

# Mixed-Drove Spatio-Temporal Co-occurrence Pattern Mining: A Summary of Results

Mete Celik<sup>1</sup> Shashi Shekhar<sup>1</sup> James P. Rogers<sup>2</sup> James A. Shine<sup>2</sup> Jin Soung Yoo<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Minnesota, MN, USA

{mcelik,shekhar,jyoo}@cs.umn.edu

<sup>2</sup>U.S. Army ERDC, Topographic Engineering Center, VA, USA

{james.p.rogers.II,james.a.shine}@erdc.usace.army.mil

## Abstract

*Mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) represent subsets of object-types that are located together in space and time. Discovering MDCOPs is an important problem with many applications such as identifying tactics in battlefields, games, and predator-prey interactions. However, mining MDCOPs 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 composite interest measure for discovering MDCOPs and a novel MDCOP mining algorithm. Analytical and experimental results show that the proposed algorithm is correct and complete. Results also show the proposed method is computationally more efficient than naïve alternatives.*

## 1. Introduction

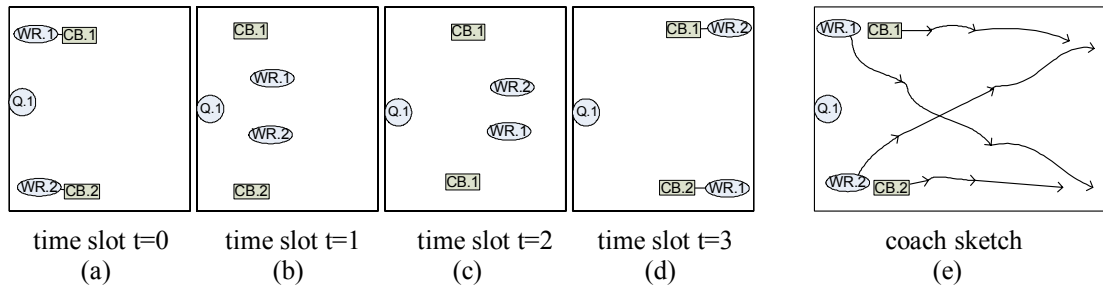
Mining MDCOPs is important for many spatio-temporal application domains, including the military (battlefield planning and strategy), ecology (tracking species and pollutant movements), homeland defense (looking for significant “events”), and transportation (road and network planning) [5, 9].

MDCOPs frequently occur during sporting events, such as in American football where two teams try to outscore each other by moving a football to the opponent’s end of the field. Various complex

interactions occur within one team and across teams to achieve this goal. These interactions involve intentional and accidental MDCOPs, the identification of which may help teams to study their opponent’s tactics. In American football object-types may be defined by the roles of the offensive and defensive players, such as quarterback, running back, wide receiver, kicker, holder, and cornerback. An MDCOP is a subset of object-types, e.g. {kicker, holder}, {wide\_receiver, cornerback}. One example MDCOP that occurs quite frequently involve wide receivers and cornerbacks of opposing teams. The objective of a wide receiver is to successfully catch the football thrown by the quarterback, whereas a cornerback attempts to prevent the catch. This interaction creates an MDCOP between these two player roles. Another pattern that may occur within the same team involves the holder and the kicker. Both are required to be together for numerous events within a single game at various locations in the field. Their objective is to kick the football to score a field goal or extra point. Other MDCOPs that frequently occur in football involve commonly used tactics such as a “double reverse,” a “blitz,” or a “fake handoff”. Identifying or mining the MDCOPs and other tactics used by opposing teams is crucial in pre-game preparation. Currently, coaches and players watch videotapes of other games to discover these patterns.

A second example of an MDCOP is in ecological predator-prey relationships. Patterns of movements of rabbits and foxes, for example, will tend to be located in the same space at the same time. The rabbit patterns will attempt to move away from the fox patterns, and the fox patterns will attempt to stay with the rabbit patterns, much like the wide receivers and cornerbacks in the American football example. In this case, other

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 (AHPARC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under contract number DAAD19-01-2-0014, and the NSF grant IIS-0208621.



**Figure 1. An example spatio-temporal dataset to compare related approaches**

**Table 1. Comparison of MDCOP with related work**

Spatio-temporal Pattern	Level		Group		Time Interval	
	Object	Object-type	Uniform	Mixed	Consecutive	Discrete
Flock Pattern [4, 11]	X		X		X	
Moving Clusters [8]	X		X	X	X	
Mixed-drove Pattern		X		X	X	X

factors such as available food and water may also affect the patterns as well.

However, discovering MDCOPs is challenging for several reasons: First, the process is computationally very expensive because the interest measures are computationally complex. Second, current interest measures (i.e. the spatial prevalence measure) are not sufficient to mine such patterns, so new composite interest measures to do so must be created and formalized [7, 13]. Third, the set of candidate patterns grows exponentially with the number of object-types. Fourth, since spatio-temporal datasets are huge, computationally efficient algorithms must be developed. We create and formalize a new monotonic composite interest measure to mine interesting and non-trivial MDCOPs out of massive spatio-temporal datasets in a computationally efficient manner.

**Related Work:** Previous studies for mining spatio-temporal co-occurrence patterns can be classified into two categories, namely, mining of uniform groups of moving objects (e.g., flock patterns [4, 11]) and mining of mixed groups of moving objects (e.g., moving clusters [8]). Our problem belongs to the latter one (Table 1). A flock pattern is a moving group of the same kind of object, such as a sheep flock or a bird flock. Gudmundsson et al. proposed algorithms for detection of the flock pattern in spatio-temporal datasets [4]. Since our problem is to mine mixed groups of objects, the proposed algorithms by Gudmundsson et al. to discover flock patterns may not be applicable to our problem. Kalnis et al. defined the problem of discovering moving clusters and proposed clustering-based methods to mine such patterns [8]. In their approach, if there is a large enough number of common objects between clusters in consecutive time

slots, such clusters are called moving clusters. Moving cluster patterns can be either uniform or a mixed group of objects [8]. However if there is no overlap between the clusters in consecutive time slots, their proposed algorithms for mining moving clusters will fail to discover MDCOPs. Table 1 shows a comparison of related work and our proposed MDCOP.

To illustrate the difference between our proposed MDCOP mining problem and related work (the flock pattern mining [4] and moving clusters mining [8]), we use the spatio-temporal dataset given in Figure 1, which gives an example of a spatio-temporal dataset of a typical play during an American football game. It shows the positions of two offensive wide receivers (WR.1 and WR.2), two defensive cornerbacks (CB.1 and CB.2), and a quarterback (Q.1) in four time slots. The objective is to have the two offensive wide receivers cross over each other and create a separation from the defensive cornerbacks to make it safer to receive a pass from the quarterback. Initially, the offensive wide receivers and the defensive cornerbacks are co-located at the time slot  $t=0$  (Figure 1(a)). In time slot  $t=1$ , the two offensive wide receivers begin their run, while the defensive cornerbacks remain in their original position, possibly due to a fake handoff from the quarterback to the running back. (Figure 1(b)). In time slot  $t=2$ , the wide receivers cross over each other and try to drift further away from their respective defensive cornerbacks (Figure 1(c)). When the quarterback shows signs of throwing the football, both defensive cornerbacks run to their respective offensive wide receivers (Figure 1(d)). The overall sketch of the game tactics can be seen in Figure 1(e).

The flock pattern algorithm [4] will not be able to find any pattern from the dataset given in Figure 1, since it is looking for uniform sets of moving objects.

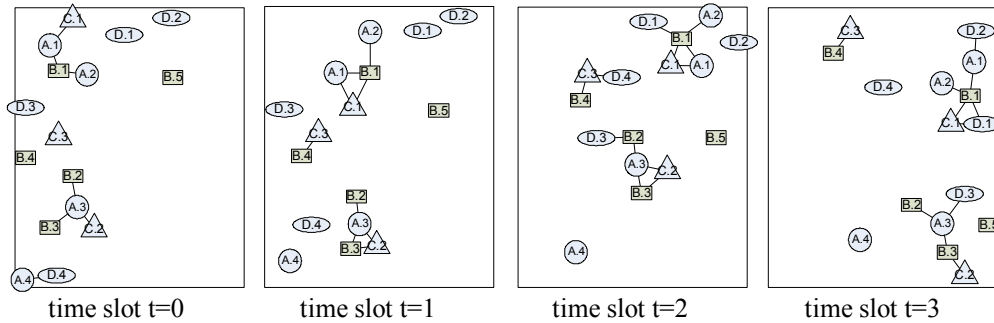


Figure 2. An input spatio-temporal dataset

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

Figure 3. A set of output mixed-drove spatio-temporal co-occurrence patterns

The moving clusters algorithms [8] will not be able to find any moving clusters in such an example because the wide receivers and cornerbacks are forming a cluster in time slots  $t=0$  and  $t=3$  but not in the intermediate time slots. Thus, there may not be any overlapping objects between clusters in consecutive time slots. In contrast, our proposed MDCOP mining approach may find MDCOP {wide\_receiver, cornerback}, if the fraction of time slots where the pattern occurs over the total number of time slots is no less than a given threshold 0.5. After all, instances of MDCOP {wide\_receiver, cornerback} are co-located in two time slots out of four. The instances of MDCOP {wide\_receiver, cornerback} are {WR.1, CB.1} and {WR.2, CB.2} in time slot  $t=0$ , and {WR.2, CB.1} and {WR.1, CB.2} in time slot  $t=3$ .

**Contributions:** This paper makes the following contributions:

- It defines mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) and the MDCOP mining problem.
- It proposes a new monotonic composite interest measure to discover and mine MDCOPs.
- It proposes a novel and computationally efficient MDCOP mining algorithm (MDCOP-Miner).
- It shows that the proposed algorithm is correct and complete in finding mixed-drove prevalent (e.g., spatial prevalent and time prevalent) MDCOPs.

- It experimentally evaluates the proposed composite interest measures and MDCOP mining algorithms using real datasets.

**Scope:** This paper focuses on the MDCOP on a typed collection of moving objects by extending interest measures for spatial co-location patterns [7, 13] given a user defined participation index threshold. The following issues are beyond the scope of this paper: (i) determining thresholds for MDCOP interest measures; (ii) similarity measures for tracking moving objects due to the focus on object-types rather than objects; (iii) indexing and query processing issues related to mining objects; (iv) discovering multisets (e.g. {A, A, B}).

**Outline:** The rest of the paper is organized as follows. Section 2 presents basic concepts to provide a formal model of MDCOPs and the problem statement of mining MDCOPs. Section 3 presents our proposed MDCOP mining algorithm. Analysis of the algorithm is given in Section 4. Section 5 presents the experimental evaluation and Section 6 discusses conclusions and future works.

## 2. Basic Concepts and Problem Statement

### 2.1 Spatial Prevalence Measure

The focus of this study is to discover mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) over a spatio-temporal framework and a neighborhood

relation R. First we will explain the modeling of mixed groups of object-types in space, e.g., spatial co-locations [13]. In the next sections, we will explain how we model modeling MDCOPs by extending the spatial co-location mining problem to include time information and then propose algorithms to mine these MDCOPs.

Spatial co-location mining algorithms are used to discover sets of mixed 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 [7, 13]. For example, in Figure 2, in time slot  $t=0$ ,  $\{A.1, C.1\}$  is an instance of a co-location if the distance between the objects is no more than a given neighborhood distance threshold. In Figure 2, the solid lines show the distance between the objects that satisfies 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 [7, 13]. Such a co-location 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 total number of instances). For example, in Figure 2,  $\{A, B\}$  is a co-location in time slot  $t=0$ , and its instances are  $\{A.1, B.1\}$ ,  $\{A.2, B.1\}$ ,  $\{A.3, B.2\}$ , and  $\{A.3, B.3\}$ . In the dataset, object-type A has 4 instances and three of them (A.1, A.2, and A.3) are contributing to the co-location  $\{A, B\}$ , so the participation ratio of A is  $3/4$ . The participation ratio of B is  $3/5$  since 3 out of 5 instances are contributing to the co-location  $\{A, B\}$ . The participation index of the co-location  $\{A, B\}$  is  $3/5$ , which is the minimum of the participation ratios of object-types A and B.

It has been shown that the participation index is anti-monotone in size of co-locations [7, 13]. In other words,  $participation\_index(P_j) \leq participation\_index(P_i)$  if  $P_i$  is a subset of  $P_j$ . In addition, [7, 13] show that the participation index has a spatial statistical interpretation as an upper bound on the cross-K function [3].

## 2.2. Modeling MDCOPs

Given a set of spatio-temporal object-types and a set of their instances with a neighborhood relationship R, an MDCOP 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 2, the total number of time slots is 4 and pattern  $\{A, B\}$  occurs in all 4 time slots, so its time prevalence is  $4/4$ . Pattern  $\{A, C\}$  occurs in 3 time slots, namely, time slots  $t=0$ ,  $t=1$ , and  $t=2$ , and its time prevalence index is  $3/4$ .

**Definition 2.2:** Given a spatio-temporal dataset  $ST$ , and a spatial prevalence threshold  $\theta_p$ , the mixed-drove prevalence measure of a spatio-temporal pattern  $P_i$  is a composition of the spatial prevalence measure and the time prevalence measure as shown below.

$$Prob_{t_m \in all\_time\_slot} (s\_prev(pattern P_i, time\_slot t_m) \geq \theta_p)$$

where *Prob* stands for probability of overall prevalence time slots and *s\_prev* stands for spatial prevalence, e.g., the participation index, described in section 2.1.

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

$$Prob_{t_m \in all\_time\_slot} [s\_prev(pattern P_i, time\_slot t_m) \geq \theta_p] \geq \theta_{time}$$

where *Prob* stands for probability of overall prevalence time slots, *s\_prev* stands for spatial prevalence,  $\theta_p$  is the spatial prevalence threshold, and  $\theta_{time}$  is the time prevalence threshold.

For example, in Figure 2,  $\{A, B\}$  is an MDCOP because it is spatial prevalent in time slots  $t=0$ ,  $t=1$ ,  $t=2$ , and  $t=3$  since its participation indices are no less than the given threshold 0.4 in these time slots, and it is time prevalent since its time prevalence index, i.e., 1, is above the time prevalence index threshold 0.5. In contrast, pattern  $\{B, D\}$  is not an MDCOP. Although it is spatial prevalent in time slot  $t=2$ , it is not time prevalent since its time prevalence index is no more than the given time prevalence index threshold 0.5.

## 2.3. Problem statement

**Given:**

- A set  $P$  of Boolean spatio-temporal object-types over a common spatio-temporal framework STF.
- A neighbor relation R over locations.
- A spatial prevalence threshold,  $\theta_p$ .
- A time prevalence threshold,  $\theta_{time}$ .

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

**Objective:** Minimize computation cost.

**Constraints:** To find a correct and complete set of MDCOPs.

**Example:** In American Football, each play (e.g., Figure 1) may represent a spatio-temporal dataset and Boolean object-types may be identified by the role of the players (e.g., wide receiver and cornerback). Each object-types are considered as Boolean because of we are interested in presence and absence at any location and time. Figure 1(a)-(d) shows the position of the Boolean object-types for four time units. The straight lines between the players show the neighboring ones. The neighbor relation  $R$  may be defined by a distance less than one meter or an average arm's length. For example, in Figure 1(a), wide receiver WR.1 is a neighbor of cornerback CB.1. However, these players are not neighbors in Figure 1(b) since they are separated by more than a meter. In this example,  $\{\text{wide\_receiver, cornerback}\}$  forms a candidate MDCOP, given  $\theta_p=0.5$ , and  $\theta_{time}=0.5$ .

### 3. Mining MDCOPs

In this section, we first discuss a **naïve approach** and then propose a novel MDCOP mining algorithm (MDCOP-Miner) to mine MDCOPs.

**Naïve approach:** A naïve approach can use a spatial co-location mining algorithm for each time slot to find spatial prevalent co-locations and then can apply a post-processing step to discover mixed-drove prevalent MDCOPs by checking their time prevalence. To mine co-locations, Huang, Shekhar and Xiong proposed a join-based approach, Yoo, Shekhar and Celik proposed a partial join-based approach and a join-less approach, and Zhang et al. proposed a multi-way spatial join-based approach [2, 7, 13, 15-18]. This study will be based on the join-based spatial co-location pattern mining algorithm proposed by Huang et al., but it is also possible to use other approaches. The naive approach will generate size  $k+1$  candidate co-locations for each time slot using spatial prevalent size  $k$  subclasses until there are no more candidate spatial co-locations. After finding all size spatial prevalent co-locations in each time slot, a post-processing step can be used to discover mixed-drove prevalent MDCOPs by pruning out time non-prevalent co-locations. Even though this approach will prune out spatial non-prevalent co-locations early, it will not prune out time non-prevalent MDCOPs before the post-processing step. This leads to unnecessary computational cost.

**MDCOP-Miner:** In contrast, we propose an MDCOP mining algorithm (MDCOP-Miner) to discover mixed-drove prevalent MDCOPs by incorporating a time-prevalence based filtering step in each iteration of the algorithm. It will generate size  $k+1$  candidate MDCOPs using size  $k$  mixed-drove prevalent MDCOPs. The participation index is used as a spatial prevalence interest measure to check if the pattern is spatial prevalent at a time slot [7]. The time prevalence (i.e., persistence measure in definition 2.1) is used as a time prevalence interest measure. First we give the pseudo code of the algorithm, and then we provide an execution trace of it using the spatio-temporal dataset from Figure 2.

<b>Pseudo code for MDCOP-Miner Algorithm</b>	
<b>Inputs:</b>	E: a set of spatial object-types ST: a spatio-temporal dataset <object_type, object_id, x, y, time> R: spatial neighborhood relationship TF: a time slot frame $\{t_0, \dots, t_{n-1}\}$ $\theta_p$ : a spatial prevalence threshold $\theta_{time}$ : a time prevalence threshold
<b>Output:</b>	MDCOPs whose spatial prevalence indices, i.e., participation indices, are no less than $\theta_p$ , for time prevalence indices are no less than $\theta_{time}$ .
<b>Variables:</b>	k: co-occurrence size t: time slots $(0, \dots, n-1)$ $T_1$ : 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 $MDP_k$ : set of mixed-drove size k co-occurrences
<b>Algorithm:</b>	1. initialization 2. co-occurrence size $k=1$ , $C_k(0)=E$ , $MDP_1(0)=ST$ 3. while ( not empty $MDP_k$ ) { 4.   For each time slot $t$ in $(0, \dots, n-1)$ { 5. $C_{k+1}(t)=\text{gen\_candidate\_co-occ}(C_k(t), MDP_k(t))$ 6. $T_{k+1}(t)=\text{gen\_co-occur\_instance}(C_{k+1}(t), T_k(t), R)$ 7. $SP_{k+1}(t)=\text{find\_spatial-prevalent\_co-occ}(T_{k+1}(t), C_{k+1}(t), \theta_p)$ 8.     } 9. $TP_{k+1}=\text{find\_time\_prevalence\_index}(SP_{k+1})$ 10. $MDP_{k+1}=\text{find\_time-prevalent\_co-occur}(TP_{k+1}, \theta_{time})$ 11. $k=k+1$ 12.    } 13. return union ( $MDP_2, \dots, MDP_{k+1}$ )

**Algorithm 1. MDCOP-Miner algorithm**

Algorithm 1 gives the pseudo code of the MDCOP-Miner algorithm. The inputs are a set of spatial object-types  $E$ , a spatio-temporal dataset  $ST$ , a spatial neighborhood relationship  $R$ , and thresholds of interest measures, i.e. spatial prevalence and time prevalence and the output is a set of mixed-drove prevalent MDCOPs.

Step 1: Generate pairs and find participation indices

Co-occurrence patterns	time slot t=0							time slot t=1							time slot t=2							time slot t=3											
	AB	AC	AD	BC	BD	CD		AB	AC	AD	BC	BD	CD		AB	AC	AD	BC	BD	CD		AB	AC	AD	BC	BD	CD						
Co-occurrence	A.1 B.1	A.1 C.1	A.4 D.4					A.1 B.1	A.1 C.1	B.1 C.1				A.1 B.1	A.1 C.1	B.1 C.1	B.1 D.1	C.3 D.4			A.1 B.1	A.1 D.2	B.1 C.1	B.1 D.1	C.1 D.1								
Co-occurrence	A.2 B.1	A.3 C.2						A.2 B.1	A.3 C.2	B.3 C.2				A.2 B.1	A.3 C.2	B.3 C.2	B.2 D.3				A.2 B.1	A.3 D.3	B.3 C.2										
Co-occurrence	A.3 B.2							A.3 B.2		B.4 C.3				A.3 B.2		B.4 C.3				A.3 B.2		B.4 C.3											
Co-occurrence	A.3 B.3							A.3 B.3						A.3 B.3						A.3 B.3													
P. ratio	3/4	3/5	2/4	2/3	1/4	1/4		3/4	3/5	2/4	2/3	3/5	3/3	3/4	3/5	2/4	2/3	3/5	3/3	2/5	2/4	1/3	1/4	3/4	3/5	2/4	2/3	3/5	3/3	1/5	1/4	1/3	1/4
P. index	3/5		2/4		1/4			3/5		2/4		3/5		3/5		2/4		3/5		2/5		1/4		3/5		2/4		3/5		1/5		1/4	
If PI threshold is 0.4	• {A,D} is pruned														• {C,D} is pruned							• {B,D} and {C,D} are pruned											

(a)

Step 2: Form time prevalence table

	time slot t=0	time slot t=1	time slot t=2	time slot t=3	time prevalence index
A B	1	1	1	1	4/4
A C	1	1	1	0	3/4
A D	0	0	0	1	1/4
B C	0	1	1	1	3/4
B D	0	0	1	0	1/4
C D	0	0	0	0	0

- If time prevalence index threshold 0.5 (50%) then prune {A,D} and {B,D}
- {A,B}, {A,C}, {B,C} are mixed-drove co-occurrence patterns

Step 3: Generate superset patterns (triplets)

	time slot t=0	time slot t=1	time slot t=2	time slot t=3		
A B C	A B C	A B C	A B C	A B C		
		A.1 B.1 C.1	A.1 B.1 C.1			
		A.3 B.3 C.2	A.3 B.3 C.2			
PR	2/4	2/5	2/3	2/4	2/5	2/3
PI		2/5		2/5		

Step 4: Find mixed-drove co-occurrence patterns

	time slot t=0	time slot t=1	time slot t=2	time slot t=3	time prevalence index
A B C	-	1	1	-	2/4

- {A, B, C} is mixed-drove co-occurrence pattern

(b)

Figure 4. Execution trace of the MDCOP-Miner algorithm

In the algorithm, steps 1 and 2 include initialization of the parameters, steps 3 through 12 give an iterative process to mine MDCOPs, and step 13 gives a union of the results of the iterative steps. Steps 3 through 12 continue until there is no candidate MDCOPs 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 mixed-drove prevalent size  $k$  mixed-drove co-occurrence patterns  $MDP_k$  [1].

**Generating spatial co-occurrence instances (step 6):** The instances of candidate  $C_{k+1}$  are generated by joining neighbor instances of mixed-drove prevalent size  $k$  MDCOPs for each time slot. This is similar to the instance generation step of the co-location miner algorithm [7].

**Finding spatial prevalent co-occurrence 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 the participation indices follows the same algorithmic ideas as those in the co-location mining algorithm [7].

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

**Forming a time prevalence table (step 9):** In step 9, the time prevalence indices of the mined spatial prevalent patterns are calculated. The time prevalence index of a spatial prevalent co-location is the fraction of the number of time slots where the pattern occurs over the total number of the time slots.

**Finding mixed-drove co-occurrence patterns (step 10):** This step discovers MDCOPs by checking the time prevalence indices of the spatial prevalent co-locations if they are no less than a given time prevalence threshold  $\theta_{ime}$ . The patterns whose time prevalence indices do not satisfy the given threshold are pruned at this stage. The remaining patterns will be mixed-drove prevalent MDCOPs and will be used to generate candidate supersets of the MDCOPs in step 5.

The algorithm will run iteratively until there are no more candidate MDCOPs to be generated. The algorithm outputs the union of all size mixed-drove prevalent MDCOPs.

**An Execution Trace:** The execution trace of the algorithm is given in Figure 4 using the spatio-temporal dataset given in Figure 2. This dataset contains four object-types A, B, C, and D and their instances in four time slots. A has 4 instances, B has 5 instances, C has 3 instances, and D has 4 instances.

The instances of each object-type have a unique identifier, such as A.1. Some of the patterns of these object-types form an MDCOP. To discover MDCOPs we propose a monotonic composite interest measure (the mixed-drove 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 when the index is greater than or equal to a given threshold [7, 13]. The time prevalence measure (time prevalence index) shows the frequency of the pattern over time.

In Figure 4(a), in step 1, candidate spatial co-location pairs of the object-types and their instances are generated for each time slot. The spatial co-locations whose participation indices are less than a given threshold are then pruned. A spatial non-prevalent co-location {A, D} is pruned in time slot  $t=0$ , {C, D} is pruned in time slots  $t=2$  and  $t=3$ , and {B, D} is pruned in time slots  $t=3$  because their participation indices are less than the given threshold 0.4.

A time prevalence table of pairs of spatial prevalent co-locations is then formed by entering a 1 if the participation index of the corresponding pattern satisfies a given participation index threshold. Time-prevalence indices are then found. For example, in the time prevalence table (step 2 in Figure 4(b)), spatial prevalent pattern {A, B} is persistent for all time slots and its time prevalence index is 4/4, and spatial prevalent pattern {A, C} is persistent in time slots  $t=0$ ,  $t=1$ , and  $t=2$  and its time prevalence index is 3/4, etc. The MDCOPs whose time prevalence indices are no less than a given threshold are selected for generating superset candidate MDCOPs. Spatial prevalent patterns {A, B}, {A, C}, and {B, C} are selected as mixed-drove prevalent MDCOPs since they are also time prevalent (their time prevalence indices satisfy the given time prevalence index threshold 0.5). In contrast, spatial prevalent patterns {A, D}, {B, D}, and {C, D} are pruned since they are time non-prevalent. Using MDCOPs {A,B}, {A, C}, and {B, C}, the next candidate MDCOP {A, B, C} is generated.

The next step is to generate instances of candidate MDCOP {A, B, C} in time slots where its subsets exist and to check its participation indices in these time slots. Since all subsets of MDCOP {A, B, C} are mixed-drove prevalent MDCOPs and exist in time slots  $t=1$  and  $t=2$ , there is no need to generate instances of them for time slots  $t=0$  and  $t=3$ . In step 3 (Figure 4(b)), the instances of candidate MDCOP {A,B,C} are generated and participation indices are found which are 2/5 for time slots  $t=1$  and  $t=2$ .

In step 4 (Figure 4(b)), the time prevalence table is formed for pattern {A, B, C} and its time prevalence index is checked to see if it satisfies the time

prevalence threshold. Candidate MDCOP {A, B, C} is an MDCOP since its time prevalence index 0.5 is equal to the time prevalence threshold 0.5. Since there are not enough subsets to generate the next superset patterns, the algorithm stops at this stage and outputs the MDCOPs union of all size mixed-drove prevalent MDCOPs, i.e., {A, B}, {A, C}, {B, C}, and {A, B, C}.

## 4. Analysis of the MDCOP-Miner

This section gives the analysis of the mixed-drove prevalence index measure, and correctness and completeness derivations for the MDCOP-Miner.

### 4.1. The Mixed-Drove Prevalence Index Measure is Monotonic

**Lemma 4.1:** *A chosen spatial prevalence measure, e.g., participation index, is monotonically non-increasing in the size of the MDCOPs at each time slot [7, 13].*

**Lemma 4.2:** *A mixed-drove prevalence index measure is monotonically non-increasing with the size of MDCOP over space and time. In other words, if MDCOP  $P_i$  is a subset of MDCOP  $P_j$  and*

$$\text{Prob}_{t_m \in \text{all\_time\_slot}} (s\_prev(\text{pattern } P_i, t_m) \geq \theta_p), \text{ and}$$

$$\text{Prob}_{t_m \in \text{all\_time\_slot}} (s\_prev(\text{pattern } P_j, t_m) \geq \theta_p)$$

where *Prob* stands for probability of overall prevalence time units, *s\_prev* stands for spatial prevalence,  $\theta_p$  is spatial prevalence threshold, and  $t_m$  is time slot.

**Proof:** The basic proof sketch follows. Let

$$TS(P_j, \theta_p) = \{t_m \mid \text{participation\_index}(P_j, t_m) \geq \theta_p\}$$

Lemma 4.1 implies that  $\text{participation\_index}(P_j, t) \geq \theta_p$  for all  $t_m \in TS(P_j, \theta_p)$ , since  $P_i$  is a subset of  $P_j$ . Thus,

$$\text{Prob}_{t_m \in \text{all\_time\_slot}} [s\_prev(\text{pattern } P_i, t_m) \geq \theta_p] \geq \theta_{time}$$

where  $\theta_{time}$  is time prevalence threshold.

### 4.2. Correctness and Completeness

**Theorem 4.1:** *The MDCOP-Miner is complete.*

**Proof:** The MDCOP-Miner is complete if it finds all prevalent mixed-drove prevalent MDCOPs that satisfy a given participation index 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 MDCOP.

The *gen\_candidate\_co-occur* function does not miss any patterns given the anti-monotone nature of the MDCOP interest measure. The input of this function is mixed-drove prevalent size  $k$  MDCOPs and the output is candidate size  $k+1$  MDCOPs. If  $c_1=\{f_1, \dots, f_k\}$  and  $c_2=\{f_1, \dots, f_{k-1}, f_{k+1}\}$  are size  $k$  mixed-drove prevalent MDCOPs, candidate size  $k+1$  pattern  $C_{k+1}=\{f_1, \dots, f_{k-1}, f_k, f_{k+1}\}$  will be produced by joining mixed-drove prevalent size  $k$  MDCOPs.

The *gen\_co-occur\_instance* function does not miss any patterns. This function generates instances of candidate size  $k+1$  MDCOPs by joining instances of mixed-drove prevalent size  $k$  MDCOPs if they are in the neighborhood distance and forming a clique.

The *find\_spatial-prevalent\_co-oc* function does not miss any patterns. It finds spatial prevalent patterns whose participation indices satisfy a given 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\_time-prevalent\_co-occur* function does not miss any MDCOP. The function finds all the mixed-drove prevalent MDCOPs whose time prevalence indices are no less than a given threshold.

**Theorem 4.2:** *The MDCOP-Miner is correct. In other words, if a MDCOP pattern  $P$  is returned by MDCOP-Miner algorithm then  $P$  is a prevalent MDCOP.*

**Proof:** The proof is easy to establish due to the pruning steps of “*find\_spatial\_prevalent\_co-occur*”, and *find\_time\_prevalent\_co-occur*” which weed out candidates not meeting the given thresholds.

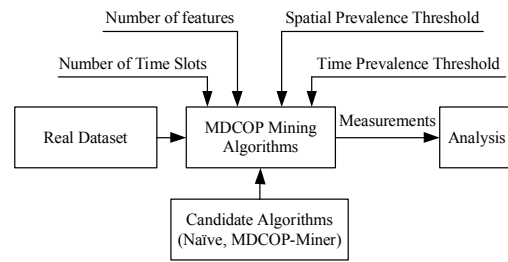
**Theorem 4.3:** *The total cost of MDCOP-Miner algorithm is no more than the total cost of naïve approach.*

**Proof:** The basic proof sketch follows. MDCOP-Miner prunes out spatial prevalent but time non-prevalent patterns early and generates no more candidates than the Naïve approach. The cost of time pruning is negligible relative to the spatial pruning.

## 5. Experimental Evaluation

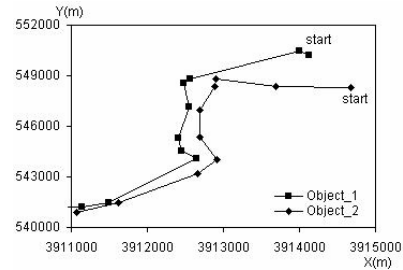
In this section, we present our experimental evaluations of several design decisions and workload parameters of our MDCOP-Miner algorithm. We used a real-world training dataset. We evaluated the behavior of the MDCOP-Miner and naïve approach by changing the number of time slots, number of object-types, and the spatial prevalence and time prevalence index thresholds. Figure 5 shows the experimental setup to evaluate the impact of design decisions of the

performance of both algorithms. Experiments were conducted on an Intel Centrino PIV 1.6 GHz computer with 512 MB of RAM.



**Figure 5. Experimental setup and design**

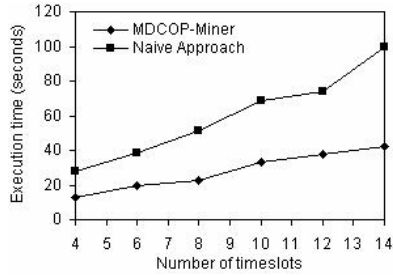
The dataset contains the location and time information of moving objects. It includes 15 time snapshots and 22 distinct vehicle types and their instances. The minimum instance number is 2, the maximum instance number is 78, and the average number of instances is 19. Figure 6 shows an instance of an MDCOP. Object 1 and object 2 are coming together, moving from top right to bottom left. Such a pattern may be of interest if it indicates an imminent offensive maneuver by object\_1 under cover from object\_2.



**Figure 6. One instance of an MDCOP**

### 5.1 Effect of Number of Time slots

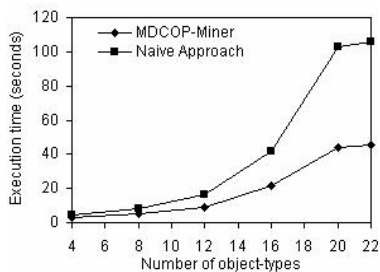
In the first experiment, we evaluated the effect of number of time slots on the execution time of both algorithms. The participation index, time prevalence index, and distance were set at 0.15, 0.5, and 100m respectively. The MDCOP-Miner requires less execution time than the naïve approach, since it prunes out mixed-drove non-prevalent MDCOPs early (Figure 7). It can also be seen that, as the number of time slots increases, the ratio of the increase in execution time is smaller for MDCOP-Miner than the naïve approach.



**Figure 7. Effect of number of time slots**

## 5.2 Effect of Number of Object-types

In the second experiment, we evaluated the effect of number of object-types on the execution time of both algorithms. The participation index, time prevalence index, number of time slots and distance were set at 0.15, 0.5, 15, and 100m respectively. The MDCOP-Miner outperforms the naive approach as the number of object-types increases (Figure 8). It is observed that the increase ratio of the execution time of the naive approach is bigger than the MDCOP-Miner as the number of object-types increases. It should also be noted that the distributions of the object-types affect the computation cost of both algorithms. The cost of both algorithms increases dramatically between 16 and 20 since the newly added 4 object-types are highly likely to have neighbor relations with nearby object-types. In contrast, between 20 and 22, the cost does not increase too much since the newly added object-types are less likely to have neighbor relations with nearby object-types.

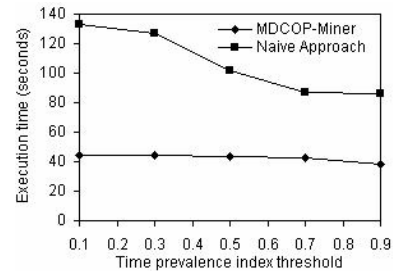


**Figure 8. Effect of number of object-types**

## 5.3 Effect of the Time Prevalence Index Threshold

In the third experiment, we evaluated the effect of the time prevalence index threshold on the execution times of both algorithms. The fixed parameters were participation index, number of time slots, and distance, and their values were 0.15, 15, and 100m respectively. For the naive approach, the effective cost in execution time to generate spatial prevalent co-locations will be

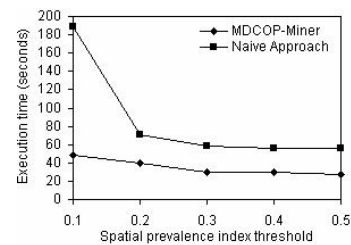
constant since it generates the same number of spatial prevalent patterns as the time prevalence index increases. In that case, the cost of the post-processing step will reflect the trend of the naive approach. Experimental results show that the MDCOP-Miner is more computationally efficient than the naive approach because of the early pruning strategy (Figure 9). It is also observed that the naive approach is computationally more expensive as the time prevalence index threshold decreases because of the increase in the number of MDCOPs to be discovered.



**Figure 9. Effect of the time prevalence index threshold**

## 5.4 Effect of the Spatial Prevalence Index Threshold

In the fourth experiment, we evaluated the effect of the spatial prevalence index threshold on the execution times of both algorithms. The fixed parameters are time prevalence index, number of time slots, and distance, with values of 0.5, 15, and 100m respectively. The MDCOP-Miner outperforms the naive approach as the spatial prevalence index threshold increases (Figure 10). The cost of the naive approach will be higher than the MDCOP-Miner for low values of the spatial prevalence index threshold.



**Figure 10. Effect of spatial prevalence index threshold**

It is also observed in all the experiments that the memory usage of the MDCOP-Miner is no more than for the naive approach (Theorem 4.3).

## 6. Conclusions and Future Work

We defined mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) and the MDCOP mining problem and proposed a new monotonic composite interest measure (the mixed-drove prevalence measure), which is the composition of the spatial prevalence and time prevalence measures. We also presented a novel and computationally efficient algorithm (the MDCOP-Miner) for mining these patterns. We compared our algorithm with a naive approach, which runs the spatial co-location mining algorithm at each time slot and then discovers MDCOPs using a post-processing step. We proved that the proposed algorithms are correct and complete in finding mixed-drove prevalent (e.g., spatial-prevalent and time prevalent) MDCOPs. Our experimental results using a real dataset provide further evidence of the viability of our approach.

In future work, we would like to evaluate effect of objects (instances) of object-types on the proposed algorithms and to explore the relationship between the proposed MDCOP interest measures and the spatio-temporal statistical measures of interaction [2]. Another problem of interest is the characterization of the probability distribution of the proposed interest measure to help in making the choice of thresholds in the proposed measures. We plan to explore other potential interest measures for MDCOPs by evaluating similarity measures for tracks of moving objects. We plan to investigate new monotonic composite interest measures and develop new computationally efficient algorithms for mining MDCOPs.

In the literature, there also other studies focused on defining spatio-temporal patterns and algorithm [4, 6, 8, 10, 12, 14]. Laube and Imfeld defined several spatio-temporal patterns, such as, leadership, convergence [11]. Query processing algorithms have been proposed to extract such patterns [11]. We plan to extend our algorithm to mine these patterns.

## 7. Acknowledgments

The authors would like to thank James Kang and Kim Koffolt for their comments.

## 8. References

- [1] R. Agarwal and R. Srikant, Fast algorithms for Mining Association Rules, *VLDB'94*, 1994.
- [2] S. Banerjee, B. P. Carlin, and A. E. Gelfrand, *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press, ISBN 158488410X, 2003.
- [3] N. A. C. Cressie, *Statistics for Spatial Data*, Wiley and Sons, ISBN 0471843369, 1991.
- [4] J. Gudmundsson, M. v. Kreveld, and B. Speckmann, Efficient Detection of Motion Patterns in Spatio-Temporal Data Sets, *ACM-GIS*, 250-257, 2004.
- [5] R. Guting and M. Schneider, *Moving Object Databases*, Morgan Kaufmans, 2005.
- [6] M. Hadjieleftheriou, G. Kollios, P. Bakalov, and V. J. Tsotras, Complex Spatio-Temporal Pattern Queries, *VLDB'05*, 877-888, 2005.
- [7] 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.
- [8] P. Kalnis, N. Mamoulis, and S. Bakiras, On Discovering Moving Clusters in Spatio-temporal Data, *9th Int'l Symp. on Spatial and Temporal Databases (SSTD)*, Angra dos Reis, Brazil, 2005.
- [9] 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.
- [10] P. Laube and S. Imfeld, Analyzing relative motion within groups of trackable moving point objects, in *In GIScience, number 2478 in Lecture notes in Computer Science*. Berlin: Springer, pp. 132-144, 2002.
- [11] P. Laube, M. v. Kreveld, and S. Imfeld, Finding REMO - detecting relative motion patterns in geospatial lifelines, *11th Int'l Symp. on Spatial Data Handling*, 201-214, 2004.
- [12] C. d. Mouza and P. Rigaux, Mobility Patterns, *GeoInformatica*, vol. 9(4), pp. 297-319, 2005.
- [13] S. Shekhar, Y. Huang, 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.
- [14] 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.
- [15] J. S. Yoo and S. Shekhar, A Joinless Approach for Mining Spatial Colocation Patterns, *IEEE Trans. on Knowledge and Data Eng. (TKDE)*, vol. 18(10), pp., 2006.
- [16] J. S. Yoo and S. Shekhar, A Partial Join Approach for Mining Co-location Patterns, *ACM-GIS'05*, Washington D.C., USA, 2005.
- [17] J. S. Yoo, S. Shekhar, and M. Celik, A Join-less Approach for Co-location Pattern Mining: A Summary of Results, *IEEE Int'l Conf. on Data Mining*, Houston, USA, 2005.
- [18] X. Zhang, N. Mamoulis, D. W. L. Cheung, and Y. Shou, Fast Mining of Spatial Collocations, *10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 384-393, Seattle, WA, 2004.