Mining SpatioTemporally Invariant Patterns

Luca Colomba
Politecnico di Torino
Torino, Italy
luca.colomba@polito.it

Luca Cagliero
Politecnico di Torino
Torino, Italy
luca.cagliero@polito.it

Paolo Garza
Politecnico di Torino
Torino, Italy
paolo.garza@polito.it

ABSTRACT

Discovering patterns that represent key spatial or temporal dependencies among data is a well-known exploratory data mining task. However, prior works either separately analyze spatial and temporal dependencies or discover joint spatiotemporal properties of specific trajectories observed over a region of interest. With the goal of generalizing the information provided by spatiotemporal patterns, in this paper we extract sequences of discrete events showing spatiotemporally invariant properties. We seek patterns whose corresponding instances in the source data differ only due to an invariant spatiotemporal transformation. We denote such a new type of patterns as SpatioTemporally Invariant. We also propose an efficient algorithm to mine STInvs and validate its efficiency and effectiveness on real data.

CCS CONCEPTS

• **Information systems** → *Spatial-temporal systems*; *Data mining*.

KEYWORDS

Spatiotemporal Data, Pattern Mining, Data Mining

ACM Reference Format:
Luca Colomba, Luca Cagliero, and Paolo Garza. 2022. Mining SpatioTemporally Invariant Patterns. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557915.3560998

1 INTRODUCTION

Desigining and applying data mining algorithms tailored to spatiotemporal data has become relevant to several application domains (e.g., [4, 7]). Sequential pattern mining is an unsupervised technique aimed at discovering recurrent patterns from sequential data, which provide actionable insights. Traditional sequential pattern mining approaches (e.g., [3]) discover sequences of discrete events. They allow end-users to enforce domain-specific temporal constraints [1, 10] while neglecting the spatial dimension. Conversely, arbitrary spatial relationships can be modelled as spatial trajectories [2] or co-location patterns [9] while ignoring the temporal aspects.

More recently, the research community has addressed the joint analysis of spatiotemporal correlations among data. Prior works focus on (1) Characterizing specific routes by means of trajectory and STAR patterns [2, 8], or (2) Finding frequent co-located and co-occurring contiguous events in a sequence by extracting propagation/influential patterns [6]. However, the discovered patterns are tailored to specific locations or regions and not easy to abstract from specific data instances.

This paper aims at generalizing geospatial event relations by introducing the notion of *spatiotemporal invariance*. Given a trigger event of interest, we look for associated sequences of events whose instances frequently show similar spatial and temporal relative gaps. For instance, we seek sequences of multiple events taking place at relatively similar distances both in space and time. To address this issue, we propose a new type of patterns, namely the SpatioTemporally Invariant Patterns (STInvs). They describe a trigger event followed by a sequence of spatiotemporally related ones. The related events are characterized by a (constrained) spatial and temporal gap with respect to the trigger event. Thus, STInvs provide a high-level description of the spatiotemporally invariant relations holding between a trigger event and those occurring in its neighborhood regardless of the specific data instance.

The proposed pattern differs from existing ones because events are not necessarily located in close spatial proximity (unlike the events described by co-location patterns [9]), are not constrained to specific trajectories (unlike trajectory patterns [2, 8]), and can represent non-contiguous sequences of co-occurring and co-located events (unlike propagation patterns [6]).

The contributions of this paper are as follows.

- **Formalization**. Introduction of the notion of spatiotemporal invariance.
- **Pattern**. Definition of a new pattern (namely the STInv patterns).
- **Algorithm**. Design and development of a new mining algorithm (namely STInvMiner).
- **Experiments**. Empirical evaluation on a real case study related to bicycle sharing system management.

2 THE SPATIOTEMPORALLY INVARIANT PATTERN

We describe the occurrences of a set of discrete events of interest according to their spatiotemporal properties. For example, the manager of a bike sharing system can be interested in recording and analyzing critical occupancy levels of the stations in terms of percentage of available docks.

Domain experts are asked to (i) define a set of possible event types E and (2) mark a subset of them, hereafter denoted by *trigger events*, as relevant to effectively support decision-making. For example, when the occupancy level of a specific station is too high a trigger event can be recorded.

SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3560998

Table 1: Summary of the notation used.

Symbol	Description
A	the geographical area under analysis (e.g., the urban area where the bike sharing system is active)
l	a geographical location within A (e.g., the location of a docking station)
r	spatial resolution, i.e., the radius of the circular neighborhood of a geographical location l
td	temporal resolution, i.e., duration of the considered time slots
$length$	max sequence length, i.e., the maximum number of time slots that appear in an input sequence
δs	spatial gap (multiple of r)
δt	temporal gap (multiple of td)
$\mathbf{E}$	set of event types under analysis (e.g., car accident, very low level of occupancy of a bike sharing station)
$\mathbf{E}^{tr}$	trigger event types ($\mathbf{E}^{tr} \subseteq \mathbf{E}$)
$\langle e,t,l \rangle$	occurrence of event type e at timestamp t in location l

When an arbitrary event is observed, we keep track of the corresponding location l and time t (see Table 1 for a detailed description of the notation used). The location indicates the geographical position where the event of type e took place. Similar to co-location patterns [9], locations are also characterized by a local neighborhood, which indicates the potential area of influence of an event occurrence. In our context, the local neighborhood of a location l is defined as a circular area of radius r centered in l. To enable a multi-resolution spatial analysis, we define various amplitudes of neighborhood represented by concentric circles of radius $k \cdot r$ centered in l ($k \in \mathbb{Z}^+$). Hereafter, r will be also denoted as *spatial resolution*.

The timestamp of occurrence of a discrete event t is discretized into discrete time slots of equal duration td. td indicates the temporal resolution at which we consider the occurrences of discrete events.

Example. In Figure 1 the temporal resolution is 15 minutes. All the events occurring in the same time slot (e.g., e_1 and e_2) are assumed to be concurrent.

A spatiotemporally annotated occurrence of event e at timestamp t in location l is denoted as a triplet $\langle e,t,l \rangle$. For the sake of brevity, we will also denote the event occurrence $\langle e_i, t_j, l_z \rangle$ where $e_i \in E$, $t_j \in T$, and $l_z \in A$, as $occ_{e_i t_j l_z}$ whenever clear from the context.

2.1 The event sequence mining task

We explore temporal sequences of discrete events [3].

Definition 2.1. An *event sequence* of length n (also denoted as event n-sequence) is an ordered list of items, each one consisting of an event occurrence: $[occ_{e_1 t_1 l_1}, occ_{e_2 t_2 l_2}, \ldots, occ_{e_n t_n l_n}]$, where $t_1 \leq t_2 \ldots \leq t_n$.□

An arbitrary sequence may include events of different types and contemporary events happening in different locations.

For our purposes, all the timestamps associated with the event occurrences are discretized into time slots (e.g., [10am, 10.15am]) with fixed duration td. We denote td as the *temporal resolution.*

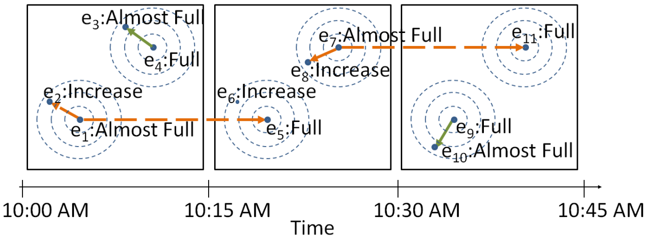

Figure 1: Examples of SpatioTemporally Invariant patterns.

We apply a time windowing approach [5] to collect a database consisting of a set of event sequences. .

Example. Let us consider the time slot TS_1 including timestamps t_1 and t_2, time slot TS_2 including timestamps t_3, and time slot TS_3 including t_4. Let $occ_{e_1 t_1 l_1}$, $occ_{e_2 t_1 l_2}$, $occ_{e_2 t_2 l_1}$, $occ_{e_4 t_2 l_3}$, $occ_{e_1 t_3 l_1}$, and $occ_{e_5 t_4 l_4}$ be the observed events.

Let us define a window of size $length \cdot td$, being $length = 2$, sliding over the time span and including pairs of consecutive time slots. The corresponding sequence database contains the following sequences $S_1, \ldots, S_3$ respectively corresponding to $TS_1, \ldots, TS_3$:

S_1: $[occ_{e_1 t_1 l_1}, occ_{e_2 t_1 l_2}, occ_{e_2 t_2 l_1}, occ_{e_4 t_2 l_3}, occ_{e_1 t_3 l_1}]$
S_2: $[occ_{e_1 t_3 l_1}, occ_{e_5 t_4 l_4}]$
S_3: $[occ_{e_5 t_4 l_4}]$

The sequence database D=$\cup_{i=1}^{3} S_i$ collects all the sequences generated by the sliding process.

2.2 The spatiotemporal invariance property

We look for the sequences of discrete events showing invariant spatiotemporal properties. To this end, we first introduce the concepts of time and spatial gaps, which indicate the sequential variation of the recorded timestamps and locations, respectively.

Definition 2.2. Let S=$[occ_{e_1 t_1 l_1}, occ_{e_2 t_2 l_2}, \ldots, occ_{e_n t_n l_n}]$ be an event n-sequence. The *temporal gap sequence* relative to S, $\delta t(S)$ in short, is the (n-1)-sequence $[t_2 - t_1, t_3 - t_1, \ldots, t_n - t_1]$. □

The temporal gaps $t_j - t_1$ are multiples of td.

Definition 2.3. Let S=$[occ_{e_1 t_1 l_1}, occ_{e_2 t_2 l_2}, \ldots, occ_{e_n t_n l_n}]$ be an event n-sequence. The *spatial gap sequence* relative to S, $\delta s(S)$ in short, is the (n-1)-sequence [dist(l_2,l_1), dist(l_3,l_1), ..., dist(l_n,l_1)], where $dist(l_x, l_y) = \left\lceil \frac{haversine\_distance(l_x,l_y)}{r} \right\rceil$.

The spatial distance among two event occurrences depends on r.

Example. In the left-hand-side picture of the example in Figure 1, the (discretized) spatial gap between events e_1 and e_2 is equal to $3 \cdot r$ because event e_2 lies in the third circle centered in e_1.

Two n-sequences are spatially/temporally invariant if their corresponding spatial/temporal gap sequences are coincident.

Definition 2.4. Let S_1 and S_2 be two event n-sequences and let $\delta t(S_1)$ and $\delta t(S_2)$ be the corresponding temporal gap sequences. S_1 and S_2 are *temporally invariant* if and only if $\delta t(S_1)$=$\delta t(S_2)$. □

Definition 2.5. Let S_1 and S_2 be two event n-sequences and let $\delta s(S_1)$ and $\delta s(S_2)$ be the corresponding spatial gap sequences. S_1 and S_2 are *spatially invariant* if and only if $\delta s(S_2)$=$\delta s(S_2)$. □

Example. Let us consider the following sequences: S_1: *a car accident occurred at 9pm in location l_1*, then *a car accident occurred at 9:30pm in l_1*. S_2: *a car accident occurred at 7.00am in location l_2*, then *a car accident occurred at 7:30am in l_2*. S_1 and S_2 are spatiotemporal invariant sequences because the corresponding pairs of events occurred within the same position (spatial gap=0) with a temporal gap of 30 minutes.

2.3 The SpatioTemporally Invariant pattern

The main purpose of STInv is to characterize relevant spatiotemporally invariant trends in sequence databases. An STInv summarizes a set of event sequences $S_1, S_2, \ldots, S_q$ such as all the pairs of summarized sequences (S_i, S_j) $1 \leq i, j \leq q$ are spatiotemporally invariant.

Definition 2.6. Let $S_1, S_2, \ldots, S_q$ be a set of q spatiotemporally invariant sequences sharing the same trigger event as first event type ($e_1 \in \mathbf{E}^{tr}$). The STInv is a sequence of triplets $[\langle\ e_1, \delta s_1, \delta t_1\ \rangle, \langle\ e_2, \delta s_2, \delta t_2\ \rangle, \ldots, \langle\ e_n, \delta s_n, \delta t_n\ \rangle]$ where e_x is x-th event type in S_1, $S_2, \ldots, S_q$, δs_x and δt_x are the x-th spatial and temporal gap values in the corresponding sequences. □

Example. Figure 1 shows an example of event sequences related to the monitoring of the occupancy of the bike sharing stations in an urban environment. Specifically, the monitored events, i.e., *fully occupied, almost fully occupied, and increasing occupancy level* indicate the levels of occupancy of the stations of a bike sharing system, expressed in terms of percentage of available docks. We consider as temporal resolution the 15-minute time slots depicted as consecutive time frames. To model the neighborhood of trigger events e_1 and e_4, we set the spatial resolution to r=100m and plot surrounding circles indicating the spatial distances $1 \cdot r$=100m, $2 \cdot r$=200m, and $3 \cdot r$=300m, respectively. Two STInvs in Figure 1 are depicted as oriented splines connecting different occurrences of spatiotemporal events. Specifically, they are represented as a dashed line and a continuous line, respectively.

Pattern example 1. Let us consider the STInv with trigger event (*Almost full*). It is observed in the time slot [10.00am, 10.15am]. It shows a spatial correlation between events e_1 (*Almost full*) and e_2 (*Increase*). More specifically, e_2 occurs in the neighborhood of the trigger event at a distance ranging between $2 \cdot r$=200m and $3 \cdot r$=300m, i.e., (200m - 300m]. Next, 15 minutes after the occurrence of the trigger event e_1, we can also observe in the same location the occurrence of event e_5 (*Full*). Such a combination of spatiotemporal events (e_1-e_2 and then e_5) is worth considering because another spatiotemporally invariant occurrence can be observed at the upper-right corner of the spatial areas in time slots [10.15am, 10.30am] and [10.30am, 10.45am] (events e_7-e_8 and then e_{11}). The STInv can be formulated as follows: [⟨ Almost Full, δs=0m, δt=0 ⟩, ⟨ Increase, δs=200m-300m, δt=0 ⟩, ⟨ Full, δs=0m, δt=15mins ⟩]

To simplify the representation and improve readability, we can omit the temporal gap indication at the event occurrences while reporting the relative temporal gap δt observed between the event sets occurring in different time slots:

$$\{\text{Almost Full}(\delta s\text{=0m}), \text{Increase}(\delta s\text{=200m-300m})\} \xrightarrow{\text{15 mins}} \{\text{Full}(\delta s\text{=0m})\}$$

This STInv can be interpreted as follows: If *a bike sharing station is almost full and the occupation level of at least one of its nearby stations (between 200m and 300m) is increasing* then *within 15 minutes the occupancy of the same station will get full.*

The pattern mining task. Let *minsup* be a minimum support threshold and D be a sequence database. A *frequent* STInv in D is an STInv whose number of corresponding event sequences in D is above *minsup*.

Given a sequence database D, our purpose is to automatically extract all the frequent STInvs in D.

3 THE STINV-MINER ALGORITHM

We present a new algorithm, namely STInv-Miner, to address the STInv pattern mining task. STInv-Miner leverages a prefix-projected-like pattern growth strategy [3]. To enable its adoption for STInv pattern mining, we first need to tailor the concept of sequence projection to STInvs.

The key idea is to transform the original event sequences, annotated with absolute temporal and spatial information, into their corresponding projected versions including relative time and spatial gaps with respect to each trigger event.

For each trigger event in the sequence we generate a projected sequence from the original one that contains the sequence of non-trigger events annotated with their relative spatial and temporal distance, i.e., the spatial and temporal gaps. The projected sequences corresponding to all event sequences in D are stored in a projected sequence database D^p.

Definition 3.1. Let be S_i be an event n-sequence and let $E^{tr}(S_i)$ be the set of trigger events in S_i (which are a subset of the whole event set $E(S_i)$ occurring in S_i). The projection function maps each S_i to a set of q distinct sequences $S^p_1, S^p_2, \ldots, S^p_q$, namely the *projected sequences*. Each projected sequence corresponds to a distinct trigger event in $E^{tr}(S_i)$ belonging to the first time slot of the sequence S_i. The j-th projected sequence S^p_j is defined as follows

$[\langle\ e^{tr}, \delta s = 0, \delta t = 0\ \rangle, \langle\ e_2, \delta s_2, \delta t_2\ \rangle, \ldots, \langle\ e_m, \delta s_m, \delta t_m\ \rangle]$

where e^{tr} is the j-th trigger event, $e_2, \ldots, e_m$ are the events in S_i that are spatially correlated with e^{tr}, and the corresponding δs and δt are the spatial and temporal gaps relative to the trigger event. □

The STInv-Miner algorithm leverages a projected version of the sequence database by tightly integrating the projection and mining phase into a scalable, distributed mining process executed using Apache Spark. Event projections are computed on top of a discretized version of the spatiotemporal event occurrences fulfilling the desired temporal and spatial resolutions.

4 EXPERIMENTS

4.1 Datasets and configurations

The *Bike Sharing Dataset* contains data related to bike sharing stations' occupancy rates of 5 different cities in the San Francisco Bay Area, sampled every minute in a 2-year period. For simplicity, all the reported results are related to San Francisco, Palo Alto and the global view (35, 5, and 70 stations).

To characterize the occupancy levels of the stations, we divided the temporal axis into timeslots of fixed size (i.e., temporal resolution) and generated an event dataset, defining the following events: *Increase* in station's occupancy level, *Decrease* in station's occupancy level, *Almost full* (at most 2 free slots available) and *Full.*

Table 2: Configuration settings.

	Config 1	Config 2
Dataset	Bike Sharing	Bike Sharing
Temporal Resolution (td)	10 min	10 min
length (nr. of time slots)	6	6
Spatial resolution (r)	100m	100m
Geographical area	Global, City	Global, City
Max pattern length	6	6
Event types	Full, Almost Full, Increase	Full, Almost Full, Increase, Decrease
Trigger event types	Full, Almost Full, Increase	Full, Almost Full, Increase

Table 3: Number of generated events for the bike sharing dataset. Values associated to events are expressed in %.

	Global		San Francisco		Palo Alto	
	Conf 1	Conf 2	Conf 1	Conf 2	Config 1	Config 2
Full	3.94	2.26	5.03	2.92	2.39	1.42
Almost Full	22.12	12.67	23.45	13.60	30.26	18.05
Increase	73.94	42.35	71.52	41.47	67.35	40.18
Decrease	N/A	42.72	N/A	42.01	N/A	40.35
Total events	1.08M	1.88M	708K	1.22M	74K	124K

We performed 2 different experiments (Conf 1 and 2 in Table 2), with the sole difference that Conf 2 also considers the *Decrease* events and analyses events taking place in the top-N departure stations instead of considering the station's neighborhood.

4.2 Pattern extraction and analysis

We performed several pattern extractions on geographical areas of different size. We adopted two configurations: (i) Conf 1 to mine local patterns and understand interactions between neighboring stations, and (ii) Conf 2 to identify long-term spatial departure-arrival patterns. *minsup* is set to 1 in both cases.

Statistics on events and STInv patterns. Table 3 reports the event type distribution, while Table 4 summarizes the characteristics of the mined STInvs when considering the global view (i.e., all cities together). Several complex and expressive patterns are mined. They are characterized by an average number of event occurrences per STInv close to 6. Since many of them show different spatial/temporal gaps, they represent non-trivial invariant patterns that human experts are unlikely to detect during a manual data exploration.

Comparison with traditional event sequences and non-spatially invariant sequences. Analyzing the percentage of STInvs with $\delta s > 0$ and $\delta t > 0$, we exclude all the patterns that disregard variations in the spatial and temporal dimension. Both the values are above 99.9% (Table 4), thus confirming the added value of exploring STInvs rather than simpler event sequences. Finally, we compare the number of STInvs with the number of traditional sequences mined by considering the absolute location of the events (Table 5), relaxing the spatial invariance constraint. The results demonstrate the summarization capability of the proposed approach.

Table 4: Statistics on the bicycle sharing dataset. Global area.

	Config 1	Config 2
Total num. of STInvs	45.3M	11.0M
Mean #triplets per sequence	5.90	5.77
Mean discrete spatial distance	2.40	8.31
Mean discrete temporal distance	2.13	2.17
#sequences with at least one $\delta s > 0$ (%)	99.95	99.89
#sequences with at least one $\delta t > 0$ (%)	99.99	99.99
#sequences with at least one $\delta s > 0$ *and at least one* $\delta t > 0$ (%)	99.29	98.36

Table 5: Compression Ratio (*CR*) of STInv compared to absolute non-spatially invariant patterns (Abs).

	Conf 1			Conf 2		
	#STInv	#Abs	CR	#STInv	#Abs	CR
San Francisco	45.3M	2350M	51.90	6.24M	28.2M	4.53
Palo Alto	111K	143K	1.29	344K	592K	1.72

5 CONCLUSIONS

We present a new type of patterns denoting spatiotemporally invariant relations among sequences of discrete events. The key contribution is the generalization of the joint spatial and temporal relations over multiple data instances. The empirical results, achieved on a real-world dataset, show that (i) STInvs averagely summarize the temporal correlations described by tens of traditional sequences and (ii) STInv patterns show non-zero spatial and temporal gaps in most cases.

ACKNOWLEDGMENTS

This work is partially funded by the SmartData@PoliTO center.

REFERENCES

[1] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. 2006. Efficient mining of temporally annotated sequences. In *SDM'06*. 348–359.

[2] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory Pattern Mining. In *ACM SIGKDD'07*. 330–339.

[3] Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. 2001. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *ICDE'01*. 215–224.

[4] Di Huang, Shuaian Wang, and Zhiyuan Liu. 2021. A systematic review of prediction methods for emergency management. *Int. J. Disaster Risk Reduction* 62 (2021), 102412.

[5] Zhenyu Liu, Zhengtong Zhu, Jing Gao, and Cheng Xu. 2021. Forecast Methods for Time Series Data: A Survey. *IEEE Access* 9 (2021), 91896–91912.

[6] Sobhan Moosavi, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath. 2019. Short and long-term pattern discovery over large-scale geo-spatiotemporal data. In *ACM SIGKDD'19*. 2905–2913.

[7] Sara Paiva, Mohd Abdul Ahad, Gautami Tripathi, Noushaba Feroz, and Gabriella Casalino. 2021. Enabling Technologies for Urban Smart Mobility: Recent Trends, Opportunities and Challenges. *Sensors* 21, 6 (2021).

[8] Florian Verhein. 2009. Mining Complex Spatio-Temporal Sequence Patterns. In *SDM'09*. 605–616.

[9] Peizhong Yang, Lizhen Wang, Xiaoxuan Wang, Lihua Zhou, and Hongmei Chen. 2021. Parallel Co-location Pattern Mining based on Neighbor-Dependency Partition and Column Calculation. In *ACM SIGSPATIAL'21*. 365–374.

[10] Mohammed J Zaki. 2005. Efficiently mining frequent embedded unordered trees. *Fundamenta Informaticae* 66, 1-2 (2005), 33–52.