \), whereas event pairs in \( M^-(E_{j_1}, E_{j_2}) \) (event pairs with negative correlation values) contribute to the negatively-correlated relationship \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_2} \).

We now define the Weighted Participation Measure (WPM) for the event set \( E_{j_1} \) in the relation \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_2} \) as

\[
WPM(E_{j_1}, (E_{j_1} \rightarrow \ldots \rightarrow E_{j_2})) = \sum_{j} \frac{M^+/-(E_{j_1}, E_{j_2})}{|E_{j_1}|} \tag{10}
\]

which is the sum of all correlation values of event pairs that contribute to the positively-correlated/negatively-correlated relationship, normalized by the total number of unique events in \( E_{j_1} \). Similarly,

\[
WPM(E_{j_2}, (E_{j_1} \rightarrow \ldots \rightarrow E_{j_2})) = \sum_{j} \frac{M^+/-(E_{j_1}, E_{j_2})}{|E_{j_2}|} \tag{11}
\]

The Weighted Participation Index (WPI), which is the interestingness measure of the size-2 sequence \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_2} \) can then be defined as

\[
WPI(E_{j_1} \rightarrow \ldots \rightarrow E_{j_2}) = \text{Minimum}\{ WPM(E_{j_1}, (E_{j_1} \rightarrow \ldots \rightarrow E_{j_2})), WPM(E_{j_2}, (E_{j_1} \rightarrow \ldots \rightarrow E_{j_2})) \} \tag{12}
\]

which is the minimum \( WPM \) of \( E_{j_1} \) and \( E_{j_2} \) in the required relationship.

For sequences of size \( \geq 3 \), we define the tail event set of a size-\( k \) sequence \( E_{j_1} \rightarrow \ldots \rightarrow E_{j_k} \) as the set of all events \( \in E_{j_{k+1}} \), participating in the given size-\( k \) sequence. We denote this tail event set as \( \text{Tail}(E_{j_1} \rightarrow \ldots \rightarrow E_{j_k}) \). The WPI of a size-\( k \) sequence, \( k \geq 3 \) is then given as:

\[
WPI(E_{j_1} \rightarrow \ldots \rightarrow E_{j_k}) = \min\{ WPI(E_{j_1} \rightarrow \ldots \rightarrow E_{j_{k-1}}), WPI(\text{Tail}(E_{j_1} \rightarrow \ldots \rightarrow E_{j_{k-1}}), E_{j_k}) \} \tag{13}
\]

### 6.4 Pattern Mining Algorithm

We observe that the interestingness measure \( WPI \) (Weighted Participation Index) exhibits weak anti-monotonic property which can be exploited to construct the following Apriori-based algorithm for mining all frequent sequential patterns of concern in our domain.
\[ k = 1 \]
\[ F_k = \{ \text{Set of all Global variables} \} \]
\[ \text{while} (F_k \neq \phi) \]
\[ F_{k+1} = \phi \]
for each pattern \( f_i \) in \( F_k \)
for each pattern \( f_j \) in \( F_1 \)
\[ C_{k+1} = \phi \]
Generate candidate pattern \( C_{k+1} = f_i \cup f_j \)
if \( WPI(C_{k+1}) \geq \text{Threshold} \)
\[ F_{k+1} = F_k \cup C_{k+1} \]
\[ T_{k+1} = T_k \cup T(C_{k+1}) \]
endif
endfor
endfor
\[ k = k+1 \]
endwhile

We start with the set of all size-1 sequences, which consists of the set of all available global variables. At the \( k^{th} \) iteration, let \( F_k \) be the set of all \( k \)-size sequences which are frequent (\( WPI \) is above a given threshold). For each such frequent sequence in \( F_k \), we add every global variable \( f_j \) to \( F_1 \), creating a candidate \( k+1 \)-size sequence \( C_{k+1} \) for each \( f_j \). If \( WPI(C_{k+1}) \) is greater than the user-defined threshold, then \( C_{k+1} \) is a frequent \( k+1 \)-size sequence and we update the set of all frequent \( k+1 \)-size sequences, \( F_{k+1} \), by adding \( C_{k+1} \) as \( F_{k+1} = F_k \cup C_{k+1} \). For every such frequent sequence retrieved, we also keep an account of the distribution of time lags for each event instance participating in the frequent sequence. The distribution of time lag for a candidate sequence \( C_{k+1} \) is denoted by \( T(C_{k+1}) \). The algorithm stops when \( F_{k+1} \) is empty and maximal frequent sequences cannot be further extended.

### 6.5 Validation Framework

For our use-case, we intend to evaluate our algorithm on the peatland forests of Indonesia, which are mostly present in the southern parts of Borneo. We would like to study the entire domain and generate global patterns at this scale, along with keeping an account of their corresponding time lag distribution. However since finding a ground truth information for the occurrence of events is difficult to obtain, the validation would depend on looking at results from news articles or satellite imagery evidencing the sequence of events extracted by the algorithm. Apart from testing the hypothesis that Deforestation leads to Forest Fires in Indonesia, the pattern mining approach also offers the possibility of forming new hypotheses, which could open up new possibilities of research for understanding such sequential patterns.

The distribution of time lags for each extracted frequent sequential pattern is also of prime interest which would inform us how closely in time did events in
the sequence influence the occurrence of the events that followed in the sequence. The statistical properties of the time lag distribution for each sequence can be looked at, such as mean and standard deviation. Also, statistical tests such as the Z-test can be performed to look at the statistical significance of the time lag, and as a measure of the influence of an event in the occurrence of the events that followed.

We would also like to note that the sequential patterns obtained at a global scale might not be truly representative of the sequences of events happening at a finer scale. By Simpson’s Paradox, patterns obtained are highly dependent on the scale or resolution of the data, and hence, sequential patterns obtained at a global scale might not be visible on a finer scale if we consider a small spatial domain. Hence, incorporating the scale at which the pattern was obtained is important and the patterns can be obtained at varying scales, from fine to coarse. Changes in the extracted patterns and their corresponding time lag information by changing the scale would be interesting to study.

7 Contributions

We explore a new domain of sequential spatio-temporal pattern mining wherein the events are spread over a time interval and consist of a series of continuous values at multiple global variables. We introduce the notion of an event in our case, and the relationship between events occurring in multiple variables at neighboring locations with a positive time lag. Two new relationships, the positively-correlated relationship and the negatively-correlated relationship were defined for our problem. We define a novel interestingness measure suited for our problem termed as the Weighted Participation Index. We exploit the weak anti-monotonicity property of this measure to devise a pattern mining algorithm which is correct and complete. This opens up a new definition of patterns and a new area of sequential pattern mining research which has to be explored.

8 Future Work

The datasets that would serve the purpose for our use-case have to be explored, such as Vegetation, Temperature, Precipitation etc. We also have to search for ground truth datasets that would assist in validating our results. Alternate techniques of event detection can be explored. Also, we can look at alternate interestingness measure definitions. We can consider pairs of sequences of events instead of event pairs in the computation of our interestingness measure. Implementation issues have to be handled and the parameters for the threshold have to be tweaked.
References


