

Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining: A Summary of Results

Mete Celik¹ Shashi Shekhar¹ James P. Rogers² James A. Shine²

¹Department of Computer Science, University of Minnesota, MN, USA
{mcelik, shekhar}@cs.umn.edu

²U.S. Army ERDC, Topographic Engineering Center, VA, USA
{james.p.rogers.II, james.a.shine}@erdc.usace.army.mil

Abstract

Sustained emerging spatio-temporal co-occurrence patterns (SECOPs) represent subsets of object-types that are increasingly located together in space and time. Discovering SECOPs is important due to many applications, e.g., predicting emerging infectious diseases, predicting defensive and offensive intent from troop movement patterns, and novel predator-prey interactions. However, mining SECOPs is computationally very expensive because the interest measures are computationally complex, datasets are larger due to the archival history, and the set of candidate patterns is exponential in the number of object-types. We propose a monotonic interest measure for mining SECOPs and a novel SECOP mining algorithm. Analytical and experimental results show that the proposed algorithm is correct, complete, and computationally faster than related approaches. Results also show the proposed algorithm is computationally more efficient than naive alternatives.

1. Introduction

Sustained emerging spatio-temporal co-occurrence patterns (SECOPs) represent subsets of object-types that are increasingly located together in space and time. Formally, given a collection of Boolean spatio-temporal (ST) features (object-types) and their instances (objects) over a common ST framework, a neighborhood relation over neighbors, and interest measure thresholds, an SECOP mining algorithm aims to discover correct and complete sets of interesting and non-trivial SECOPs while minimizing computational cost.

Discovering and characterizing SECOPs is an important problem with many application domains [13], including public health (predicting emerging infectious disease), the military (battlefield planning and strategy), ecology (tracking species and pollutant movements), and homeland defense (looking for significant “events”, biodefense)

However, discovering SECOPs poses several challenges. The first challenge is that the process is computationally very expensive since the interest measures

are computationally complex. The second challenge is creating and formalizing composite interest measures to mine interesting and non-trivial SECOPs, since current interest measures such as the spatial prevalence measure are not sufficient to mine such patterns [10, 11]. The third challenge is the exponential number of object-types in the set of candidate patterns. The fourth challenge is to develop computationally efficient algorithms to mine massive spatio-temporal datasets. This paper describes an approach which meets all of these challenges.

An Application Domain Example: SECOPs are of great concern in public health, where there is a frequent need to identify emerging infectious diseases in order to take timely action [7, 14]. Emerging infectious diseases (EIDs) are diseases whose incidence has increased within the past two decades and that threaten to increase in the near future [14]. EIDs can be caused by previously undetected microbes (i.e. SARS, AIDS), the evolution of previously known microbes (Influenza), the spreading of known microbes to new locations or populations (West Nile Virus), and the re-emergence of old infections such as tuberculosis. The World Health Organization reported that approximately 26 percent of the 57 million annual deaths worldwide in 2002 were caused by infectious diseases, the second cause of death after cardiovascular disease [3]. Examples of EIDs throughout the world can be seen in Figure 1.

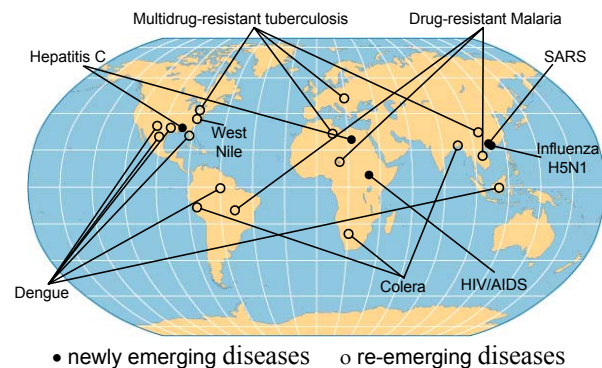


Figure 1: Examples of EIDs. Adapted from [7]

Populations around the world have had to contend with a number of recent outbreaks of EIDs. In a 2005 outbreak of dengue fever in Singapore, the number of cases (instances) rose rapidly throughout the year. Because of the sudden emergence of this disease, hospitals were forced to cancel some surgeries to provide more beds for dengue patients [1]. Figure 2 shows the weekly dengue fever cases in 2005 in Singapore.

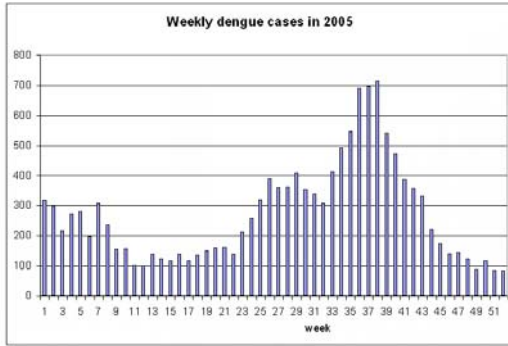


Figure 2: Weekly dengue fever cases in 2005 in Singapore [1]

Scientists are trying to predict or to prevent EIDs. The main focus is to discover the patterns that have cause or effect on EIDs. For example, it is found out that there is a strong relationship between increase of dengue fever cases, stagnant water (i.e., flowerpots), and high temperature. These phenomena are highly likely to co-occur. Similarly, influenza and migrating birds (i.e., waterfowl, shorebirds) frequently co-occur. SECOPs are subsets of object-types e.g., {dengue_disease, stagnant_water, high_temperature} and {influenza_disease, migrating_birds}, whose instances are increasing over time. Discovering such SECOPs is crucial for timely response to outbreaks of diseases and to be fully prepared for such outbreaks. For example, the dengue fever outbreak was controlled after eliminating stagnant water in Singapore; stagnant water is a breeding ground for mosquitoes transmitting the disease.

Another kind of emerging pattern can occur when certain invasive species are introduced at a location where they did not previously occur naturally. For example, the increase in the population of Brown Tree Snake that was accidentally transported to the snake-free Guam Island after World War II, increased the presence of insect pests, which has caused forest defoliation and decrease in crop yields [2]. The invasive Brown Tree Snake preyed upon most of the insectivorous birds, bats and lizards on the island. The decimation of insectivorous birds, bats and lizards caused an increase in presence of insect pests. It is found out that there is a strong relationship between increase of Brown Tree Snake population and increase of presence of insect pests. Discovering such SECOP i.e. {brown_tree_snake,

insect_pests} is important to prevent harmful affects on the ecology and environment.

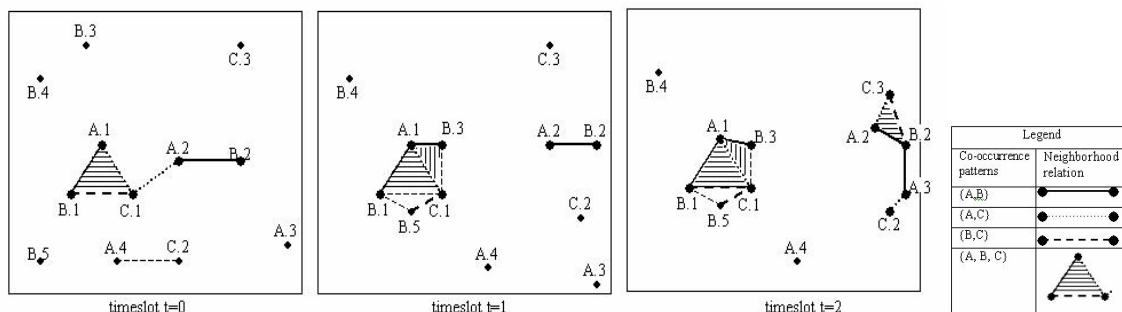
Related Work: The closest related work focuses on spatio-temporal episode “formation” to identify pattern characterizing evolution of spatial relationship objects (or features) over time [16]. A “formation “event occurs when the number of instances of a pattern changes from zero to non-zero. The algorithm to efficiently mine “formation” events is not specified explicitly. This model is limited due to several reasons. First “formation” is different from “emerging”. While a sustained emerging pattern may imply a “formation”, there may be a long time-lag between “formation” and emergence. Furthermore, a formation may not lead to sustained emergence. In addition, detection of “formation” may require tracking extremely rare patterns (i.e. pattern with very low number of instances), thus hampering prevalence-based filtering [4] and leading to exorbitant computational costs. Support and realization were used in [16] as interest measures to characterize the importance of the “formation” pattern. The support of “formation” pattern p is defined as the number of timeslots (snapshots) in the database where p occurs. The realization of “formation” pattern p is defined as the minimum of the number of instances of p in each timeslot. However, scaling the algorithm to large spatio-temporal databases is challenging since the interest measure used in [16], i.e. realization, is not monotonic.

In contrast, this paper defines sustained emerging co-occurrence patterns, formulates new monotonic interest measures, and proposes algorithms to mine SECOPs from massive spatio-temporal datasets in a computationally efficient manner.

Contributions: The paper makes the following contributions:

1. It defines sustained emerging spatio-temporal co-occurrence patterns (SECOPs) and the SECOP mining problem.
2. It proposes a new monotonic composite interest measure to discover and mine SECOPs.
3. It proposes a novel and computationally efficient SECOP mining algorithm (SECOPs -Miner).
4. It shows that the proposed algorithm is correct and complete in finding sustained emerging prevalent (e.g., spatial prevalent and time prevalent) SECOPs.
5. It experimentally evaluates the proposed composite interest measures and SECOP mining algorithms using real datasets.

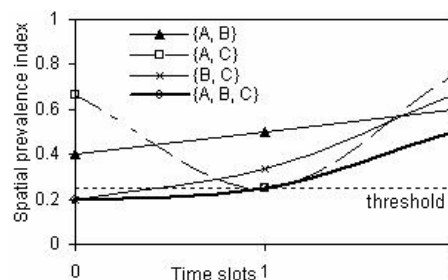
Scope: This paper focuses on the sustained emerging co-occurrence pattern on a typed collection of moving objects extending interest measures for spatial co-location patterns given a user defined participation index threshold [10, 11]. The following issues are outside the scope of this paper: (i) Determining thresholds for SECOP interest measure, (ii) similarity measures for tracking moving



(a) An input spatio-temporal dataset

Co-occurrence Patterns	Spatial prevalence index values			Time prevalence index values
	time slot 0	time slot 1	time slot 2	
A B	2/5	2/4	3/5	3/3
A C	2/3	1/4	3/4	2/3
B C	1/5	1/3	2/3	2/3
A B C	1/5	1/4	2/4	2/3

(b) A set of output SECOPs



(c) Trend of spatial prevalence indices of SECOPs

Figure 3. Spatio-temporal dataset

objects due to the focus on object-types rather than objects, and (iii) indexing and query processing issues related to mining objects.

Outline: The rest of the paper is organized as follows. Section 2 presents basic concepts and the problem statement of mining SECOPs. Section 3 presents our proposed SECOP mining algorithm. Analysis of the SECOP mining algorithm is given in Section 4. Section 5 presents the experimental evaluation and Section 6 discusses the conclusions and future work.

2. Basic concepts and problem statement

2.1 Spatial prevalence measure

The focus of this study is to discover sustained emerging spatio-temporal co-occurrence patterns (SECOPs) over a spatio-temporal framework and a neighborhood relation R . First we explain the modeling of groups of object-types in space, e.g., spatial co-locations, and then we explain modeling SECOPs and propose algorithms to mine SECOPs [11].

Spatial co-location mining algorithms are used to discover sets of object-types that are frequently located together in a spatial framework for a given set of spatial object-types, their instances and a spatial neighbor relationship R [10, 11]. For example, in Figure 3(a), in time slot $t=0$, $\{A.1, C.1\}$ is an instance of a co-location if the distance between the objects is less than or equal to a given neighborhood distance threshold. The lines show the distances between objects which satisfy the

neighborhood distance threshold. The participation index is used to determine the strength of the co-location pattern, that is, whether the index is greater than or equal to a threshold [10, 11]. Such a co-location pattern is called spatial prevalent. The participation index is defined as the minimum of the participation ratios (the fraction of the number of instances on object-types forming co-location instances to the number of instances). For example, in Figure 3(a), $\{A, B\}$ is a co-location for time slot $t=0$, and its instances are $\{A.1, B.1\}$ and $\{A.2, B.2\}$. In the dataset, object-type A has 4 instances and two of them ($A.1$ and $A.2$) are contributing to the co-location $\{A, B\}$, so the participation ratio of A for the co-location $\{A, B\}$ is $2/4$. The participation ratio of object-type B is $2/5$ since 2 out of 5 instances are contributing to form the instances of the co-location $\{A, B\}$. The participation index of the co-location $\{A, B\}$ is $2/5$, which is the minimum of the participation ratios of object-types A and B .

It has been shown that the participation index is anti-monotonic with respect to the size of co-location patterns [10, 11]. In other words,

$participation_index(P_j) \leq participation_index(P_i)$ if P_i is a subset of P_j . In addition, shows that the participation index has a spatial statistical interpretation as an upper bound on the cross- K function [5, 10, 11].

2.2 Modeling SECOPs

Given a set of spatio-temporal object-types and a set of their instances with a neighborhood relationship R , an

SECOP is a subset of spatio-temporal object-types whose instances are neighbors in space and time.

Definition 2.1: *Given a spatio-temporal pattern and a set T of time slots, such that $T=[T_0, \dots, T_{n-1}]$, the time prevalence or persistence measure of the pattern is the fraction of time slots where the pattern occurs over the total number of time slots.*

For example, in Figure 3(a), the total number of time slots is 3 and pattern $\{A, B\}$ occurs in all 3 time slots, so its time prevalence index is $3/3$.

Definition 2.2: *Given a spatio-temporal dataset ST , and a spatial prevalence threshold θ_p , the sustained emergence prevalence measure of a spatio-temporal pattern P_i is a composition of the spatial prevalence measure and the time prevalence measure where the spatial prevalence measure is getting stronger (monotonically increasing) over time, such as*

$Prob_{t_m \in all_time_slot} (s_prev(pattern\ P_i, t_m) \geq \theta_p)$ and $s_prev(pattern\ P_i, t_m)$ monotonically increasing over time

where s_prev stands for spatial prevalence index, e.g., the participation index (described in section 2.1), of pattern P_i and θ_p is a spatial prevalence threshold. $Prob$ stands for the probability of overall prevalence time slots.

Definition 2.3: *Given a spatio-temporal dataset ST and a threshold pair $(\theta_p, \theta_{time})$, a SECOP P_i is a sustained emergence prevalent patterns, if its sustained emergence prevalence measure satisfies the following.*

$Prob_{t_m \in all_time_slot} [s_prev(pattern\ P_i, t_m) \geq \theta_p] \geq \theta_{time}$ and $s_prev(pattern\ P_i, t_m)$ monotonically increasing over time

where s_prev stands for spatial prevalence index, e.g. the participation index (described in section 2.1) of pattern P_i . $Prob$ stands for the probability of overall prevalence time slots, θ_p is a spatial prevalence threshold, and θ_{time} is a time prevalence threshold.

For example, in Figure 3(a), $\{A, B\}$ is a SECOP because: (a) it is spatial prevalent in time slots $t=0, t=1$, and $t=2$ since its participation indices are above the participation index threshold 0.25; and (b) it is time prevalent since its time prevalence index, i.e., 1, is above the time prevalence index threshold 0.5. SECOPs are extensions of traditional spatial co-location patterns, whose instances are increasing over time.

2.3 Problem statement

Given:

- A set P of Boolean spatio-temporal object-types over a common spatio-temporal framework.
- A neighbor relation R over locations.
- A spatial prevalence threshold, θ_p .
- A time prevalence threshold, θ_{time} .

Find: $\{P_i \mid P_i \text{ is a subset of } P \text{ and } P_i \text{ is a sustained emergence prevalent SECOP as in Definition 2.3}\}$.

Objective: Minimize computation cost.

Constraints: To find a correct and complete set of SECOPs.

Example: The spatio-temporal dataset given in Figure 3 (a) contains 3 Boolean object-types, A, B, and C, for 3 time slots. A distance between the objects may define the neighborhood relation R . For example, A.4 is a neighbor of C.2 in time slot 0, but not in time slots 1 and 2. In this example dataset $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, and $\{A, B, C\}$ form a candidate SECOP, given $\theta_p=0.25$, and $\theta_{time}=0.5$. Figure 3(c) gives the trend of the spatial prevalence indices, i.e. participation indices, of the SECOPs. As can be seen pattern $\{A, B\}$ is above the threshold for the time interval $[0,2]$ and rest of the patterns are above the threshold for time interval $[1,2]$.

3. Mining SECOPs

In this section, we first discuss a naïve approach to mining SECOPs and then propose a novel SECOP mining algorithm (SECOP-Miner).

Naïve approach: A naïve approach can generate all spatial co-locations for each time slot and then can apply a post-processing step to discover sustained emergence prevalent co-occurrence patterns by checking their spatial and time prevalence indices. The naïve approach will generate size $k+1$ candidate co-location patterns for each time slot using size k subclasses until there are no more candidate spatial co-locations. After finding all spatial co-location patterns in each time slot, a post-processing step can be used to discover sustained emergence prevalent SECOPs by pruning out spatial and time non-prevalent co-location patterns. This approach will lead to unnecessary computational costs since it does not prune out sustained emergence non-prevalent SECOPs before the post-processing step.

SECOP-Miner: In contrast, we propose a SECOP mining algorithm (SECOP-Miner) to discover sustained emergence prevalent SECOPs by incorporating a pruning step in to the algorithm. It will generate size $k+1$ candidate SECOPs using size k sustained emergence prevalent subclasses. The participation index is used as a spatial prevalence interest measure to check if the pattern is spatial prevalent at a time slot [10]. The time

prevalence (i.e. persistence measure in definition 2.1) is used as a time prevalence interest measure. First we give the pseudocode of the algorithm, and then we provide an execution trace of the algorithm using the spatio-temporal dataset from Figure 3(a).

Algorithm 1 gives the pseudocode of the SECOP-Miner algorithm. The inputs to the algorithm are a set of spatial event types E , a spatio-temporal dataset ST , a spatial neighborhood relationship R , and thresholds of interest measures such as, spatial prevalence and time prevalence. The output of the algorithm is a set of SECOPs.

In the algorithm, steps 1 and 2 include initialization, steps 3 through 12 give an iterative process to mine SECOPs, and step 13 gives a union of the results of the iterative steps. Steps 3 through 12 continue until there is no candidate SECOP to be generated (mined). The functions of the algorithm are explained below.

Generation of candidate co-occurrence patterns (step 5): This function uses an apriori-based approach to generate size $k+1$ candidate co-locations C_{k+1} for each time slot, using all sustained emergence prevalent size k SECOPs EP_k [4].

Generation of spatial co-occurrence instances (step 6): The instances of candidate C_{k+1} are generated by joining neighbor instances of sustained emergence prevalent size k patterns for each time slot. This is similar to the instance generation step of the co-location miner algorithm [10].

Finding spatial prevalent co-location patterns (step 7): All spatial prevalent size $k+1$ patterns SP_{k+1} are found by pruning the patterns whose spatial prevalence indices, i.e., participation indices, are less than a given threshold for each time slot. Computation of participation indices follows the same algorithmic ideas as those in the co-location mining algorithm [10].

In steps 5 through 7, the algorithm finds size $k+1$ spatial prevalent co-location patterns for each time slot.

Finding time prevalence index (step 9): This step checks the behavior of the spatial prevalence index of a pattern over time, which can be classified into three categories: 1) monotonically increasing, such that the prevalence index is getting stronger over time, 2) monotonically decreasing, such that the prevalence index is getting weaker over time, and 3) has extremums oscillating over time. To find SECOPs, the algorithm checks the behavior of the spatial prevalence index of each size $k+1$ co-occurrence pattern. If the spatial prevalence index of a pattern is monotonically increasing over time, the pattern is recorded in candidate SECOP TP_{k+1} ; otherwise it is eliminated. If it is monotonically decreasing over time (that is, the participation index is getting weaker), the pattern is eliminated and is not included in the set TP_{k+1} . If it has extremums (one or more

roots), it is divided into time intervals such that the spatial prevalence index is monotonically increasing or decreasing. The monotonically decreasing parts are eliminated but the rest are saved as candidate SECOPs. After the elimination of the irrelevant patterns, the time prevalence indices of candidate SECOPs are calculated.

Finding sustained emerging co-occurrence patterns (step 10): This step discovers SECOPs by checking the time prevalence indices of the patterns if they are above or equal to a given time prevalence threshold θ_{time} . The patterns whose time prevalence indices do not satisfy the given threshold are pruned. The remaining patterns will be sustained emergence prevalent SECOPs and will be used to generate candidate supersets of the SECOPs in step 5.

The algorithm will run iteratively until there are no candidate SECOPs to be generated. The algorithm outputs the union of all size sustained emergence prevalent SECOPs.

Pseudocode for SECOP-Miner Algorithm	
Inputs:	
E:	a set of spatial event types
ST:	a spatio-temporal dataset <event_type, instance_id, x, y, time>
R :	spatial neighborhood relationship
TF :	a timeslot frame $\{t_1, \dots, t_{n-1}\}$
θ_{time} :	a t-prevalence threshold
S_{time} :	a t-support threshold
Output:	Sustained emerging spatio-temporal co-occurrence patterns (SECOPs) whose spatial prevalence indices, i.e., participation indices, are greater than θ_p and increase monotonically over time, for time prevalence indices is greater than θ_{time} .
Variables	
k:	co-occurrence size
T_1 :	set of instances of size-k co-occurrences
C_k :	set of candidate size-k co-occurrences
SP_k :	set of spatial prevalent size-k co-occurrences
TP_k :	set of time prevalent size-k co-occurrences
EP_k :	set of sustained emerging size-k co-occurrences
Algorithm	
1.	Initialize parameters
2.	Co-occurrence size $k=1$, $C_1=E$, $EP_1=ST$
3.	while (not empty EP_k) {
4.	For each time slot {
5.	$C_{k+1}=\text{gen_candidate_co-occur}(C_k, EP_k, \text{time_slot})$
6.	$T_{k+1}=\text{gen_co-occur_instance}(C_{k+1}, T_k, \text{time_slot}, R)$
7.	$SP_{k+1}=\text{find_spatial_prevalent_co-occur}(C_{k+1}, T_{k+1}, TF, \theta_p)$
8.	}
9.	$TP_{k+1}=\text{find_time_prevalence_index}(SP_{k+1})$
10.	$EP_{k+1}=\text{find_emergence_prevalent_co-occur}(TP_{k+1}, \theta_{time})$
11.	$k=k+1$
12.	}
13.	return union (EP_2, \dots, EP_{k+1})

Algorithm 1. SECOP-Miner algorithm

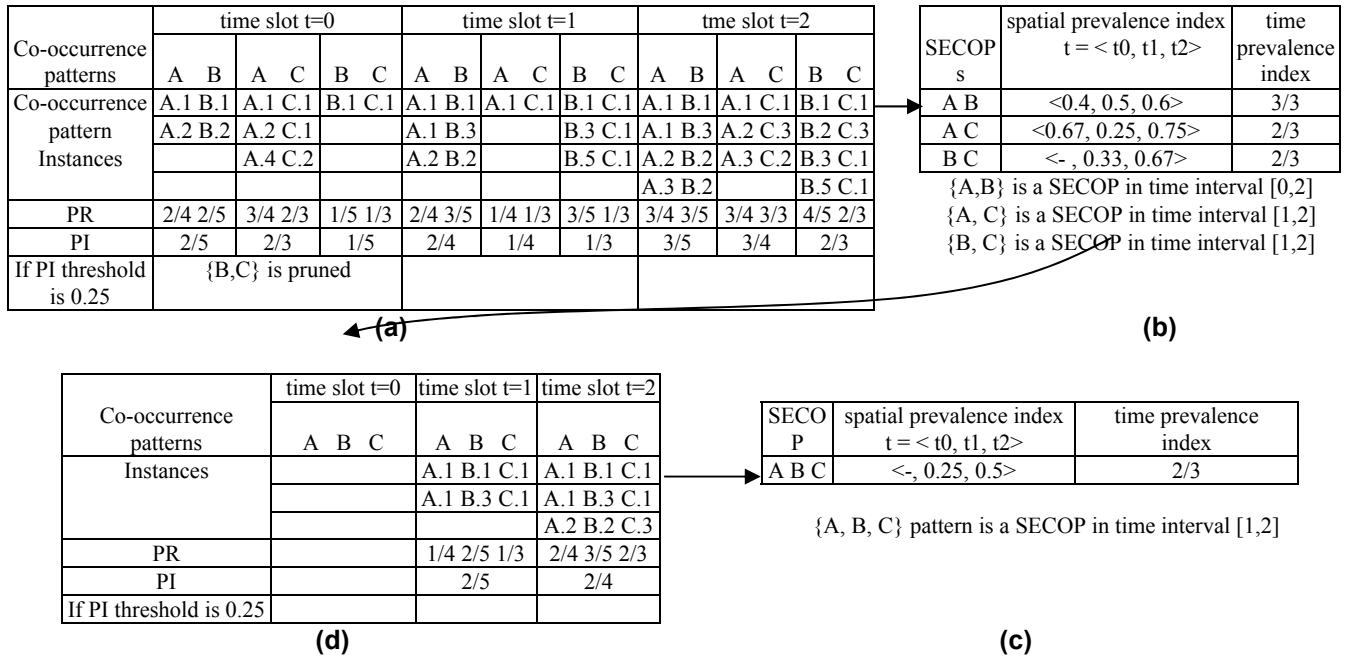


Figure 4. An example of mining spatio-temporal co-occurrence patterns (STCOPs).

An Execution Trace: The execution trace of the algorithm is given in Figure 4 using the spatio-temporal dataset given in Figure 3(a). The dataset contains three object-types A, B, and C and their instances in three time slots. A has 4 instances, B has 5 instances, and C has 3 instances. Each instance has a unique identifier, such as A.1. Some of the patterns of these object-types form a SECOP. To discover SECOPs we compute the sustained emergence prevalence measure, which is a composition of the spatial prevalence and time prevalence measures. The spatial prevalence measure, (participation index) shows the strength of the spatial co-location pattern and whether the index is greater than or equal to a given threshold. The time prevalence measure, (time prevalence index), shows the frequency of the pattern over time.

In Figure 4(a), candidate spatial co-location pairs of the object-types and their instances are generated for distinct time slots and then the spatial co-location patterns whose participation indices are less than a given threshold are pruned since they are spatial non-prevalent. A spatial non-prevalent co-location pattern {B, C} is pruned in time slot t=0 because its participation index is less than the given threshold 0.25.

SECOPs whose participation indices are increasing over time are then determined and their time prevalence indices are calculated. For example, in Figure 4(b), the time prevalence index of pattern {A, B} is 3/3 because it is spatial prevalent in all time slots and its participation indices increase monotonically over time. The time prevalence indices of pattern {A, C} and {B, C} are 2

since they are SECOPs in time interval [1,2]. Pattern {A, C} is not a SECOP in time interval [0,1] since its participation index is decreasing from time slot t=0 to time slot t=1. Pattern {B,C} is not a SECOP in time interval [0,1] since it is not spatial prevalent in that interval. The SECOPs whose time prevalence indices are greater than or equal to a given time prevalence index threshold are selected for generating superset candidate patterns. Spatial prevalent patterns {A, B}, {A, C} and {B, C} are selected as sustained emergence prevalent SECOPs since they are also time prevalent (they exceed the time prevalence index of 0.5). Using SECOPs {A,B}, {A, C}, and {B, C}, the next candidate pattern {A, B, C} is generated.

The next step is to generate instances of candidate pattern {A, B, C} in time slots where its subsets exist and to check its participation indices in the corresponding time slots. Since all subsets of SECOP {A, B, C} are sustained emergence prevalent and exist in time slots t=1 and t=2, there is no need to generate instances for time slot t=0. The instances of candidate SECOP {A,B,C} are generated and participation indices of pattern {A, B, C} are found, which are 2/5 and 2/4 for time slots t=1 and t=2 respectively (Figure 4(b)). Since both participation indices are greater than the spatial prevalence threshold 0.25, pattern {A, B, C} is spatial prevalent in these time slots.

In Figure 4(b), SECOP {A,B,C} is determined since its spatial prevalence indices, i.e., participation indices, are increasing over time over given threshold 0.5. Since

there are not enough subsets to generate the next candidate SECOPs, the algorithm stops at this stage and outputs the union of all sustained emergence prevalent SECOPs: $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, and $\{A, B, C\}$. Figure 3(c) gives the trend of the spatial prevalence indices, i.e. participation indices, of the SECOPs. As can be seen pattern $\{A, B\}$ is above the threshold for the time interval $[0,2]$ and rest of the patterns are above the threshold for time interval $[1,2]$.

4. Analysis of the SECOP mining algorithm

4.1 Sustained emergence prevalence measure is monotonic

Lemma 4.1: *A chosen spatial prevalence measure, such as, participation index, is monotonically non-increasing in the size of the SECOPs at each time slot.*

Proof: Let a SECOP P_i be a subset of a SECOP P_j . Then, $participation_index(P_j, t) \leq participation_index(P_i, t)$

This follows from the anti-monotone property of the participation index for co-location patterns using the data subset for time slot t [10]. If an instance I of object-type O in intersection (P_j, P_i) participates in any instance of P_j , I must participate in some instance of P_i , as well. Thus,

$$participation_index(O, P_j, t) \leq participation_index(O, P_i, t)$$

for all object-types $O \in intersection(P_j, P_i)$. This implies the lemma. \square

Lemma 4.2: *A sustained emergence prevalence measure is monotonically non-increasing with the size of SECOP over space and time. In other words, if SECOP P_i is a subset of SECOP P_j then*

$Prob_{t_m \in all_time_slot} (s_prev(pattern P_i, t_m) \geq \theta_p)$, and $s_prev(pattern P_i, t_m)$ monotonically increasing over time

$Prob_{t_m \in all_time_slot} (s_prev(pattern P_j, t_m) \geq \theta_p)$ and $s_prev(pattern P_j, t_m)$ monotonically increasing over time

where $Prob$ stands for the probability of overall prevalence time units, s_prev stands for spatial prevalence, θ_p is the spatial prevalence threshold, and t_m is the time slot.

Proof: Let $TS(P_j, \theta_p) = \{t_m \mid participation_index(P_j, t_m) \geq \theta_p\}$ and

$participation_index(P_j, t_m)$ is monotonically increasing over time.

Lemma 4.1 implies that the $participation_index(P_j, t) \geq \theta_p$ for all $t_m \in TS(P_j, \theta_p)$ and is monotonically increasing, since P_i is a subset of P_j . Thus,

$Prob_{t_m \in all_time_slot} [s_prev(pattern P_i, t_m) \geq \theta_p] \geq \theta_{time}$ and $s_prev(pattern P_i, t_m)$ monotonically increasing over time

where θ_{time} is time prevalence threshold. \square

The participation ratio and participation index have anti-monotonic properties as the number of co-occurrences increases, and both have been successfully used in previous studies [10].

4.2. Correctness and completeness

Theorem 4.1: *The SECOP-Miner algorithm is complete.*

Proof: The SECOP-Miner is complete if it finds all sustained emergence prevalent SECOPs that satisfy a given spatial prevalence threshold and time prevalence threshold. We can show this by proving that none of the functions of the algorithm miss any patterns, i.e., filter out a prevalent SECOP. \square

The $gen_candidate_co-occur$ function does not miss any patterns given the anti-monotonic nature of the SECOP interest measure. The input to this function is the sustained emergence prevalent size k SECOPs and the output is candidate size $k+1$ SECOPs. If $c_1 = \{f_1, \dots, f_k\}$ and $c_2 = \{f_1, \dots, f_{k-1}, f_{k+1}\}$ are size k sustained emergence prevalent co-occurrence patterns, candidate size $k+1$ pattern $C_{k+1} = \{f_1, \dots, f_{k-1}, f_k, f_{k+1}\}$ will be produced by joining sustained emergence prevalent size k SECOPs.

The $gen_co-occur_instance$ function does not miss any patterns. This function generates instances of candidate size $k+1$ SECOPs by joining instances of sustained emergence prevalent size k SECOPs if they are in the neighborhood distance and forming a clique.

The $find_spatial-prevalent_co-occur$ function does not miss any patterns. It calculates spatial prevalence indices of the patterns for each time slot and finds spatial prevalent patterns whose participation indices are greater than a given participation index threshold.

The $find_time_prevalence_index$ function does not miss any patterns. This function calculates time prevalence indices of the patterns found in steps 4 through 8 and does not do any pruning.

The $find_emergence-prevalent_co-occur$ function does not miss any SECOPs. The function finds all sustained emergence prevalent SECOPs whose time prevalence index is greater than or equal to a given time prevalence

threshold. SECOPs whose time prevalence indices do not satisfy the given threshold are pruned.

Theorem 4.2: The SECOP-Miner is correct. In other words, if an SECOP pattern P is returned by the SECOP-Miner algorithm, then P is a sustained emergence prevalent SECOP.

Proof: This is easy to establish due to the pruning steps of “find_spatial_prevalent_co-occur” and “find_emergence_prevalent_co-occur” which weed out candidates not meeting the given thresholds. □

5. Experimental evaluation

In this section, we present our experimental evaluations of several design decisions and workload parameters of our SECOP-Miner algorithm. We use a real-world training dataset. We evaluated the behavior of the SECOP-Miner algorithm and naive approach by changing the number of time slots, the number of object-types, and the value of the spatial prevalence and time prevalence measures. Figure 5 shows the experimental setup to evaluate the impact of design decisions of the performance of the proposed algorithms. Experiments were conducted on an Intel Centrino PIV 1.6 GHz computer with 512 MB of RAM.

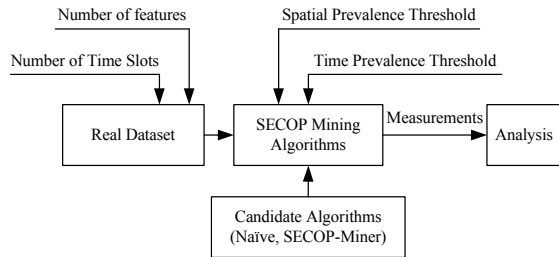


Figure 5. Experimental setup and design

The training dataset contains location and time information of vehicle moving objects. This dataset includes 14 time snapshots and 20 distinct object-types and their instances in each time slot. The minimum instance number is 2, the maximum instance number is 78, and the average number of instances is 18. Figure 6 shows an instance of a SECOP where object_1, and object_2 are coming together, moving from top right to bottom left. Initially the objects are far away from each other but they get relatively close to each other, that is, the pattern emerges over time. Such a pattern may be of interest to a planner if it indicates an emerging imminent maneuver by object_1 under the protection of object_2.

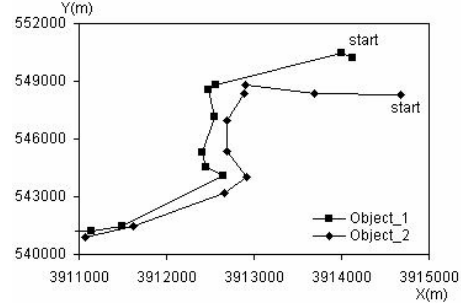


Figure 6. One instance of a SECOP

Figure 7 gives a statistical summary of the participation index of a SECOP {object_1, object_2} for 10 time slots. The participation of the SECOP increases over time between time intervals [4,5] and [6,9]. If the time prevalence threshold is 0.3, the time interval [4,5] will be pruned and the time interval [6,9] will be used for further data generation.

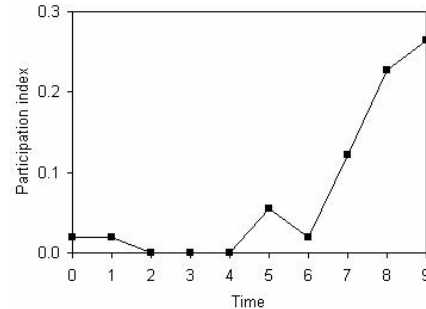


Figure 7. Statistical summary of participation index of SECOP

5.1 Effect of number of timeslots

In the first experiment, we evaluated the effect of the number of time slots on the execution time of the SECOP-Miner algorithm and naive approach. The participation index threshold, time prevalence index threshold, and neighborhood distance threshold were set at 0.2, 0.5, and 100m respectively. As shown in Figure 8, the execution time of the SECOP-Miner algorithm outperforms the naive approach, since it prunes out emergence non-prevalent SECOPs early. It can also be seen that, as the number of time slots increases, the execution time for the naive approach increases drastically.

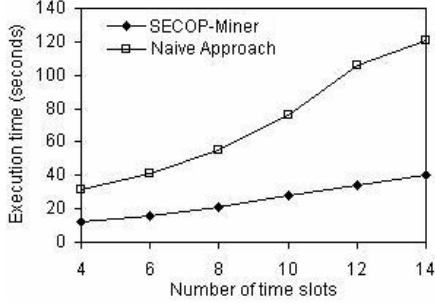


Figure 8. Effect of number of timeslots

5.2 Effect of number of object-types

In the second experiment, we evaluated the effect of the number of object-types on the execution time of two algorithms. The fixed parameters were the participation index threshold, time prevalence index threshold, and neighborhood distance and their values were 0.2, 0.5, and 100m respectively. As shown in Figure 9, the SECOP-Miner algorithm outperforms the naive approach as the number of object-types increases. It can also be seen that, as the number of time slots increases, the execution time for the naïve approach increases drastically.

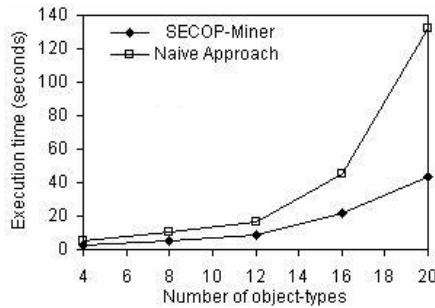


Figure 9. Effect of number of object types

5.3 Effect of time prevalence index threshold

In the third experiment, we evaluated the effect of the time prevalence index threshold value on the execution time of both algorithms. The fixed parameters were participation index threshold, number of time slots, and neighborhood distance and their values were 0.2, 11, and 150m respectively. The effective cost of generation of spatial prevalent co-location patterns on the execution time of the naive approach will be constant since it generates the same number of spatial prevalent patterns as the time prevalence index increases. In that case, the cost of the post-processing step will reflect the trend of the naive approach as the time prevalence index threshold increases. Experimental results show that the SECOP-Miner algorithm is more computationally efficient than the naive approach because of the early pruning strategy.

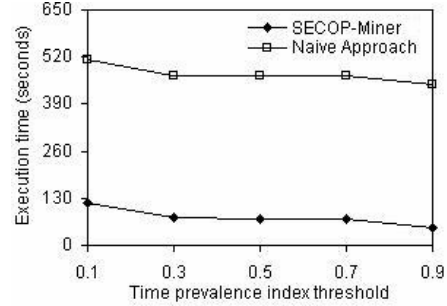


Figure 10. Effect of time prevalence index threshold

5.4 Effect of spatial prevalence index threshold

In the fourth experiment, we evaluated the effect of the value of the spatial prevalence index threshold on the execution times of both algorithms. The fixed parameters are time prevalence index threshold, number of time slots, and distance threshold; the values are 0.5, 11, 150m respectively. As can be seen in Figure 11, the SECOP-Miner algorithm outperforms the naive approach as the spatial prevalence index threshold increases.

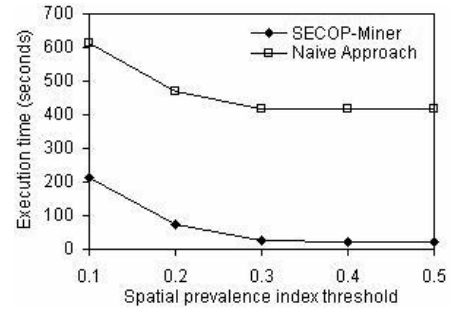


Figure 10. Effect of spatial prevalence index threshold

6. Conclusions and future work

We defined sustained emerging spatio-temporal co-occurrence patterns (SECOPs) and the SECOP mining problem and proposed a new monotonic composite interest measure, the sustained emergence prevalence measure, which is the composition of the spatial prevalence and time prevalence measures. We also devised a novel, computationally efficient SECOP mining algorithm, SECOP-Miner, for mining these patterns. We compared our algorithm with the naive approach, which finds all spatial co-locations of each time slot and then discovers SECOPs by applying sustained emergence prevalence to prune irrelevant co-locations using a post-processing step. We proved that the proposed algorithms are correct and complete in finding sustained emergence prevalent (e.g., spatial-prevalent and time prevalent)

SECOPs. Our experimental results using a real dataset provide further evidence of the viability of our approach.

There are also other studies defining the emerging pattern mining problem in classical data mining, notably [6]. However, the problem defined by Dong et. al. is different than what we are dealing with [6]. The problem they define is to capture the significant differences between two classes, i.e., normal tissues vs cancer tissues. The significant change is measured by a growth rate ratio (ratio of the supports of two classes) [6, 15]. The proposed approaches in [6, 15] are not applicable to the mining of SECOPs from spatio-temporal datasets since they cannot handle the spatial and temporal characteristics of spatio-temporal datasets.

Other studies in the literature, however, have defined spatio-temporal patterns of interest to us. For example, Kalnis et. al. proposed a moving clusters problem and clustering-based algorithms to mine this pattern [12]. Hadjieleftheriou et. al. defined spatio-temporal pattern queries which can use various types of spatial predicates (range search, nearest neighbor, etc) associated with temporal constraints (time-instant or time-interval) and proposed spatio-temporal index structures and algorithms [9]. Gudmundsson proposed algorithms to query flock, leadership, convergence, and encounter patterns from spatio-temporal databases [8]. Yoo et. al. proposed a method to query co-evolving spatial event sets [17]. In the future we plan to apply our algorithms to mine these patterns. We also plan to investigate other pruning methods and test our algorithm on different datasets and to develop new computationally efficient algorithms for mining SECOPs

Acknowledgments

This work was partially supported by the US Army Corps of Engineers under contract number W9132V-06-C-0011, the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, and the Army Research Laboratory (ARL) under contract number DAAD19-01-2-0014.

The authors would like to thank the members of the Spatial Database Group, The authors would also like to thank Kim Koffolt for helping improve the readability of this paper.

References

- [1] 2005 Dengue Outbreak in Singapore, http://en.wikipedia.org/wiki/2005_dengue_outbreak_in_Singapore.
- [2] Invasive species, http://en.wikipedia.org/wiki/Invasive_species.
- [3] The World Health Report 2004 - Changing History, World Health Organization, 2004.
- [4] R. Agarwal and R. Srikant, Fast algorithms for Mining Association Rules, *20th Int'l Conf. on Very Large Data Bases (VLDB)*, 1994.
- [5] N. A. C. Cressie, *Statistics for Spatial Data*, Wiley and Sons, ISBN 0471843369, 1991.
- [6] G. Dong and J. Li, Efficient mining of emerging patterns: Discovering trends and differences, *In Proc Int'l. Conf. Knowledge Discovery and Data Mining KDD'99*,43-52, San Diego, CA, USA, 1999.
- [7] A. S. Fauci, Emerging and Re-emerging Infectious Diseases: The Perpetual Challenge, *2005 Robert H. Ebert Memorial Lecture*, 2006.
- [8] J. Gudmundsson, M. v. Kreveld, and B. Speckmann, Efficient Detection of Motion Patterns in Spatio-Temporal Data Sets, *ACM-GIS*,250-257, 2004.
- [9] M. Hadjieleftheriou, G. Kollios, P. Bakalov, and V. J. Tsotras, Complex Spatio-Temporal Pattern Queries, *VLDB*,877-888, 2005.
- [10] Y. Huang, S. Shekhar, and H. Xiong, Discovering Co-location Patterns from Spatial Datasets: A General Approach, *IEEE Trans. on Knowledge and Data Eng. (TKDE)*, vol. 16(12), pp. 1472-1485, 2004.
- [11] Y. Huang, S. Shekhar, and H. Xiong, Discovering Spatial Co-location Patterns: A Summary of Results, *7th Int'l Symp. on Spatial and Temporal Databases (SSTD)*, L.A., CA, 2001.
- [12] P. Kalnis, N. Mamoulis, and S. Bakiras, On Discovering Moving Clusters in Spatio-temporal Data, *SSTD*, 2005.
- [13] M. Koubarakis, T. Sellis, A. Frank, S. Grumbach, R. Guting, C. Jensen, N. Lorentzos, H. J. Schek, and M. Scholl, *Spatio-Temporal Databases: The Chorochronos Approach*, LNCS 2520, vol. 9, Springer Verlag, 2003.
- [14] D. M. Morens, G. K. Folkers, and A. S. Fauci, The Challenge of Emerging and Re-emerging Infectious Diseases, *Nature*, vol. 430(242-249), 2004.
- [15] L. Wang, H. Zhao, G. Dong, and J. Li, On the complexity of finding emerging patterns, *COMPSAC'04*, 2004.
- [16] H. Yang, S. Parthasarathy, and S. Mehta, A Generalized Framework For Mining Spatio-temporal Patterns in Scientific Data, *ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD)*,716-721, 2005.
- [17] J. S. Yoo, S. Shekhar, S. Kim, and M. Celik, Discovery of Co-evolving Spatial Event Sets, *SIAM Int'l Conf. on Data Mining (SDM)*, Maryland, 2006.