GeoInformatica
https://doi.org/10.1007/s10707-023-00506-4

Discovery of multi-domain spatiotemporal associations

Prathamesh Walkikar[1] · Lei Shi[1] · Bayu Adhi Tama[1] · Vandana P. Janeja[1]

Received: 1 July 2021 / Revised: 19 June 2023 / Accepted: 14 September 2023

Abstract
This paper focuses on the discovery of unusual spatiotemporal associations across multiple phenomena from distinct application domains in a spatial neighborhood where each phenomenon is represented by anomalies from the domain. Such an approach can facilitate the discovery of interesting links between distinct domains, such as links between traffic accidents and environmental factors or road conditions, environmental impacts and human factors, disease spread, and hydrological trajectory, to name a few. This paper proposes techniques to discover spatiotemporal associations across distinct phenomena using a series of anomalous windows from each domain that represent a phenomenon. We propose a novel metric called influence score to quantify the associated influence between the phenomena. In addition, we also propose spatiotemporal confidence, support, and lift measures to quantify these associations. Two novel algorithms for finding multi-domain spatiotemporal associations across phenomena are proposed. We present experimental results across real-world phenomena that are linked and discuss the efficacy of our approach.

Keywords Spatiotemporal associations · Anomalies · Spatial neighborhood · Spatiotemporal confidence and support

1 Introduction

A fundamental law of geography says that everything is related to everything else [1]. Many studies show this in single application areas, for example, rents in a community are similar, and traffic in a vicinity is largely similar. This fundamentally also becomes true across multiple areas. For example, (a) weather conditions at a location will impact traffic [2], (b) oil spills in oceans will adversely impact underlying aquatic animal population [3], (c) pollution at locations can affect disease spread [4], and many more. In addition, well-known phenomena in polar science could also be drawn from such regularity. Changes in ice sheet mass are deemed to be the key to understanding present and future sea level rise [5, 6], in which the contribution to sea level rise depends on the interactions between ice, ocean, and atmosphere. It is estimated that surface meltwater runoff, basal melting, and precipitation will all increase

✉ Vandana P. Janeja
vjaneja@umbc.edu

[1] Department of Information Systems, University of Maryland, Baltimore County, MD 21250, USA

Published online: 03 October 2023

in a warmer climate, as evidenced for both Greenland and Antarctica ice sheets [7]. However, observing ice mass changes associated with atmospheric and ocean forcing of the ice sheets with a dataset has remained difficult despite advances in understanding ice sheet reactions to climate change [8].

These are some common regularities that have long been established and validated by long and rigorous studies. There are possibilities of links among domains in the underlying space due to the fact that in a region of space a single process can govern the behavioral changes across multiple domains; for example, childhood poverty and unemployment in a region are related mainly because they are both influenced by lack of education. One thing that makes studying such domain influences difficult is the vast amount of data generated even for a single application domain. In addition to this, combining data from different application domains can be challenging due to data heterogeneity. Now, what if we could avoid that by not looking at individual raw data from distinct application domains but instead utilizing extracted knowledge from each domain and further mining it to identify associated links between the domains using space as the common point of reference.

The focus of this paper is to provide a mechanism by which we can study spatiotemporal associations across two distinct application domains (such as traffic and weather or childhood poverty and unemployment) in the same spatial region by combining the knowledge derived in each domain. This is a subset of a bigger problem of looking at several associations in space, which is currently out of scope in this paper. However, we intend to build the larger model by currently considering the process of associations across two domains and later on considering associations across several domains simultaneously.

We study these domain associations in the form of some phenomena captured in each domain dataset. For example, in studying poverty, we look at the phenomena of unusual poverty rates in a region. Thus, the knowledge we consider discovering the domain associations is derived from the anomalous windows representing the phenomenon in each domain, for example, an unusual spatial window, which depicts highly unusual poverty rates. An anomalous window comprises a set of points in a region that are unusual with respect to the rest of the data points in that region in terms of some attribute of interest. Current methods in spatial and spatiotemporal scan statistics study the unusualness of the phenomenon and detect anomalous hotspots, which are anomalous spatial and spatiotemporal windows measured at single or multiple time intervals.

Existing studies in co-location pattern mining [9] aim to discover the association between two or more spatial objects with respect to non-spatial attributes. However, they mainly look at spatial co-locations and not necessarily spatiotemporal co-locations. Moreover, existing approaches to multi-domain link discovery [10] do not quantify the strength of the relationships and also do not address the discovery of spatiotemporal relationships.

In this paper, we propose to identify the relationships between domains based on proximity and overlap patterns of the anomalous windows representing the phenomena in each domain. In other words, the co-occurrence of different anomalous windows from multiple phenomena in areas of proximity determines spatial association. However, instead of using geographic distance measures only, we propose a novel influence measure that utilizes knowledge from the underlying phenomenon of study. Furthermore, we also discover spatiotemporal associations between phenomena based on spatial associations. This measures the links between phenomena with spatial associations, which are long-lasting across a certain time period.

Our key contributions are as follows:

(a) We present a novel framework of algorithms and metrics for the discovery of spatiotemporal associations between phenomena from distinct application domains.

(b) We introduce a metric called influence score for the measurement of spatial association of phenomena, as well as a novel variation of spatiotemporal confidence, support and lift measures for the measurement of spatiotemporal associations across phenomena.
(c) We present a method for evaluating the statistical significance of the spatiotemporal associations using Monte Carlo simulations.
(d) We conduct detailed experiments on synthetic and real-world datasets, discovering spatiotemporal associations among phenomena indicating strong influence relationships between them, which demonstrates the efficacy of our algorithm.

2 Related work

Mining statistically significant associations across multiple domains has related works in co-location pattern mining [9] and trajectory mining method [11]. Existing literature presents two broad approaches in co-location mining, namely statistical approaches and data mining approaches. Statistical approaches can be sub-divided into spatial as well as temporal analysis techniques. Spatial analysis techniques discover the frequent co-location rules based on correlation measures such as Ripley's Cross-K function [12, 13]. Extensive studies have been done in analyzing the temporal analysis techniques for co-location patterns, which include first-order and second-order autocorrelation [14] and periodic pattern discovery methods like periodicity transform proposed in [15]. However, spatial correlation measures suffer from the disadvantage of expensive computation due to the exponential generation of candidate subsets for large spatial Boolean features [16]. Data mining approaches include map overlay-based clustering techniques [17, 18] which employ a layered approach for point-data to obtain spatial association rules [19]. Spatial association rule mining, on the other hand, employs general association rule mining together with spatial predictors to find interesting spatial associations [20].

Some of the other interesting works that deal with spatiotemporal patterns in the spatiotemporal domain include [21, 22]. Tao et al. [21] uses a brute-force approach in the detection of spatiotemporal association rules (STAR). However, despite creating simplified STAR definitions, the underlying spatial and temporal semantics like interest neighborhood and time interval width are ignored. The work mentioned above is quite different from ours, which involves studying underlying anomalies in potentially interrelated spatiotemporal phenomena across multiple time intervals.

Spatial and spatiotemporal scan statistics methods [23, 24], hotspot detection methods [25, 26], and also multivariate scan statistics approaches [14, 27] address the detection of anomalous windows. However, spatiotemporal scan statistics-based methods typically focus on a single phenomenon of study, and in contrast, the multivariate methodologies do not address the linked association discovery among application domains. We address this issue of discovering associations across multiple domains by proposing a novel framework that utilizes resultant individual domain anomalies and ties them together with space as the reference point but also considers time series into account.

Other approaches for the identification of co-location patterns, such as [9] utilize monotonic composite interest measures with space and use time prevalence thresholds to eliminate candidate sets for associations. In contrast to this, our approach utilizes the underlying multi-domain spatiotemporal anomalies and finds linkages between them, thus reducing candidate phenomena instead of analyzing complete large datasets. Also, on similar lines to the concept of R-proximity neighborhood [16], we use a network of phenomena connected by influence

distances which are governed according to the underlying phenomena instead of only the spatial distances.

Using the basic concepts of trajectory mining from [11] and [28], we design a hierarchical strategy for creating a trajectory of anomalous window geographic centers over time to analyze the movement of anomalous windows. We find trajectory clustering approaches [11, 28] to be closely related to our approach; however, with one significant difference. Anomalous windows typically consist of multiple spatial objects, whereas trajectory-clustering approaches tend to work on single-point trajectories. Also, in contrast to the prevalence measure of participation index (or support) and conditional probability (or confidence) proposed in [9, 16], we propose a significant variation of spatiotemporal support, confidence, and lift measures for each of the proposed approaches to quantify the relationships across phenomena.

The rest of the paper is organized as follows: in Section 3 the approach and associated terminologies are discussed in detail. Section 5 discusses the experimental evaluation and results. Section 6 discusses the overall system architecture for mining phenomena-related associations, and finally, we conclude in Section 7.

3 Methodology

Figure 1 describes our overall approach, and the terminologies used throughout the paper are shown in Table 1. The discovery of unusual spatiotemporal associations starts with the phenomenon representation in each domain, which is done by the discovery of anomalous windows. These windows capture the unusualness of spatial nodes in terms of an attribute of interest. For instance, a disease outbreak can be measured in terms of the number of disease cases with respect to the base population. Thus, we identify and quantify these anomalous windows in the initial step for different domains. Our approach is not fixed to a particular type of anomalous window detection method and is adaptable to multiple such methods. Next, we discover the spatial associations between phenomena where the anomalous windows build the basis for detecting spatiotemporal associations. We next explain each step in detail.

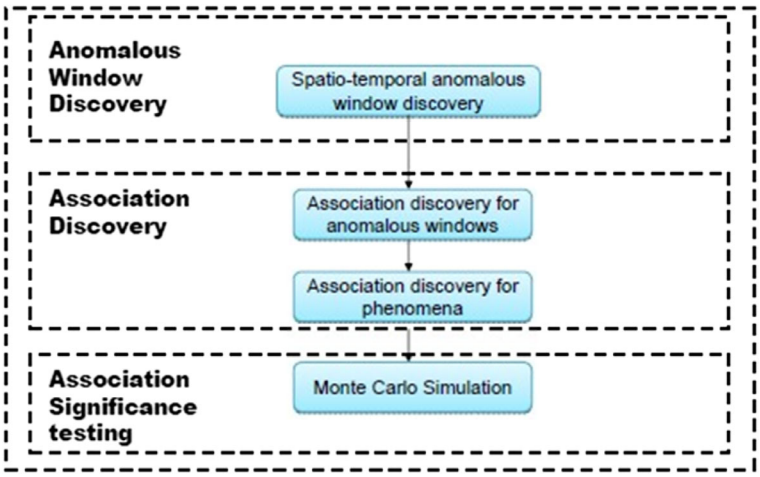

Fig. 1 Discovering unusual spatiotemporal associations

Table 1 Summary of terminologies used in the paper

v	Phenomenon of paths
$d^v_{p\to q}$	Influence distance from spatial object p to q for phenomena v
$s^v_{p\to q}$	Influence score from spatial object p to q for phenomena v
δ	Influence decay rate
A	Anomalous window
A^v	Anomalous window for phenomena v
s^v_{pa}	Influence score with primary anomalies only
s^v_{aa}	Influence score with primary and secondary anomalies both
s^v_{wc}	Influence score considering anomalous window centers
STC_p	spatiotemporal confidence considering all anomalies
STS_p	spatiotemporal support considering all anomalies
STL_p	spatiotemporal lift considering all anomalies
$\overline{STC_p}$	spatiotemporal confidence considering only primary anomalies
$\overline{STS_p}$	spatiotemporal support considering only primary anomalies
$\overline{STL_p}$	spatiotemporal lift considering only primary anomalies

3.1 Anomalous window discovery

Our aim is to quantify links between domain datasets that have some phenomena taking place in them; for example, a domain dataset of child poverty data could be capturing data on counts of children under poverty and total population counts. The phenomenon in this dataset is the clusters of child poverty, which can be measured by the unusualness of the number of children who are poor in a contiguous region. We might want to consider linking this poverty data to another domain of unemployment where the phenomena are clusters of unusual unemployment rates in the region. We want to study the relationships between these two distinct sets of clusters. We utilize anomalous window discovery to find these clusters in the domain datasets.

Definition 1 (Anomalous Window) An anomalous window A^v is a collection of spatial objects $s_{A^v} = s_1, \cdots, s_n$ in proximity, where each node has associated spatial and non-spatial attributes, such that s_{A^v} has a quantified unusual behavior as compared to that of the other spatial objects in the data. The unusualness is quantified as a measurement of the phenomena being observed.

Timely detection of such unusual phenomena is crucial; hence, appropriate solutions can be devised and implemented to mitigate further other risks. As another illustration, in the polar science domain, spatial variability in warming (e.g., measured by mean annual temperature)

might be helpful to understanding the loss of land and sea ice, extreme weather at lower latitudes, and low biodiversity in a polar region [29].

We consider phenomena as represented by a set of anomalous windows, which capture unusual behavior in terms of an attribute of interest, for example, the count of cases of childhood poverty. The first step is to employ available and existing techniques to detect anomalous windows. Spatial scan statistics [23, 24, 30, 31] and spatiotemporal scan statistics [32] techniques can be utilized in this regards to identifying anomalous windows for each phenomenon of study. It is important to note that our approach is not specific to one such technique and is adaptable to multiple types of anomalous window detection techniques.

Let us consider child poverty data, which is represented in terms of spatial objects with distinct interest measures such as the number of children under the poverty level, age group, and sex. This data can be captured in terms of location coordinates (spatial attributes) and the number of children meeting the poverty criteria (non-spatial attributes). After discovering spatiotemporal anomalous windows using any of the available detection techniques, we then look for spatial associations between the phenomena. Table 1 summarizes some of the terminologies used throughout this paper.

We utilize these anomalous windows to define spatiotemporal phenomenon. A spatial phenomenon v is represented by an anomalous window A^v which represents a set of spatial objects $s_{A^v} = s_1, ..., s_n$ where each node has associated spatial and non-spatial attributes. Extending this notion to spatiotemporal windows where phenomena vary over time can be represented by a set of anomalous windows for each phenomenon $A^v = \{A^v_{t_1}, A^v_{t_2}, ..., A^v_{t_n}\}$ over discrete intervals of time t_1 to t_m. These anomalous windows across application domains can hold information about significant hidden influence relationships between phenomena, which we aim to quantify. We first outline the measures we use to quantify associations and then explain the discovery of associations between anomalous windows.

3.2 Influence distance and influence score

We first start by defining the basic terminologies we use to quantify spatial associations. Then, starting with anomalous windows associated with distinct phenomena, we propose that spatial associations are quantified by two relationships, namely proximity and overlap of the anomalous windows in each domain, suggesting strongly associated spatial phenomena. Proximity represents how close the windows are in the geographic vicinity, and overlap describes the identical nature of anomalous windows in terms of spatial locations. Figure 2 describes the basic terminologies used in this paper to quantify the spatial associations.

We introduce the notion of influence distance to measure these spatial properties. Unlike the traditional spatial distances, the influence distance is associated with the underlying phenomena of study which governs the anomalous windows. Influence distance thus indicates the hardship of influence of spread between any two spatial objects. The network of phenomena describes the flow of influence among spatial objects in a neighborhood. In this network, vertices represent spatial objects, and edges represent the possible influence flow. The weight on each edge indicates the influence reaches from one spatial object to another. Such a weight takes into account the spatial distance and other underlying phenomena governed by related factors such as barriers, which can even be geopolitical in nature. For example, suppose we have an interrelated network of phenomena for two distinct domains located in the same spatial neighborhood, as shown in Figure 3.

Let $L = \{L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8, L_9, L_{10}\}$ represent set of geographical locations in a particular region of study. For instance, Caroline and Allegany counties, even

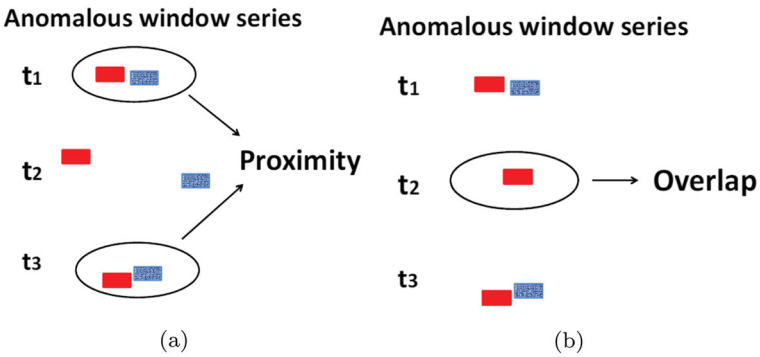

Fig. 2 Two different spatial properties, illustrating the spatial associations between spatial objects, such as proximity (a) and overlap (b)

though they are in close proximity to each other, being in the same state of Maryland, will have different rates of child poverty due to different demographics and county-level initiatives taken to eliminate child poverty. Thus, despite their proximity, their link weights can be small due to consideration of underlying phenomena-related factors. Figure 3 represents a network of child poverty cases and unemployment cases. We can see that locations $\{L_1, L_2, L_3, L_4, L_5, L_8, L_9\}$, share common edges due to detection of child poverty clusters in the adjacent areas. We can calculate the distance between any two spatial objects by considering the minimum number of hops required to traverse to that particular object in a network. Considering every edge has a weight equal to 1, we can calculate the influence distance between any two spatial objects, such as $d_{1\to4}^{\text{child poverty}} = 2$ and $d_{1\to3}^{\text{unemployement}} = 1$.

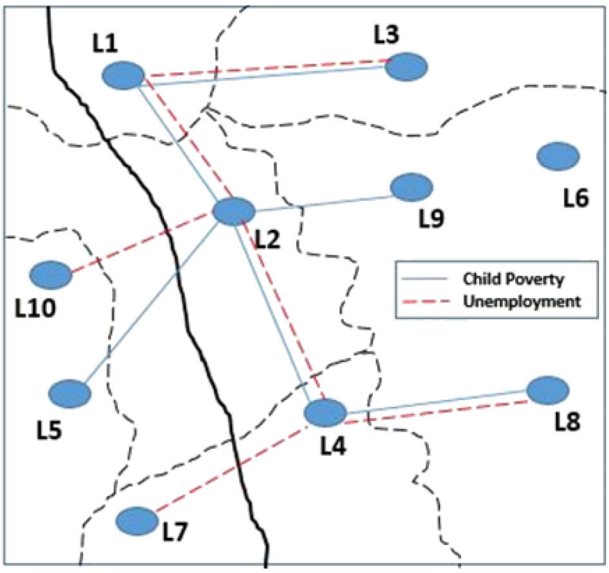

Fig. 3 Example of an interrelated network of phenomena

We use influence distance to capture the underlying phenomena-related factors in capturing proximity and overlap, such as location-related factors, the directionality of spread, and others, as discussed in prior work [33]. We define influence distance as follows:

Definition 2 (Influence distance) Let v be the given phenomenon, and p and q be two spatial objects. We define $d^v_{p\to q}$ as the influence distance from spatial object p to q for the phenomenon v. $d^v_{p\to q}$ is the sum of the weights of the constituent edges of the shortest path from p to q in the network of v. If p and q are one spatial object, then $d^v_{p\to q} = 0$. If p and q are not connected, then $d^v_{p\to q} = \infty$.

Based on influence distance, we calculate the influence score, which quantifies the proximity and overlap of anomalous windows. The influence score is defined as follows:

Definition 3 (Influence score) Let $d^v_{p\to q}$ be the influence distance from spatial object p to q for the phenomenon v. We define $s^v_{p\to q}$ as the influence score that measures the influence of v that spatial object p has on q. $s^v_{p\to q}$ is then computed as $s^v_{p\to q} = e^{-(d^v_{p\to q}\times(1/\delta^v))}$ where $\delta > 0$ is the influence decay rate for a phenomena v. If p and q are one spatial object, then $s^v_{p\to q} = 1$. If p and q are not connected, then $s^v_{p\to q} = 0$.

The influence score ranges from 0 to 1. Thus, we can say that the larger the influence scores, the greater the influence of a spatial object on another. Since the influence score is dependent on the influence distance between spatial objects, an increase in distance between two spatial objects can cause a diminishing effect on the influence between them. We use the influence decay rate δ to model the diminishing speed. The greater the value of δ, the quicker the influence decreases as the distance from the origin of influence increases. Table 2 describes an example calculation for computing influence score.

Extending this idea to the concept of anomalous windows, we can compute the influence score between anomalous windows, given influence scores between constituting spatial objects. Let A^v be the anomalous window of phenomenon v and $A^{v'}$ be the anomalous window of phenomenon v'. We quantify the influence that a phenomenon v (represented by A^v) has on phenomenon v' (represented by $A^{v'}$) through the influence score using Equation 1.

$$s^v_{A^v\to A^{v'}} = \frac{\sum_{q\in A^{v'}} \mathbf{max}(s^v_{p\to q})_{p\in A^v}}{\mid A^{v'} \mid} \tag{1}$$

where $\mathbf{max}(s^v_{p\to q})_{p\in A^v}$ is calculated for every spatial object q in anomalous window $A^{v'}$ given any spatial object p in anomalous window A^v. We use the maximum function to capture all possible outlier behavior. That is, we aim at preserving striking examples of behavior. The average of this maximum influence score of all spatial objects q in anomalous window $A^{v'}$ is then the influence that A^v has on $A^{v'}$, denoted by $s^v_{A^v\to A^{v'}}$. If A^v and $A^{v'}$ are identical, there must be the $s^v_{A^v\to A^{v'}} = 1$ for every $q \in A^{v'}$, there must be same $q \in A^v$ that $s^v_{p\to q} = 1$. If A^v and $A^{v'}$ are not connected by any of their spatial objects, then $s^v_{A^v\to A^{v'}} = 0$, since for any

Table 2 Example Influence Score Computation

Phenomena	Decay Rate	Influence distance	Influence score
Child Poverty	10	0.8	0.92
Unemployment	50	0.2	0.99

$p \in A^{v}$ and $q \in A^{v'}$, p is not connected to q that $s^{v}_{p \to q} = 0$. The proximity of anomalous windows can be indicated by an influence score larger than a certain threshold s_T, which acts as a lower bound for the proximity of anomalous windows. Thus, a maximum influence score of 1 indicates a strong overlap between anomalous windows.

Given a set of anomalous windows for two distinct phenomena (v) within the same spatial neighborhood and same time period, the network of phenomena is created as a two-dimensional adjacency matrix A_{network} structure with each cell $A_{\text{network}}[i, j]$ represents the individual influence distance between each individual spatial node. Next, we compute the influence scores as explained above and thus obtain the influence score matrix $M_{\text{influence}}$ where $M_{\text{influence}}[i, j]$ represents the influence score between each spatial objects taken into consideration. A sample influence score matrix is shown in Figure 4 below between the counties L_1 to L_{10}. This sample influence matrix is built based on the network of phenomena depicted in Figure 3. It shows influence scores computed only between corresponding locations defined in the network of phenomena.

4 Associating anomalous windows

The process of associating multiple domains starts with the anomalous window discovery within each domain, for example, child poverty and unemployment in the State of Maryland. We first discretize the time series into distinct temporal intervals $T_1, T_2, T_3, \cdots, T_n$ using various data discretization strategies such as equal frequency discretization, equal width discretization, and hierarchical time-series clustering-based discretization strategy. We then apply multi-domain association discovery algorithms (MDA), shown in Figure 5, in each temporal bin interval to get the overall influence between the two domains. We consider the proximity and overlap patterns (MDAWnU-CP) of the anomalous windows while quantifying the influence relationships across domains. Table 3 illustrates an example of detected single domain anomalies at each temporal interval along with the decay rate for the phenomena. We next discuss two methods of finding associations between anomalous windows: (1) pairwise approach, which utilizes pairwise relationships between location pairs in anomalous windows, and (2) windows centers-based approach, which utilizes the geographic center of the anomalous windows at each time interval.

4.1 Pairwise influence relationship approach

This approach utilizes pairwise relationships between location pairs in the anomalous windows at any time instant T_i based on influence scores to compute the associations between two

Location	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
L1	1	0.904837418	0.904837418	0.818730753	0	0	0	0.740818	0.818731	0
L2	0	1	0.818730753	0.904837418	0.904837	0	0	0.818731	0.904837	0
L3	0.904837418	0.818730753	1	0	0.904837	0	0	0.67032	0.740818	0
L4	0.818730753	0.904837418	0.740818221	1	0	0	0.818731	0.904837	0.818731	0
L5	0.818730753	0	0.740818221	0.818730753	1	0	0.904837	0.740818	0.818731	0
L6	0	0	0	0	0	1	0	0	0	0
L7	0	0	0	0	0	0	1	0	0	0
L8	0.740818221	0.818730753	0.670320046	0.904837418	0.740818	0	0	1	0.740818	0
L9	0.818730753	0	0.740818221	0.818730753	0.818731	0	0	0.740818	1	0
L10	0	0	0	0	0	0	0	0	0	1

Fig. 4 Example of influence score matrix

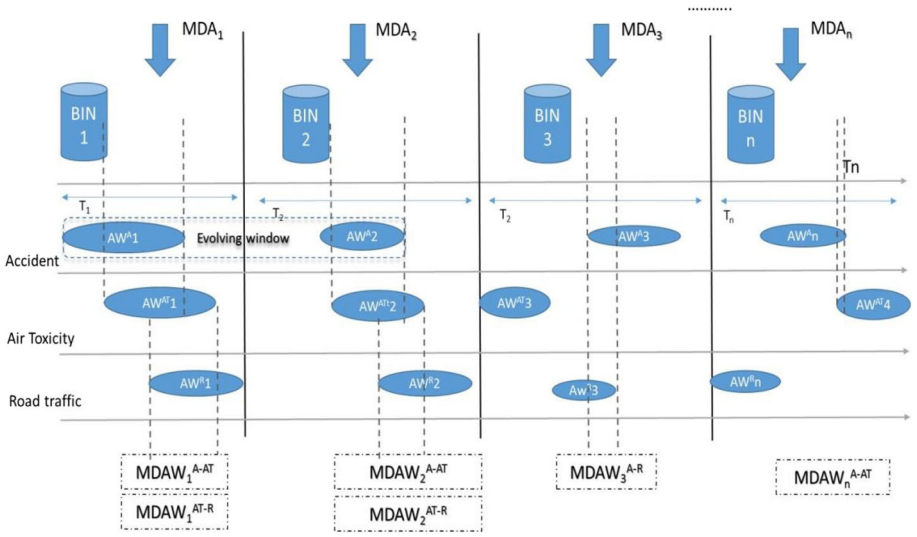

Fig. 5 Multi-domain association framework

anomalous windows. Figure 6 illustrates this approach where $AWA_1^U = \{L_1, L_2, L_3\}$ and $AWA_1^{CP} = \{L_2, L_3, L_4, L_5\}$ represent anomalous windows for unemployment and child poverty respectively at the discrete time interval T_1 (where $t_1 \in T$ and $T = \{T_1, \cdots, T_m\}$) and $L = \{L_1, L_2, L_3, L_4, L_5\}$ represent constituent locations in the anomalous windows. At time interval T_i, we can associate anomalous windows A_{pd1} and A_{pd2} by considering all the best possible combinations of location pairs present in anomalous windows across both the domains based on influence scores. We compute the total influence of $A_{\text{unemployment}}$ on $A_{\text{child poverty}}$ using Equation 1 to find overlaps between anomalous windows across domains.

We later compute the overall influence score between the two anomalous windows given by the Equation 1, which constitutes the best pairing between the location pairs. A sample illustration of this is shown in Table 4 below. For example, in time T_1 we see the window for child poverty is $\{L_1, L_2, L_3\}$ and unemployment is L_2, L_3, L_4, L_5. For each time slice, we compute the influence scores. We then aggregate the influence scores obtained across each temporal interval to get the final influence between unemployment rates and child poverty.

We show the process for this complete approach in Algorithm 1. The algorithm takes a series of anomalous windows from distinct, interrelated phenomena. It computes the overall influence of one domain over another at a particular time interval t_i by using the best pairing between location pairs in the anomalous windows. Line 1 computes influence distance across

Table 3 Sample anomalous windows at each temporal interval

T	Child poverty	Unemployment	Decay rate
T_1	$\{L_1, L_2, L_3\}$	$\{L_2, L_3, L_4, L_5\}$	10
T_2	$\{L_3, L_4, L_5, L_8, L_{10}\}$	$\{L_2, L_6, L_7, L_8\}$	10
T_3	$\{L_6, L_7, L_3, L_2\}$	$\{L_1, L_6, L_8, L_9\}$	10
T_4	$\{L_3, L_5, L_4, L_7, L_2\}$	$\{L_4, L_6, L_9, L_2, L_3\}$	10

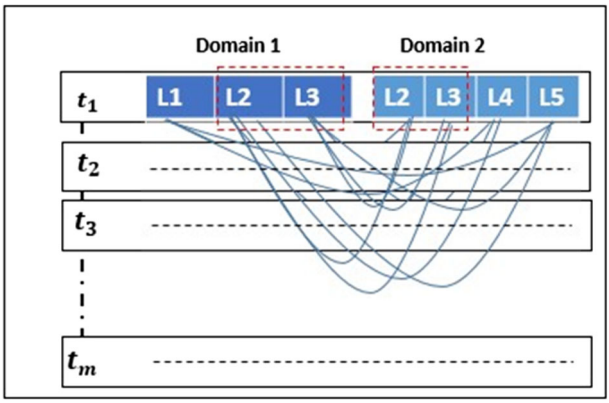

Fig. 6 Pairwise influence relationship approach

anomalous windows for spatial location combinations between them. Lines 2 to 6 are used to find the maximum influence score for each spatial location present in A_{d2} for every location available in A_{d1}. Line 7 computes the overall influence score of A_{d1} on A_{d2}. Lines 8 to 12 compute the best available pairing of locations in A_{d2} given each spatial location in A_{d1}. The goodness of pairing is determined by the total influence score of the pairs. Lines 13 to 15 quantify these discovered associations across domains.

We quantify these associations by proposing a novel variation of spatiotemporal confidence (STC_p), support (STS_p), and lift measure (STL_p) for our approach. The quantification of these associations involves considering results based on primary (highly significant) only or both primary and secondary (relatively less significant) anomalies. This leads to two significant variations of formulae for computation of these measures: a) Primary Influence Relationship: One considering only the primary anomalies with respect to all the obtained anomalies (both primary and secondary), and b) Complete Influence Relationship: Considering all the obtained anomalies (considering both primary and secondary). The formulae for both versions are explained in Tables 5 and 6 below.

(a) Primary influence relationship
We utilize the primary anomalous windows (which is highly statistically significant) and analyze their influence with respect to all the discovered anomalous windows (both primary and secondary). Table 5 explains various quantification measures regarding primary influence relationships between two phenomena, v and v'.
(b) Complete influence relationship

Table 4 Sample influence score between anomalous windows

T	Child poverty	Unemployment	Decay rate	Influence score
T_1	$\{L_1, L_2, L_3\}$	$\{L_2, L_3, L_4, L_5\}$	10	0.72620
T_2	$\{L_3, L_4, L_5, L_8, L_{10}\}$	$\{L_2, L_6, L_7, L_8\}$	10	0.90710
T_3	$\{L_6, L_7, L_3, L_2\}$	$\{L_1, L_6, L_8, L_9\}$	10	0.70241
T_4	$\{L_3, L_5, L_4, L_7, L_2\}$	$\{L_4, L_6, L_9, L_2, L_3\}$	10	0.76193

Algorithm 1 Pairwise influence relationship approach

Require: Phenomena for analysis $\{v1, v2, v3, \ldots.., vn\}$, series of anomalous windows. $\{A^{v1}, A^{v2}, A^{v3}, A^{v4}, ..., A^{vn}\}$ each at time intervals $\{t1, t2, ..., tn\}$.
Ensure: Each anomalous window A^{vm} is a set of spatial locations S_j^{vm} where $S_j^{vm} = \{S_1^{vm}, S_2^{vm}, S_3^{vm}, ..., S_n^{vm}\}$.
Ensure: Associated Network of Phenomena.

1. For any two anomalous window pairs A^{vm} and A^{vn} at time t_i, compute influence distance $d_{S_i^{vm} \rightarrow S_i^{vn}}$ for every spatial location pair combination across A^{vm} and A^{vn}.
2. **for** S_i in A^{vm} **do**
3. **for** S_j in A^{vn} **do**
4. Compute maximum_influence_score for each S_j^{vn}.
5. **end for**
6. **end for**
7. Anomalous window influence = maximum_influence_score/length (A^{vm})
8. **for** S_j in A^{vm} **do**
9. **for** every S_j in A^{vn} **do**
10. Compute best pairing for every spatial object S_j^{vn} given S_i^{vm}.
11. **end for**
12. **end for**
13. **for** $vm, ..., vn$ **do**
14. Identify the spatiotemporal associations by using confidence and support measures.
15. **end for**

We utilize all the anomalous windows (both primary and secondary) and analyze their influence with respect to all the discovered anomalous windows (both primary and secondary), thus considering all the obtained anomalies (considering both primary and secondary). Table 6 explains various quantification measures regarding complete influence relationships, which consider both primary as well as secondary windows.

Thus, the main difference between the two sets of formulae lies in the type of anomalies being considered, which aids in effectively analyzing the influence relationships from two different perspectives, thus helping in efficient reasoning and analysis of spatiotemporal associations across phenomena. These approaches, in turn, also yield different versions of influence scores: influence score considering only highly significant primary anomalies(s_{pa}^v) and influence score considering both primary and secondary anomaly associations (s_{aa}^v) which form an important result for determination of influence relationships across domains.

4.2 Window centers-based approach

Considering the variety and tremendous amount of data generated at a spatial location in our pairwise approach, testing for significant associations across all the phenomena can become costly. Therefore, we propose a variation of the pairwise approach, which utilizes the geographic center of the anomalous windows at each time interval t_i instead of pairwise computations of all locations in the anomalous windows, as is illustrated in Figure 7. We quantify these associations between window centers across domains using the influence scores between window centers across anomalous windows (s_{wc}^v). We also utilize the respective variations of spatiotemporal confidence (STC_d), support (STS_c) and lift (STL_c) for window centers that are similar to 5, 6, and 7 mentioned above to quantify the associations using the window centers-based approach. Algorithm 2 illustrates the overall approach for window center associations.

Table 5 Quantification measures including only primary influence relationships

Quantification measures	Formula	Description
Confidence ($\overline{STC_p}$)	$\overline{STC_p} = \frac{\text{Count of primary pairs}(A^v_{ti}, A^{v'}_{ti}) \mid s^v_{A^v_{ti} \rightarrow A^{v'}_{ti}} \geq s_{min}}{\lvert\text{Count of all pairs from} A^v\rvert}$	Proportion of count of primary pairs which satisfies the influence score threshold criteria with respect to count of all pairs from anomalous windows in phenomena v (i.e. A^v).
Support ($\overline{STS_p}$)	$\overline{STS_p} = \frac{\text{Count of primary pairs}(A^v_{ti}, A^{v'}_{ti}) \mid s^v_{A^v_{ti} \rightarrow A^{v'}_{ti}} \geq s_{min}}{\text{Total number of possible pairs}(A^v_{ti}, A^{v'}_{ti})}$	Proportion of count of primary pairs which satisfy the influence score threshold criteria with respect to total number of all possible pairs between respective anomalous windows across both the phenomena given by A^v and $A^{v'}$.
Lift ($\overline{STL_p}$)	$\overline{STS_p} = \frac{\text{Count of primary pairs}(A^v_{ti}, A^{v'}_{ti}) \mid s^v_{A^v_{ti} \rightarrow A^{v'}_{ti}} \geq s_{min}}{P_{uc}(A^v) \times P_{uc}(A^{v'})}$	Proportion of count of primary pairs which satisfies the influence score threshold criteria with respect to product independent occurrence probabilities for each domain windows given by A^v and $A^{v'}$.

Table 6 Quantification measures including complete relationship

Quantification measures	Formula	Description
Confidence (STC_p)	$STC_p = \frac{\text{Count of all pairs}(A_{ti}^{v}, A_{ti}^{v'}) \mid s^{v}_{A_{ti}^{v} \rightarrow A_{ti}^{v'}} \geq s_{min}}{\lvert\text{Count of all pairs from} A^{v}\rvert}$	Proportion of count of all pairs (both primary and secondary anomalies) which satisfies the influence score threshold criteria with respect to count of all pairs from anomalous window for phenomena v given by A^{v}.
Support (STS_p)	$STS_p = \frac{\text{Count of all pairs}(A_{ti}^{v}, A_{ti}^{v'}) \mid s^{v}_{A_{ti}^{v} \rightarrow A_{ti}^{v'}} \geq s_{min}}{\text{Total number of possible pairs}(A_{ti}^{v}, A_{ti}^{v'})}$	Proportion of count of all pairs (both primary and secondary anomalies) which satisfies the influence score threshold criteria with respect to total number of all possible pairs between anomalous windows across phenomena given by A^{v} and $A^{v'}$.
Lift (STL_p)	$STS_p = \frac{\text{Count of all pairs}(A_{ti}^{v}, A_{ti}^{v'}) \mid s^{v}_{A_{ti}^{v} \rightarrow A_{ti}^{v'}} \geq s_{min}}{P_{uc}(A^{v}) \times P_{uc}(A^{v'})}$	Proportion of count of all pairs (both primary and secondary anomalies) which satisfies the influence score threshold criteria with respect to product of unconditional probabilities of individual domain windows given by A^{v} and $A^{v'}$.

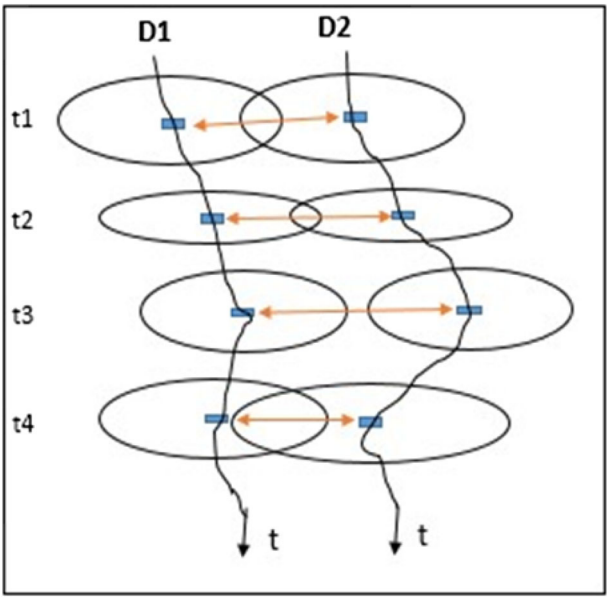

Fig. 7 Window centers influence relationship approach

Algorithm 2 Spatiotemporal association discovery - window center influence relationship approach.

Require: Phenomena for analysis $\{v1, v2, v3, \ldots.., vn\}$, series of anomalous windows A^{v1}, A^{v2}, A^{v3}, A^{v4}, ..., A^{vn} each at time intervals $\{t1, t2, ..., tn\}$.
Ensure: Each anomalous window A^{vm} is a set of spatial locations S_j^{vm} with C_j^{vm} be the window centers given by $C_j^{vm} = \{C_1^{vm}, C_2^{vm}, C_3^{vm}, ..., C_n^{vm}\}$.
Ensure: Associated Network of Phenomena.

1. For any two anomalous window pairs A^{vm} and A^{vn} at time t_i, compute influence distance $d_{S_i^{vm} \rightarrow S_i^{vn}}$ for every anomalous window center combination across A^{vm} and A^{vn}.
2. **for** C_i in C^{vm} **do**
3. **for** C_j in C^{vn} **do**
4. Compute maximum_influence_score for each S_j^{vn}.
5. **end for**
6. **end for**
7. Anomalous window influence = maximum_influence_score/length (A^{vm}).
8. **for** C_j in A^{vm} **do**
9. **for** every C_j in A^{vn} **do**
10. Compute best pairing for every window centers C_j^{vn} given C_i^{vm}.
11. **end for**
12. **end for**
13. **for** vm, ..., vn **do**
14. Identify the spatiotemporal associations by using confidence and support and lift measures.
15. **end for**

Algorithm 3 Spatiotemporal association significance testing

Require: Replica value k, original data points. Discovered associations between phenomena v and v'.
Ensure: p-value
Ensure: Associated Network of Phenomena.

1. **for** 1, ..., k **do**
2. Randomize the network of phenomena links between the original data points.
3. Repeat the association discovery process to measure the current spatiotemporal
4. Association of phenomena v and v' given the same influence score threshold value s_{min}.
5. **if** the confidence of the current association is greater than those of discovered association
6. **then**
7. $p - \text{value} = p - \text{value} + 1/k$
8. **end if**
9. **end for**

The algorithm takes in anomalous window centers for distinct time periods from each phenomenon. It computes the overall influence of one domain over another at a particular time interval ti by using the best pairing between anomalous window centers. Line 1 computes the influence distance across anomalous windows for window center combinations between them. Lines 2 to 6 are used to find the maximum influence score for each window center present in A_{d2} for every location available in A_{d1}. Line 7 computes the overall influence score of A_d1 on A_{d2}. Lines 8 to 12 compute the best available pairing of center locations in A_{d2} given each center in A_{d1}. The goodness of pairing is similarly determined by the total influence score of the pairs. Lines 13 to 15 quantify these discovered associations across domains using spatiotemporal matrices. We quantify these associations between window centers across domains using the influence scores between window centers across anomalous windows (s^v_{wc}). We also utilize the respective variation of spatiotemporal confidence (STC_d), support (STS_c), and lift (STL_c) for window centers that are similar to quantification measures mentioned above in the case of the pairwise influence relationship approach thus resulting in the quantification of associations using the window centers-based approach.

4.3 Spatiotemporal association significance test

Monte Carlo simulation is a widely used evaluation technique to determine the statistical significance of an approach. A Monte Carlo simulation compares the findings from the original data with several randomly generated samples. This process produces p-values to quantify the statistical significance. The lower the p-value, the more significant the finding is. The purpose of the Monte Carlo test is to evaluate the statistical significance of our proposed algorithm in comparison to links that are randomly generated. Under the null hypothesis, the Monte Carlo test is practically beneficial because it does not adhere to the normal distribution, t-distribution, or chi-square distribution. For evaluating if the expectation of samples $\mathcal{D}$ is the sample average μ using a Monte Carlo test, a bootstrap resampling of $\mathcal{D}$ and computing their average is repeated multiple times, and a histogram of the average is constructed. The testing of the hypothesis can be accomplished by determining if the target value μ falls inside the critical region α [34].

More specifically, we developed a permutation-based sampling method by randomizing the links between the network of phenomena. This means, that we generate the new sample in each replica by regenerating the network of phenomena, which links the spatial locations, and finally, re-computing the influence relationships in each replica. This results in randomized locations of anomalous windows for the same phenomena in each sample. The purpose is to

see how often spatiotemporal associations discovered for those phenomena in the original data can be discovered in the randomized samples with similar confidence. Thus, the less often it appears in the sample, the p-value becomes, the more significant the association is. Algorithm 3 explains the complete process of significance testing using Monte Carlo simulations.

5 Experiments and results

We performed detailed experiments on synthetic as well as real world data sets to test the efficacy of our approach.

5.1 Datasets

We consider two multi-domain datasets for our experimental results which are explained in the sections below.

(a) MATCH dataset
In our experiments on a real-world dataset, we study unusual associations between child poverty and unemployment cases in the state of Maryland. We conducted the experiments to check for potential associations between child poverty rates and the unemployment rate for the state of Maryland. "Mobilizing Action Towards Community Health" (MATCH), which is a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute, provides extensive and rich multi-domain health ranking data for counties in the United States of America. We took a subset of this extensive data and analyzed child poverty data against unemployment data for the state of Maryland. The detailed child poverty data contains small area income and poverty statistics for all 24 counties of Maryland and ranges from the year 2010 to 2016 [35]. We aim to analyze the evident potential associations of child poverty with other phenomena, such as unemployment rates. The unemployment rates, which were obtained from the Bureau of Labor Statistics for all 24 counties in the state of Maryland and consists of extensive unemployment statistics like the number of unemployed in the county for the years 2010 to 2016. The list of 24 counties, along with the county code, is provided in Table 7 below.

(b) Synthetic Dataset
We utilized a synthetically generated spatiotemporal dataset with point anomalies for distinct application domains to test our approach. We generated anomalous windows for specific locations in order to test potential associations between domains. The dataset consists of point anomalies for spatial locations and had a time range from 1997 to 2012. The major aim of this data set was to test our approach for multi-domain associations with labeled data and validate our approach findings based on already known associations across domains.

5.2 Experimental results

We performed detailed experiments on the above-discussed datasets to test our approach to finding unusual spatiotemporal associations. We achieved a temporal discretization of the datasets using the available discretization strategies and individually applied our multi-

Table 7 The code of each county utilized in the MATCH dataset

County	Code	County	Code
Allegany County	AL	Harford County	HR
Anne Arundel County	AA	Howard County	HW
Baltimore County	BL	Kent County	KN
Baltimore City	BC	Montgomery County	MG
Calvert County	CV	Prince George's County	PG
Caroline County	CC	Queen Anne's County	QA
Carroll County	CL	Somerset County	SS
Cecil County	CC	St. Mary's County	SM
Charles County	CH	Talbot County	TB
Dorchester County	DR	Washington County	WA
Frederick County	FR	Wicomico County	WC
Garrett County	GR	Worcester County	WR

domain association algorithm to detect anomalous windows in each temporal interval. Finally, we quantify the combined association relationships using influence score, spatiotemporal confidence, support, and lift measures. The experimental results are organized as follows: 1) We first present detailed results for each dataset based on our influence metrics. 2) we discuss the performance of our approach using accuracy, precision, and recall measures. 3) we perform ground truth validation on the obtained results.

5.2.1 Influence score results

We derived anomalous windows for analysis using spatiotemporal scan statistics using Satscan. However, our approach is not limited to using Satscan-based anomalies, and we have tested multiple methods for detecting anomalous windows. We outline our results with Satscan windows due to the wide use and intuitive findings of Satscan-based anomalies. The results were obtained using retrospective scan statistics employing a space-time permutation model over the entire time scan. In the experiments, we set the influential score threshold $S_T = 0.7$, influential decay rate for all phenomena as $\delta = 10$, and the Jaccard similarity coefficient = 0.3.

Results for MATCH dataset Table 9 shows the associated anomalies discovered across both domains using the 4-bin data discretization. We used the multi-domain framework and applied retrospective space-time scan statistics within each bin interval to detect anomalies based on the bin data. For example, we obtain anomalous window {SM, WC, WR, BC, GA, AL, WA, FR, KN, QA, HR, CC, SM, WC, WR, BC, WA} for child poverty and {SM, WC, WR, BC, WA, TB, KN, QA, HR, CC, CH, CL, SM, SM, WC, WR, WA, BC, TB} for unemployment domains respectively for temporal bin ranging from 2010-2011 which contains anomalous counties in the State of Maryland. The definitions for all the above-mentioned abbreviations are provided in Section 5.1.

Later, our approach applies multi-domain association algorithms in each temporal bin and then takes the combined aggregated measure of the discovered common associations across temporal intervals and quantifies these associations using influence relationship metrics. After initial analysis, we found that a significant part of the anomalous windows without temporal

Table 8 Comparison of all levels of discretization for the MATCH dataset

# of bins	s^v_{aa}	s^v_{aa}	s^v_{aa}	p-value
2	0.959	0.5	0.94	0.001
3	0.970	0.88	0.97	0.001
4	0.988	0.97	0.98	0.001

discretization appeared in the 4-bin resultant anomalies (for example, counties like Frederick (FR), Washington (WA), Carroll (CL), and Baltimore City (BC)) (Table 8).

We detected strong influence relationships from all three influence scores - influence score with primary anomalies into consideration(s^v_{pa}), influence score considering both primary and secondary anomaly associations (s^v_{aa}), and influence scores between window centers across anomalous windows (s^v_{wc}) as shown in Table 9. These high influence scores were also quantified by high values of spatiotemporal confidence and lift measures, which indicate a strong influence relationship between the domains. Also, the Monte Carlo simulation results show a strong p-value of 0.001, indicating the statistical significance of these influence relationship results. We also detected significant increases in the influence scores with significant p-values on increasing the number of temporal bin intervals, which imply strong associations with the child poverty and unemployment data as depicted in Table 8.

Results for Synthetic dataset We detected strong influence relationships from all three influence scores - influence score with primary anomalies into consideration(s^v_{pa}), influence score considering both primary and secondary anomaly associations (s^v_{aa}), and influence scores between window centers across anomalous windows (s^v_{wc}) as shown in Table 10. These high values of influence scores were also quantified by high values of spatiotemporal confidence (given by ($\overline{STC_p}$), STC, and STC_w) and lift measures (given by $\overline{STL_p}$), STL, and STL_w) values that indicate comparative results obtained from our approaches explained above. Also, the Monte Carlo simulation results show a strong p-value of 0.001 ($p < 0.05$),

Table 9 Domain anomalies with 4-bin temporal discretization for MATCH dataset

Anomalies from child poverty	Time period range	Anomalies from unemployment	s^v_{aa}	s^v_{pa}	s^v_{wc}
SM, WC, WR, BC, GA, AL, WA, FR, KN, QA, HR, CC, SM, WC, WR, BC, WA	2010-2011	SM, WC, WR, BC, WA, TB, KN, QA, HR, CC, CH, CL, SM, SM, WC, WR, WA, BC, TB			
KN, QA, MG, PG, HW, HR, CC, BL, SM, CV, CH, DH, AL, GA, CL, SM, WC, WR, HW, BC, KN, QA	2012-2013	MG, PG, HW, KN, QA, HR, CC, BL, AL, GR, SM, CL, CH, DR, CR, SM, WC, WR, KN, QA	0.98	0.97	0.98
DH, CL, AL, BC, CV	2014-2015	DL, AL, BL			
SM, WC, WR, CL, AL, BC, AA, CV, CR	2015-2016	CR, AA, CL, AL			
	Child Poverty → Unemployment $\overline{STC_p} = 0.625$; $\overline{STS_p} = 0.010$; $\overline{STL_p} = 0.20$; $STC_p = 0.44$; $STS_p = 0.11$; $STC_w = 0.62$; $STS_w = 0.15$; $STL_w = 23$; $p - \text{value} = 0.001$				

Table 10 Domain anomalies with 4-bin temporal discretization for synthetic dataset

Anomalies from child poverty	Time period range	Anomalies from unemployment	s^v_{aa}	s^v_{pa}	s^v_{wc}
CR, FR, HW, BL, BC, MG, HR, WH, CR, CC, MG, PG, BL	1997-2002	PG, BL, AA, FR, WH, CR, MG, HW, MG			
AL, GR, WH, FR, CR, MG, HW, CL, TB, QA, DR, KN, WC, AA, CV, CC, BC, HR, WR, SS, CL, TB	2003-2005	MG, KN, QA, HR, CC, CL, BC,BL	0.95	0.50	0.93
PG, AL, GR, WH, FR, CR, MG, HW, QA, DR, KN, WC, AA, CV, CC, BC, HR, WR, SS, BL	2006-2008	AA, CV, SM, CH, TB, DR, PG, BC, BL			
GR, AL, WH, FR, MG, CR, HW, PG, CL, TB, QA, DR, KN, WC, AA, CV, CC, BC, HR, WR, SS, BL	2009-2012	AA, MG, PG			
	Child Poverty → Unemployment $\overline{STC_p} = 1$; $\overline{STS_p} = 0.012$; $\overline{STL_p} = 20.5$; $STC_p = 1$; $STS_p = 0.28$; $STC_w = 0.5$; $STS_w = 0.11$; $STL_w = 140$; $p - \text{value} = 0.001$				

which indicates the statistical significance of these influence relationship results. We also detected significant increases in the influence scores with significant p-values on increasing the number of temporal bin intervals, which imply strong associations with the child poverty and unemployment data as depicted in Table 10. We also found a significant part of the common anomalous windows appears in the discovered 4-bin resultant anomalies (for example, counties like Prince George (PG), Baltimore City (BC), and Baltimore County (BL)).

5.2.2 Ground truth and accuracy evaluations

In order to test our approach results, we performed ground truth validation by testing our approach findings against the known anomalies for each of the two datasets. These are explained in detail in subsequent sections below.

MATCH Dataset We utilized the health rankings of the counties as a ground truth validation measure for the respective domains of child poverty and unemployment, which are provided based on the health outcome model proposed on the match website. The known anomalies for the child poverty dataset were Allegany (AL), Worcester (WR), Wicomico (WC), Garrett (GR), Dorchester (DR), Somerset (SS), and Baltimore City (BC). Similarly, the known anomalies for the unemployment data set were Allegany (AL), Worcester (WR), Wicomico (WC), Washington (WA), Garrett (GR), Carroll (CL), Dorchester (DR), Somerset (SS), Kent (KN), and Baltimore City (BC). Thus, the typical multi-domain anomalous counties across both domains were Allegany (AL), Worcester (WR), Wicomico (WC), Garrett (GR), Dorchester (DR), Somerset (SS), and Baltimore City (BC). These we accurately detected by our approach with significant influence. We also compared our method with existing methods to detect spatial outlier associations [10], which can be illustrated in Figure 8 below. Despite

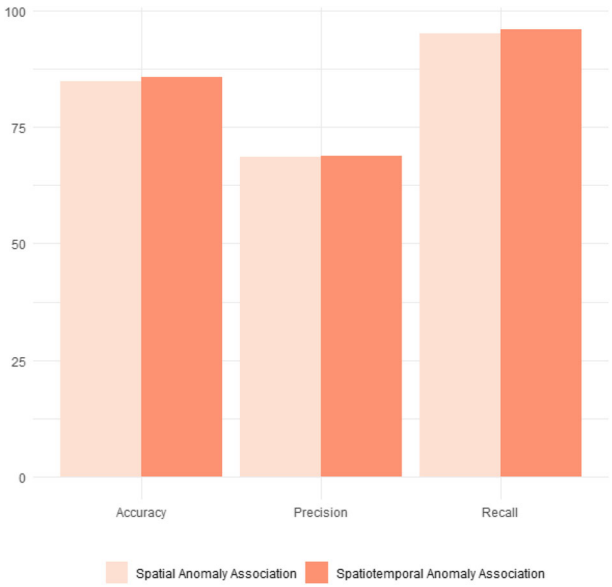

Fig. 8 Performance comparison between the proposed approach and spatial anomaly detection [10]

having roughly similar performance measures as shown in Figure 8, we were able to find two significant achievements. First, as compared to [10], our method could detect significant anomalous window associations across domains and significantly quantify these associations based on a novel metric of influence - influence score. Second, previous work in [10] address the problem of outlier associations in a spatial context, whereas our current method extends this problem to detect spatiotemporal associations.

Synthetic Dataset We test our approach on known pre-defined multi-domain anomalies and check whether the approach is able to detect significant associations across domains. We pre-defined and modeled the known anomalies for the data set at locations Prince George (PG), Baltimore County (BL), and Baltimore City (BC) for Domain 1 and Prince George (PG), Baltimore City (BC), Baltimore County (BL), Montgomery (MG), and Ann Arundel (AA) for Domain 2. Thus, the common multi-domain anomalous counties for Domain 1 and Domain 2 were Prince George (PG), Baltimore County (BL), and Baltimore City (BC). It was observed that our approach was able to detect these known anomalies across all the temporal discretized intervals.

Accuracy Evaluations Considering Locations This approach evaluates accuracy evaluations based on the anomalous window locations detected across the domains. The performance measures for the approach in the form of accuracy, precision, and recall values across datasets are plotted in Figure 9. It was observed that our approach was able to detect anomalies across multi-domain datasets. However, it was observed that, as we increase the number of bins, the accuracy value decreases significantly. This was associated with the increase in false positives generated from the anomalous window discovery process. However, it was observed that our approach was able to detect all the associated anomalies across domains. It is also observed that the same trend for accuracy, precision, and recall measures is observed for both datasets, which justifies the efficacy of our approach. We also observe a

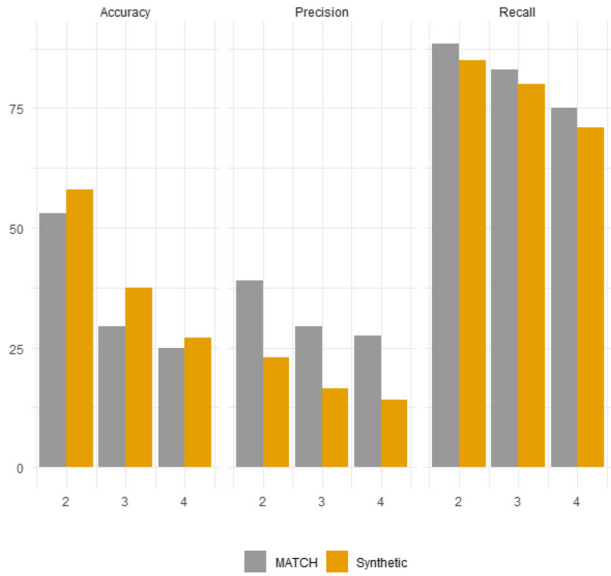

Fig. 9 Performance evaluation across datasets with respect to locations

higher recall value in the performance results, indicating that the model has a greater ability to correctly identify positive instances or true positives, i.e., it is effective at capturing a larger proportion of relevant or positive cases within both datasets.

Accuracy Evaluations Considering Phenomena Linkages Here we consider accuracy evaluations considering the discovery of phenomena linkages which are in the form of common anomalous locations (space-time linkages between the domains) present across both domains. The performance measures for the approach in the form of accuracy, precision, and recall values across datasets are shown in Figure 10. It was observed that our approach was able to detect linkages across multi-domain datasets with significantly high accuracy, precision, and recall values for both the tested datasets. This finding resembles that of the evaluation that took locations into account, but it is clear that there is not much of a trade-off between recall and precision values.

6 System architecture

We developed a complete analysis dashboard application for analyzing these intersecting spatiotemporal associations between anomalies across multiple spatiotemporal datasets to identify interesting phenomena relationships. Figure 11 illustrates the overall system architecture for this dashboard application. The core functionality of this dashboard application constitutes of a set of novel association algorithms developed in *R*. These algorithms utilize data mining and statistical methodologies to find unusual spatiotemporal associations across distinct inter-related domains such as traffic conditions and environmental factors, disease spread epidemic trajectory patterns to name a few.

We first discover the single domain anomalies by interfacing with existing powerful tools for scanning statistics-based anomalies. Subsequently, we find associations between the spa-

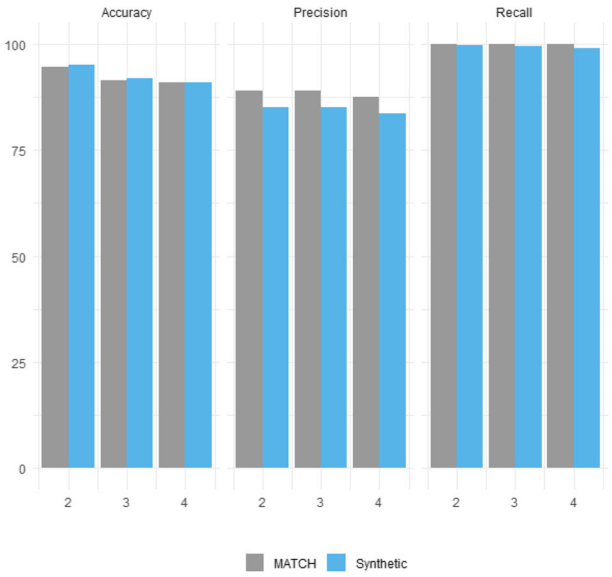

Fig. 10 Performance evaluation across datasets with respect to phenomena locations

tial and spatiotemporal anomalies using a novel metric that is determined based on the spatial and non-spatial overlaps between the anomalies. Each component of the above architecture is explained in detail below, and the overall architecture is shown in Figure 11.

(a) Spatiotemporal datasets
This component accepts the spatiotemporal data from distinct domains as input for analysis. Developed in *R Shiny* package, it provides inputs, which are specifically *.csv* files, which contain the datasets from the individual domains for analysis. We also provide the functionality to format the data required for the single-domain anomaly detection software. For example, *SatScan* input format files generated include a case file, which

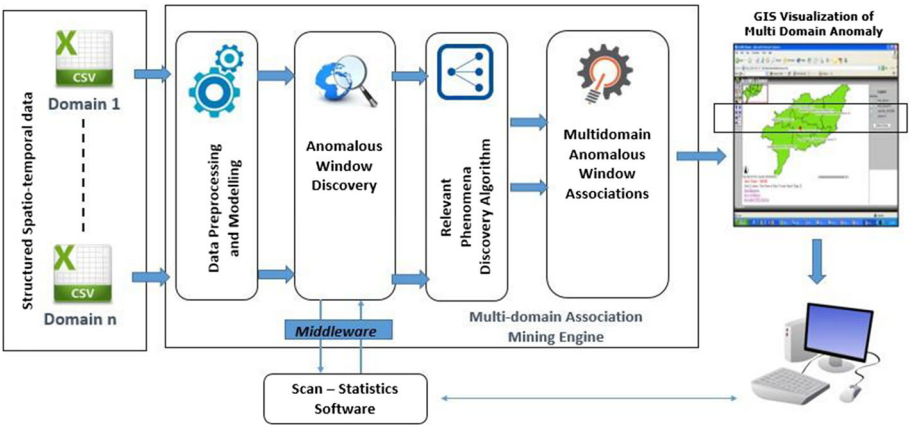

Fig. 11 Overall system architecture for spatiotemporal dashboard applications

contains the information regarding the number of cases by location (attribute of interest in which we are discovering the anomalies) and associated time attribute (date or year), the *geo* file, which contains the location information including the latitude and longitude information of the location.

(b) Multi-domain association engine

This forms the core component of the application, which provides all the necessary application functionality, including data preprocessing and modeling, interface with scan statistics software for anomalous window discovery, and novel association algorithms, which associate the multi-domain anomalies. The Data Preprocessing and Modelling component involves effective pre-processing of the data required to be given as input to scan statistics software. For handling the temporal element of the data, we discretize the data by time and then perform anomaly detection in each of the time bins created. Our application is equipped with three different categories of data discretization techniques which include equal frequency binning, equal width binning, and hierarchical time series

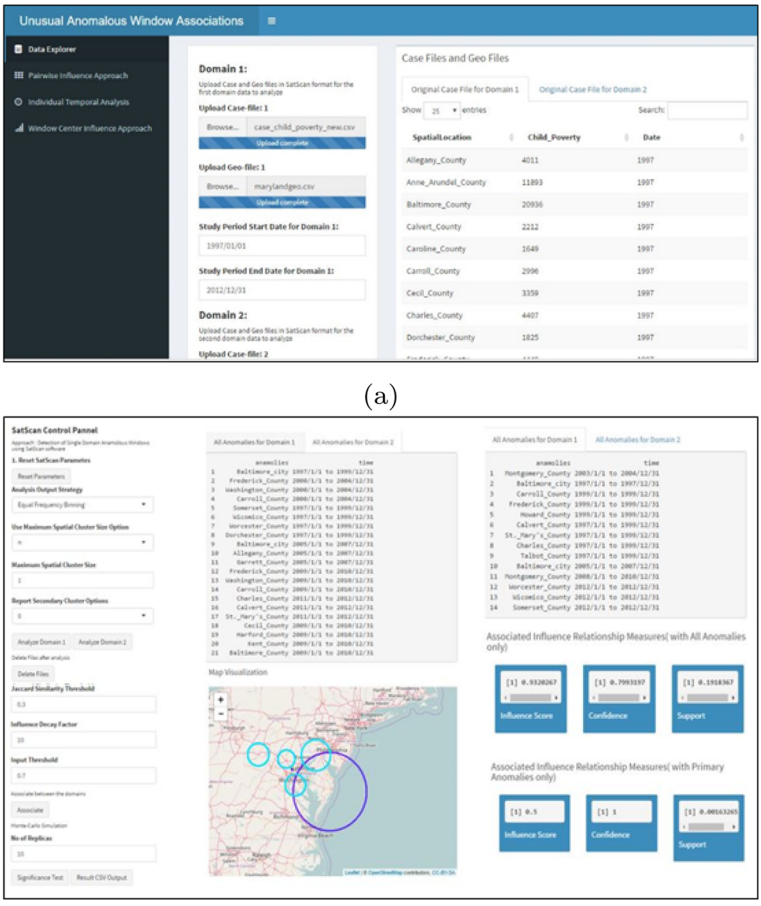

Fig. 12 Spatiotemporal data processing and modeling component

clustering-based binning strategies. The modeling component handles the generation of specific input files from individual discretized instances of datasets, which are ready inputs for associations, such as specific formats required for *SatScan*. A glimpse of this application component is shown in Figure 12a.

The Anomalous Window Discovery component takes in the respective input files modeled from the previous component and processes it using the scan statistics methodologies to obtain single domain anomalies. The *rsatscan* package in *R* acts as a wrapper for interfacing *R* with SatScan software which needs to be installed on the user machine. This package is responsible for interfacing the flow of data as input and generated output between the *R* application and *SatScan* software, which runs on the local machine where analysis needs to be conducted. It also provides multiple action controls that can be used to set the parameters of the *SatScan* processing. The interface with *SatScan* is much more tightly coupled in our current prototype, as it is a widely-used software. The Multi-Domain Anomalous Window Association component discovers the associations between the single-domain anomalies and forms the core component of this application. We identify unusual spatiotemporal associations across distinct domains based on the novel influence relationships algorithms explained in the above sections, which form a significant part of this research. We then utilize novel matrices for the measurement and quantification of spatiotemporal associations, which can thus act as supplementary information for domain experts in order to obtain useful phenomena-centric relationships. This is depicted in Figure 12b.

(c) GIS visualizations

This component provides a mechanism for interactive analysis and visualizations of anomalous windows. The analysis is depicted in the form of visualization of the unusual clusters and association quantification methods, which include support and confidence measures that are adapted for influence the score-based computation of associations. The application also incorporates actual GIS visualization of domain anomalies using

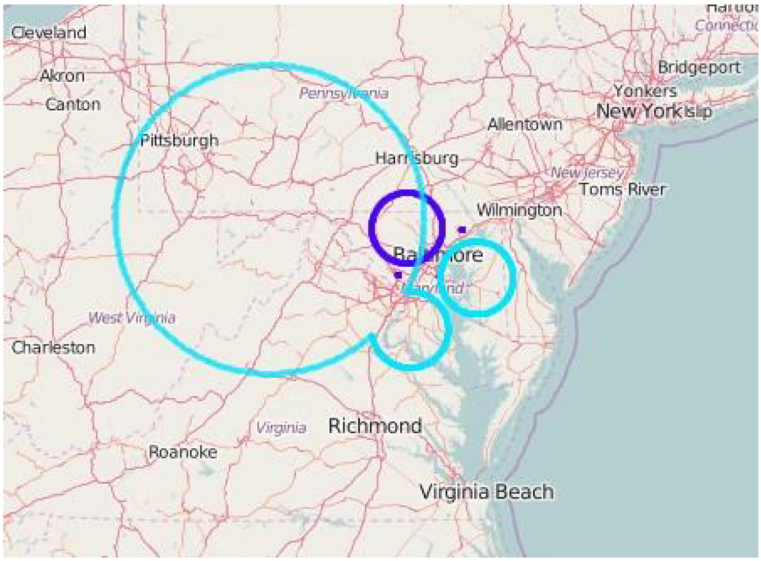

Fig. 13 GIS visualization for anomalous windows

SRgoogleMapsS package and leaflet package providing actual maps to visualize anomalies which can prove effective in analytical insights. Figure 13 shows a sample visualization obtained for the anomalous windows for the child poverty and unemployment dataset for the State of Maryland.

7 Conclusion

In this paper, we proposed a novel approach to detecting and associating unusual spatiotemporal associations across distinct, interrelated application domains. We discovered spatiotemporal associations between phenomena represented by anomalous windows. We proposed a novel measure of influence score to find the influence between these phenomena. In addition, we also quantified these associations using a novel variation of spatiotemporal confidence and support measures. In our future work, we plan to present the associated influence between 'n' possible domains applicable to any possible real-world phenomena using a fast phenomenon discovery algorithm to identify potentially associated phenomena while keeping space as a common reference point. This will utilize the idea of window centers for clustering trajectories and then apply the influence score metrics to phenomena clustered together.

Acknowledgements This work is supported in part by the US Army Corps of Engineers, Engineers Research and Development Center, agreement number: W9132V-15-C-0004 and by the National Science Foundation (iHARP, Award #2118285).

Data Availability Datasets used in this paper are derived from public sources, links to which are provided in the article. The code and sample dataset are available at the Github repository: https://github.com/MultiDataLab/Multi-Domain-Spatiotemporal-Associations

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Econ Geogr 46(sup1):234–240
2. Cools M, Moons E, Wets G (2010) Assessing the impact of weather on traffic intensity. Weather Clim Soc 2(1):60–68
3. Zhang B, Matchinski EJ, Chen B, Ye X, Jing L, Lee K (2019) Marine oil spills-oil pollution, sources and effects. In World seas: an environmental evaluation, p 391-406. Elsevier
4. Xu R, Rahmandad H, Gupta M, DiGennaro C, Ghaffarzadegan N, Amini H, Jalali MS (2021) Weather, air pollution, and SARSCoV-2 transmission: a global analysis. Lancet Planet Health 5(10):e671–e680
5. Mass balance of the antarctic ice sheet from 1992 to 2017 (2018) Nature, 558(7709):219-222
6. Bamber JL, Westaway RM, Marzeion B, Wouters B (2018) The land ice contribution to sea level during the satellite era. Environ Res Lett 13(6):063008
7. Rignot E, Mouginot J, Scheuchl B, Van Den Broeke M, Wessem MJV, Morlighem M (2019) Four decades of antarctic ice sheet mass balance from 1979–2017. Proc Natl Acad Sci 116(4):1095–1103
8. Smith B, Fricker HA, Gardner AS, Medley B, Nilsson J, Paolo FS, Holschuh N, Adusumilli S, Brunt K, Csatho B et al (2020) Pervasive ice sheet mass loss reflects competing ocean and atmosphere processes. Science 368(6496):1239–1242
9. Celik M, Shekhar S, Rogers JP, Shine JA (2008) Mixed-drove spatiotemporal cooccurrence pattern mining. IEEE Trans Knowl Data Eng 20(10):1322–1335

10. Janeja VP, Palanisamy R (2013) Multidomain anomaly detection in spatial datasets. Knowl Inf Syst 36(3):749–788
11. Lee J-G, Han J, Whang K-Y (2007) Trajectory clustering: a partition-and-group framework. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, p 593-604
12. Cressie N (2015) Statistics for spatial data. John Wiley & Sons
13. Schabenberger O, Gotway CA (2017) Statistical methods for spatial data analysis: Texts in statistical science. Chapman and Hall/CRC
14. Chuang A (1991) Time series analysis: univariate and multivariate methods. Taylor & Francis
15. Cao H, Cheung DW, Mamoulis N (2004) Discovering partial periodic patterns in discrete data sequences In Pacific-Asia conference on knowledge discovery and data mining, p 653-658. Springer
16. Huang Y, Shekhar S, Xiong H (2004) Discovering colocation patterns from spatial data sets: a general approach. IEEE Trans Knowl Data Eng 16(12):1472–1485
17. Lee I, Estivill-Castro V (2011) Exploration of massive crime data sets through data mining techniques. Appl Artif Intell 25(5):362–379
18. Estivill-Castro V, Lee I (2001) Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In Proc. of the 6th International Conference on Geocomputation, p 24-26. Citeseer
19. Huang Y, Zhang P (2006) On the relationships between clustering and spatial co-location pattern mining. In 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-06), p 513-522. IEEE
20. Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In International Symposium on Spatial Databases, p 47-66. Springer
21. Tao Y, Kollios G, Considine J, Li F, Papadias D (2004) Spatio-temporal aggregation using sketches. In Proceedings. 20th International Conference on Data Engineering, p 214-225. IEEE
22. Tsoukatos I, Gunopulos D (2001) Efficient mining of spatiotemporal patterns. In International Symposium on Spatial and Temporal Databases, p 425–442. Springer
23. Kulldorff M (1997) A spatial scan statistic. Commun Stat - Theory Methods 26(6):1481–1496
24. Neill DB, Moore AW (2006) Chapter 16 - methods for detecting spatial and spatio-temporal clusters. In: Wagner MM, Moore AW, Aryel RM (eds) Handbook of Biosurveillance. Academic Press, Burlington, pp 243–254
25. Xie Y, Shekhar S, Li Y (2022) Statisticallyrobust clustering techniques for mapping spatial hotspots: A survey. ACM Comput Surv (CSUR) 55(2):1–38
26. Fitzpatrick D, Ni Y, Neill DB (2021) Support vector subset scan for spatial pattern detection. Comput Stat Data Anal 157:107149
27. Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R (2007) Multivariate scan statistics for disease surveillance. Stat Med 26(8):1824–1833
28. Tao Y, Pi D (2008) A neighborhood-based trajectory clustering algorithm. In 2008 Workshop on Power Electronics and Intelligent Transportation System, p 272–275. IEEE
29. Post E, Alley RB, Christensen TR, Macias-Fauria M, Forbes BC, Gooseff MN, Iler A, Kerby JT, Laidre KL, Mann ME et al (2019) The polar regions in a 2 c warmer world. Sci Adv 5(12):eaaw9883
30. Shi L, Janeja VP (2009) Anomalous window discovery through scan statistics for linear intersecting paths (sslip). In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, p 767–776
31. Tango T, Takahashi K (2012) A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. Stat Med 31(30):4207–4218
32. Mohammadi SH, Janeja VP, Gangopadhyay A (2009) Discretized spatio-temporal scan window. In Proceedings of the 2009 SIAM International Conference on Data Mining, p 1197–1208. SIAM
33. Janeja VP, Adam NR, Atluri V, Vaidya J (2010) Spatial neighborhood based anomaly detection in sensor datasets. Data Min Knowl Discov 20(2):221–258
34. Sugiyama M (2016) Introduction to statistical machine learning. Morgan Kaufmann
35. The county health rankings, a key component of the mobilizing action toward community health 1034 (match) project, 2010. http://www.countyhealthrankings.org/. Last Accessed 01-March-2011