International Journal of Data Science and Analytics (2024) 17:275–288
https://doi.org/10.1007/s41060-023-00411-x

REGULAR PAPER

Granger causality-based cluster sequence mining for spatio-temporal causal relation mining

Nat Pavasant[1] · Takashi Morita[2] · Masayuki Numao[2] · Ken-ichi Fukui[2]

Received: 12 December 2022 / Accepted: 15 June 2023 / Published online: 3 July 2023

Abstract
We proposed a method to extract causal relations of spatial clusters from multi-dimensional event sequence data, also known as a spatio-temporal point process. The proposed Granger cluster sequence mining algorithm identifies the pairs of spatial data clusters that have causality over time with each other. It extended the cluster sequence mining algorithm, which utilized a statistical inference technique to identify the occurrence relation, with a causality inference based on the Granger causality. In addition, the proposed method utilizes a false discovery rate procedure to control the significance of the causality. Based on experiments on both synthetic and semi-real data, we confirmed that the algorithm is able to extract the synthetic causal relations from multiple different sets of data, even when disturbed with high level of spatial noise. False discovery rate procedure also helps to increase the accuracy even more under such case and also make the algorithm less-sensitive to the hyperparameters.

Keywords Relation mining · Granger causality · Spatio-temporal relation · Spatio-temporal point process

1 Introduction

Many of the data being collected today are spatio-temporal in nature; for example, meteorological data, earthquakes, crime occurrence, road traffic patterns, epidemic outbreaks, social network posts are being used by real-world organizations on a day-to-day basis [1]. Analysis and understanding of these data are crucial to the applications of the ecology and environmental management, crime analysis, transport route analysis, disease management, precision agriculture, and many more.

In this work, we considered a *point-process spatio-temporal data* [2]. A point process data consist of discrete observations, for example, list of points in a Euclidean space. A point process spatio-temporal data are point process data that have freedom in both temporal and spatial dimensions, for example, a series of Euclidean vectors. Many real-world spatio-temporal data can be represented as a point process, such as earthquake epicenters as a list of latitude and longitude as the spatial part and the occurrence time as the temporal part; or social network posts can be considered as features extracted by natural language processing (NLP) algorithm for the spatial part and the post time as the temporal part. A spatial cluster of these data can represent a meaningful concept, and the causal relationships between these clusters over the time series indicate the mechanism of operation. Thus, our objective is to find two spatial clusters of the data that have a causal relationship with each other.

There were many existing works for identifying causal relationships within purely temporal or spatial data. Granger causality [3], for example, can identify causal relations between time series, or PC algorithm [4] for discrete random variables. When extended to spatio-temporal data, even though several works can identify non-causal relation [1], none can identify causal relations. We believe that with causal relations, we can gain a more thorough understanding of the occurrence mechanism.

Recently, co-occurrence cluster mining (CCM) [5] and cluster sequence mining (CSM) [6] algorithm were proposed for extracting relationships between spatial clusters from the point-process spatio-temporal data, namely a co-occurrence relationship. However, since correlation does not

✉ Nat Pavasant
p-nat@ai.sanken.osaka-u.ac.jp

✉ Ken-ichi Fukui
fukui@ai.sanken.osaka-u.ac.jp

[1] Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

[2] SANKEN (The Institute of Scientific and Industrial Research), Osaka University, Osaka, Japan

imply causation, we cannot conclude that the result from those algorithms is causal relations.

In this work, we proposed the Granger cluster sequence mining (G-CSM) algorithm, which is an extension to the cluster sequence mining (CSM) algorithm. We integrated the Granger Causality [3] method for temporal causality inference. Granger causality is one of the most commonly used temporal causality analysis techniques. It originated from the field of economics, where it is being used to analyze the relationships between different time series. It was also extensively used in the neuro-science applications [7]. The principle of Granger causality is that if **A** causes **B**, then **B** must be easier to predict using all available data than to predict using all available data except **A**. Additionally, we have extended our conference paper [8], a false discovery rate (FDR) procedure was also used to quantify the significance of the detected causality, allowing us to be certain of the statistical significance and to eliminate false-positive result. The proposed algorithm can extract causal relations between spatial clusters within a point process spatio-temporal data according to the causality proposed by Granger.

We validated the performance of our proposed G-CSM algorithm against the original CSM algorithm using a synthetic data. The result showed that the proposed G-CSM algorithm can detect causal relations more accurately and is more robust against noise. We also analyzed the hyperparameters of the G-CSM algorithm and found that the G-CSM algorithm is less sensitive to them, unlike the original CSM which required a careful setting of its hyperparameters. The usage of FDR for statistical testing also increases the accuracy of the algorithm. We also applied the G-CSM algorithm on the semi-real-world data, namely, we used an existing real-world spatial data and the synthetic temporal relationships.

To summarize, the contribution of this paper is as follows:

1. We proposed a causal relation mining model for a spatio-temporal point process data by extending the cluster sequence mining algorithm with Granger causality.
2. We also propose the application of FDR procedure for evaluation of the significance of the causality.
3. The experiments showed that our proposed algorithm especially with the addition of the FDR procedure can extract relations with higher F-score and is less sensitive to the hyperparameters.

2 Literature review

2.1 Spatio-temporal point process

A point process is a framework used for modeling spatial and/or temporal distributions of discrete events. Specifically, the spatial/temporal distributions of an event are represented by the probability of the event occurring in a certain spatial/temporal domain, called the intensity function. The intensity function may be in the form of simple Poisson distributions or more complex distributions [9]. The spatio-temporal point process is a type of point process that jointly models spatial and temporal distributions of events together.

There are many spatio-temporal models that have fixed spatial points; thus, the spatial model can be ignored entirely. The example includes neural activities modeling [10].

For works with arbitrary spatial points, the intensity function of the spatio-temporal point process can be roughly separated into two groups: with or without first-order spatio-temporal separability [2]. This distinction is based on weather or not the time and space are statistically modeled independently.

The models with first-order spatio-temporal separability are usually simpler and have previously been used to model and predict real-world phenomena, such as earthquakes [11, 12], whose intensity function is defined by the epidemic type aftershock-sequence (ETAS) model. Other earthquake prediction models such as [13] use the Hawkes model for temporal modeling and a kernel function of the spatial location for spatial modeling. There is also the work [14], which uses neural network to model the spatial and temporal intensity function.

On the other hand, the assumption of time-space separability usually hinders the accuracy of the model, so there were also many types of research with non-separatable intensity functions. A Marked Recurrent Temporal Point Process model [15] represented the spatial information using a feature vector and use it as one of the input to the temporal model. Some models use Gaussian mixtures as representative points and model the intensity function based on these points [16]. There are also attempts to use deep learning for modeling the intensity function [17] by creating representative points in the spatio-temporal space, assign intensity to these points, and calculating the final intensity as a neural network function over the representative points. Other methods include using reinforcement learning [18].

While adopting the framework of the spatio-temporal point process is useful to study the occurrence mechanism and predicting future events, it does not describe the relations between the events. These existing works focused on modeling the actual point process, which is a mathematical model that captures the occurrence of each data point in the spatio-temporal domain. They are useful for studying the mechanics of each occurrence or for predicting future events, but does nothing about relationships within the data.

Only a few works address the extraction of relations among events in spatio-temporal point process modeling. Higuchi et al. proposed a model [19] with the Gaussian mixtures (the spatial part) and their temporal influences on each other (the temporal part) to discover latent influences between each mixture. Zhu et al. [20] proposed a deep learning model with intensity function for influence between each spatial region of interest. However, both models derive the relationships from the latent feature and thus could not be considered a causality.

2.2 Relation mining

Relation mining is a type of data mining where a relation between each random variable is to be determined [1]. Target relations vary depending on data: similarity, dissimilarity, causality, or co-occurrence, to list a few.

For temporal relation mining (e.g., time-series data), the definition of relation is relatively well-defined, and there were many existing well-researched methods to find relationships, e.g., dynamic time warping [21]. However, when spatial elements are involved, which is regarded as spatio-temporal relation mining, the types of possible relations increase. In the case where no time difference is observed between the associating entities, there are many works, especially in neuroscience. One such work is to identify regions and relations within the human brain from fMRI data [22].

By contrast, the exploration of relation mining when with the element of time difference is extremely limited. Ebert-Uphoff and Deng [23] proposed a framework of constraint-based structure learning of graphical models, modeling the temporal parts as time-lagged variables, to identify the relationship, however as stated in the paper, their introduction of time lag variable violated the probabilistic independability assumption of the constraint-based structure, and moreover, the spatial components are quantized into predefined grids, unlike our proposed method which works directly with the spatial data.

Other methods proposed for determining time-lagged relation included a co-occurrence cluster mining (CCM) [5] and cluster sequence mining (CSM) [6]. Both algorithms can extract relation between a *spatial cluster* in the spatio-temporal point process data. By using a spatial cluster, a similar concept is grouped together. CCM is designed to extract a non-directional occurrence correlation from the spatio-temporal event sequence by trying to first cluster the data spatially and then evaluate the co-occurrence coefficient of each pair of clusters. CSM extends the CCM algorithm by adding a directional requirement (so that one cluster contains a prior event and another the posterior event) and using a probability inference of the time difference. However, both CCM and CSM algorithms were for occurrence correlation, and not causal relations.

2.3 Time-series causality detection

Detection of actual causality from time-series data is a challenging problem. For a causality with the relation on the temporal domain, a well-adopted way of measuring causality is based on the predictability of one event from another, which is known as Granger causality [3].

Definition 1 *Granger Causality* A causes B if it is easier to predict B using all available data than to predict B with all available data except A.

When A is said to cause B according to Granger causality definition, we said that A *granger-cause* (g-cause) B. Since in real data, no one knows the true causality, henceforth we will use the word causality to refer to Granger causality. Granger causality has been used especially intensively in the neuroscience field [24].

There were another similar method called transfer entropy [25]; however, it reduced to the same model for vector autoregressive model [26]. The transfer entropy was also further developed to Causation entropy [27]. Convergent Cross-Mapping [28] was a similar approach based on predictability, but use attractor manifolds to model the history and check whether such model can predict the target time series. These model was designed to overcome one of the main limitation of Granger causality: it does not perform well when the causation graph is complex.

Meanwhile, there was the adaption of Granger causality to work with temporal point process data. Kim et al. [29] applied the Granger causality to a spike train of neuron activation to analyze causal relationships between brain neural activities. This was further extended by Casile et al. [30] by also taking background activities into account during causality estimation. We have chosen to use Granger causality in this work for it is widely used especially in neuroscience and its simplicity.

2.4 Spatio-temporal clustering

Spatio-temporal clustering [31] is a process of clustering a spatio-temporal data. They are difference from normal spatial clustering algorithms in a sense that there are needs to handle the temporal parts, either by also considering temporal proximity such as temporal closeness, movement behavior/trajectory, or parameter changes over time.

On the other hand, out work are trying to detect causality over time between events within the *two* spatial clusters. We do not require temporal similarity within each cluster; hence, regular spatio-temporal clustering does not fit our objective.

3 Methodology

3.1 Granger cluster sequence pattern

The proposed Granger cluster sequence mining (G-CSM) algorithm is based on the original CSM algorithm. It takes an input of a multi-dimensional event sequence and outputs a list of *cluster sequence patterns*. Each event is an n-dimensional vector, which is considered to be a point in n-dimensional spatial space.

Definition 2 A *Multi-dimensional Event Sequence* is a sequence of length N of n-dimensional vectors of real numbers representing *events*, each with an associated *timestamp*, ordered sequentially:

$$X = \{\mathbf{x}^{(i)} \in \mathbb{R}^n\} \ (|X| = N) \tag{1}$$

$$(X) = \langle t(\mathbf{x}^{(1)}), t(\mathbf{x}^{(2)}), \ldots, t(\mathbf{x}^{(N)})\rangle \quad \text{s.t. } t(\mathbf{x}^{(1)}) \leq t(\mathbf{x}^{(2)}) \leq \cdots \leq t(\mathbf{x}^{(N)}). \tag{2}$$

The objective of the G-CSM algorithm is to identify a set of *prior* event clusters, along with the respective *posterior* event clusters, that fit the following conditions:

1. **Causality** The occurrence of the prior event **A** must have a causal relation with the posterior event **B** according to a causality measure.
2. **Frequency** The more number of pairs of prior events $\mathbf{x}^{(a)} \in \mathbf{A}$ with respective posterior events $\mathbf{x}^{(b)} \in \mathbf{B}$, the better the cluster sequence. The number of pairs of events in the cluster sequence pattern must be larger than some hyperparameter $Supp_{\min}$.
3. **Spatial proximity** The variance of the event within each cluster **A** or **B** must be low. This was evaluated using SSW (sum of square within) measure. **A** and **B** were evaluated independently.

Conditions (2) and (3) are the same as the original CSM algorithm. Thus, *cluster sequence pattern* can be defined by the following:

Definition 3 A *Cluster Sequence Pattern* is a pair of spatial clusters of the event sequence, called *prior* cluster and *posterior* cluster, that satisfy the three conditions outlined above.

$$S_{A\to B} = \langle \mathbf{A} = \{\mathbf{x}^{(i)} | A^i = 1\}, \mathbf{B} = \{\mathbf{x}^{(i)} | B^i = 1\}\rangle, \quad (\mathbf{A} \cap \mathbf{B} = \emptyset), \text{ and must satisfy the 3 conditions,} \tag{3}$$

where A and B are an assignment vector for the set **A** and **B**, respectively. The set **A** is a *prior cluster*, while the set **B** is a *posterior cluster*.

3.2 Algorithm overview

To extract the cluster sequence patterns from the input event sequence, this paper substantially modified the cluster sequence mining (CSM) algorithm [6]. Specifically, the time proximity evaluation of the original CSM was replaced with a Granger causality-based measure. Pairwise point-process Granger causality was adapted from [29] to determine the Granger causality between the prior and posterior event cluster.

The original CSM algorithm works in 3 steps, as shown in Fig. 1.

1. **Candidates Generation**: Potential pattern candidates were generated from all pairs of clusters as identified by using the agglomerative hierarchical clustering (AHC) algorithm over the spatial space X. Only pairs of clusters that contained at least $Supp_{\min}$ (the pattern minimum support) pairs of correspondence events were considered.
2. **Evaluation**: The candidates were evaluated using both the temporal and spatial proximity measures.
3. **Elimination of Inclusive Relation**: Eliminate patterns that have inclusion relation with other patterns, keeping only the pattern with the best score.

The evaluation of the original CSM was based on spatial proximity (using SSW measure) and temporal proximity (based on the likelihood function of the time interval). The final evaluation must be higher than $\mathcal{L}_{\min}$ (a minimum sequence threshold) to be considered for the final output.

In this work, the temporal proximity is modified to use the Granger causality measure instead of the original Bayesian inference method. The process is detailed in the next section.

3.3 Pairwise point-process Granger causality

Traditionally, the Granger causality works on time series or spectral data (data that is in the form of frequency spectrum). We adopted a temporal point process Granger causality [29] to evaluate causality of pattern candidates of cluster pair. The main difference from the original work is that we simplified the algorithm to work with pairwise event sequence only.

Basically, a Granger causality of the cluster sequence pattern $S_{A\to B}$ is whether **A** *g-causes* **B** or not. A cumulative incidence function (CIF) for the point process of occurrence of event **B** can be defined as:

$$\lambda_b(t|H_b(t)) = \lim_{\Delta\to 0} \frac{Pr[(N_b(t+\Delta) - N_b(t)) = 1]}{\Delta}, \tag{4}$$

where $N_b(t)$ is a counting measure of event b within the time of $(0, t]$, and $H_b(t)$ is an occurrence history of all event occurrences up to time t for event b. The probability of the

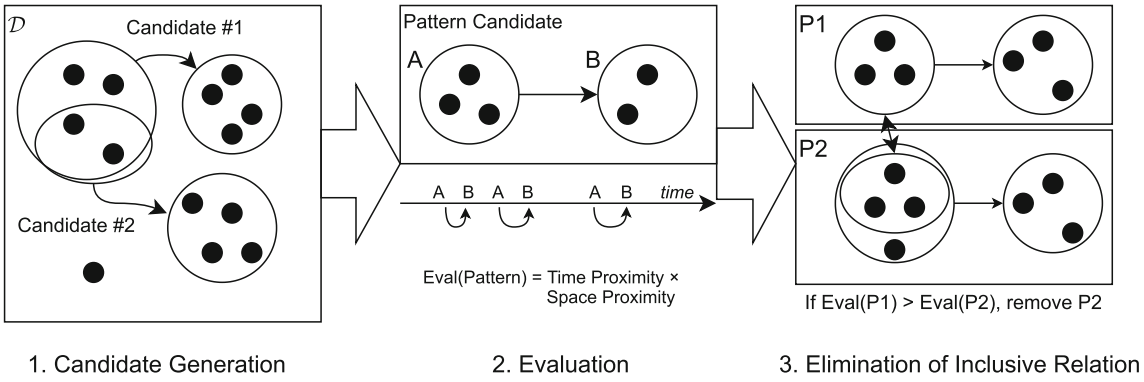


Fig. 1 The overview of cluster sequence mining algorithm [6]

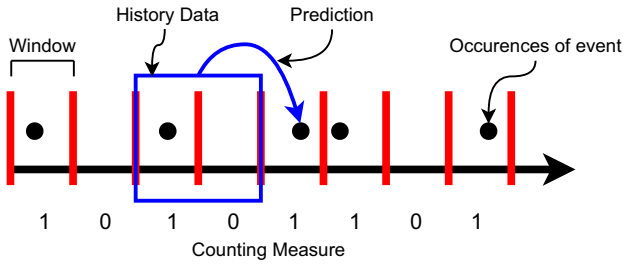


Fig. 2 Simplified view of the GLM model for point-process Granger causality

event b occurring in a small time window $[t, t+\Delta)$ can be written as $\lambda_b(t|H_b(t))\Delta$.

Since it is not feasible to consider the entire history, only the history from time $[t - M_b W, t]$ was considered. The time range was divided into M_b equal windows of length W. The number of occurrences of event q in the time window $[t-mW, t-(m-1)W]$ is denoted as $R_{q,m}(t)$ for $q \in \{\mathbf{A}, \mathbf{B}\}$, a prior or posterior event. $m = 1, \ldots, M_b$ is the window number.

To model the predictor for the event occurrence, a generalized linear model (GLM) framework was used to model the CIF. In GLM, the logarithm of the CIF was modeled using a linear combination of the occurrence history. We currently assume linearity of the data by setting a small window size; however, extension to nonlinear model is currently a future work. The simplified model is shown in Fig. 2. In this case, the log-CIF is modeled as:

$$\log \lambda_b(t|\theta_b, H_b(t)) = \theta_{b,0} + \sum_{q\in\{\mathbf{A},\mathbf{B}\}} \sum_{m=1}^{M_b} \theta_{b,q,m} R_{q,m}(t), \quad (5)$$

where $\theta_{b,0}$ is a background activity, and $\theta_{b,q,m}$ is the effect of $R_{q,m}(t)$ to the event b. The parameter vector, θ_b, is defined as:

$$\theta_b = \{\theta_{b,0}, \theta_{b,a,1}, \ldots, \theta_{b,a,M_b}, \theta_{b,b,1}, \ldots, \theta_{b,b,M_b}\} \quad (6)$$

A point process likelihood function [10] was used to fit the GLM model. As [10] shows that for the point process, both the Binomial and Poisson estimation in GLM are equivalent; this work chose the former.

To make the calculation easier, the entire timeline $[0, T]$ was divided into K equal non-overlapping windows, each with the length of W. The time window k would represent the time window $(t_{k-1} = (k-1)W, t_k = kW]$. To represent this discretized time, the history $H_b(t)$ is written as $H_b[k]$, and $R_{q,m}(t)$ is written as $R_{q,m}[k]$. $\Delta N_b[k] = N_b[k] - N_b[k-1]$ is the number of event occurrences within the time $(t_{k-1}, t_k]$, and the CIF in Eq. (4) is written as $\lambda_b(t_k|\theta_b, H_b[k])$. W should be chosen to be very small so that $\Delta N_b[k]$ can only be either 0 or 1.

Thus, the likelihood function for the binomial GLM model is given as:

$$L_b(\theta_b) = \prod_{k=1}^{K} [\lambda_b(t|\theta_b, H_b[k])\Delta]^{\Delta N_b[k]} [1 - \lambda_b(t|\theta_b, H_b[k])\Delta]^{1-\Delta N_b[k]}. \quad (7)$$

As per the Granger causality in Definition 1, event **A** is considered to *Granger-cause* event **B** if there was a reduction in the likelihood of predicting the occurrence using the history of only **B** instead of using the history of both **A** and **B**. The log-likelihood ratio, $\Gamma(S_{A\to B})$, is defined as:

$$\Gamma(S_{A\to B}) = \log \frac{L_b(\theta_b^a)}{L_b(\theta_b)}, \quad (8)$$

where the likelihood $L_b(\theta_b)$ was obtained from model fitting Eq. (5), and the likelihood $L_b(\theta_b^a)$ was obtained using new CIF with history of **A** cut:

$$\log \lambda_b^a(t|\theta_b^a, H_b^a(t)) = \theta_{b,0}^a + \sum_{m=1}^{M_b} \theta_{b,b,m}^a R_{b,m}(t). \quad (9)$$

The log-likelihood ratio $\Gamma(S_{A\to B})$ in Eq. (8) is considered to be a Granger causality strength for pattern $S_{A\to B}$.

3.4 Significant testing using false discovery rate

The Granger causality strength from Eq. (8) cannot tell us whether the relation is significant enough to be considered a causality. Thus, the following null and alternative hypotheses were formed:

$$H_0 : \theta' = \theta_b^a \quad \text{(the limited predictor is better)}, \tag{10}$$
$$H_1 : \theta' = \theta_b \quad \text{(the full predictor is better)}. \tag{11}$$

To test H_0 against H_1, we use a likelihood-ratio test [29, 32]. The likelihood-ratio test evaluated the difference of deviance ΔD flowing the χ^2 distribution, which is given by:

$$\Delta D = -2\Gamma(S_{A\to B}) \sim \chi^2_w, \tag{12}$$

where w is the degree of freedom; in this case, the difference in dimensionality of the two predictors is equal to the history length of Granger causality M_b.

Because we were performing tens of thousands of significant testings over the course of the algorithm, we also need a way of controlling the type-I error rate. We utilized the false discovery rate (FDR) procedure [33], specifically the Benjamini–Hochberg procedure, which can be summed up as:

1. Perform all significant testings and calculate the p-values.
2. Rank all p-values from low to high. So that $p_1 \le p_2 \le p_3 \le \cdots \le p_n$
3. Find maximum k such that $p_k \le \frac{k}{n}\alpha$, where α is the acceptable ratio of type-I error.
4. Accept first k testings.

In our algorithm, we applied FDR over all cluster sequence pattern candidates. The p value of each candidate was calculated according to Eq. (12). We calculated the threshold Γ_0 for the Granger causality strength such that:

$$P(\Gamma(S_{A\to B}) \ge \Gamma_0) = p_k, \tag{13}$$

where p_k is the highest p-value accepted by the FDR algorithm above, meaning Γ_0 is the likelihood ratio of the kth candidate pattern, sorted by the likelihood ratio.

3.5 Evaluation using Granger causality

The evaluation is separated into two parts: temporal and spatial evaluation. The temporal evaluation is based on the strength and significance of the Granger causality of the sequence. We proposed two different methods of temporal evaluation: threshold strength and scaled strength.

1. **Threshold strength**. Use the significant threshold as a cutoff, resulting in the sequence that is deemed to have significant causality to be evaluated using spatial features only.

$$\mathcal{F}_{TH}(S_{A\to B}) = \begin{cases} 0 & (\Gamma(S_{A\to B}) < \Gamma_0) \\ 1 & (\Gamma(S_{A\to B}) \ge \Gamma_0) \end{cases}. \tag{14}$$

2. **Scaled strength**. The strength is scaled from 0 to 1

$$\mathcal{F}_{SC}(S_{A\to B}) = \max\left(0, 1 + \frac{\Gamma_0}{\Gamma(S_{A\to B})}\right). \tag{15}$$

The final evaluation score, $\mathcal{L}$, in Eq. (17) use the same formula as the original CSM (with original temporal evaluation $\mathcal{F}$ replaced by proposed Granger causality-based $\mathcal{F}$). $\mathcal{G}$, in Eq. (16), is a spatial evaluation using SSW, also the same as the original CSM.

$$\mathcal{G}(\mathbf{A}, \mathbf{B}) = \exp\left(-\frac{\mathrm{SSW}(\mathbf{A})^2 + \mathrm{SSW}(\mathbf{B})^2}{2\sigma^2}\right), \tag{16}$$

$$\mathcal{L}(S_{A\to B}) = \mathcal{F}_i(S_{A\to B})^\gamma \cdot \mathcal{G}(\mathbf{A}, \mathbf{B})^{(1-\gamma)}, i \in \{TH, SC\}. \tag{17}$$

The candidates were ranked and eliminated in the same manner as in the original CSM as described in Sect. 3.2. The entire G-CSM algorithm is also detailed in Algorithm 1. PatternInclusive($S_{A\to B}$, $S_{C\to D}$) is *true* if $(A \subset C \vee C \subset A) \wedge (B \subset D \vee D \subset B)$, and $\|X\|$ is the cardinality of set X.

4 Experiments

To validate our algorithm, we performed multiple experiments with synthetic data and semi-real data. First, we compared the performance between our proposed G-CSM algorithm with and without FDR, and the original CSM algorithm. We also tested against various patterns of synthetic data. Second, we analyzed the hyperparameters of our proposed algorithm. Lastly, we tested the algorithm against semi-real data. To the best of our knowledge, no other method can be compared other than CSM as discussed in Sect. 2.

4.1 Data generation

In many of our experiments, synthetic data with embedded true relation and noise were used. The synthetic data contain an *embedded relation*, which is a pair of spatial clusters

Algorithm 1 G-CSM algorithm

Input List of event X and timestamps X_t
Output List of cluster sequence patterns

```
# Step 1: Candidate Generation
Perform AHC on X
C ← ∅                                              ▷ Set of candidates
for all A ∈ AHC(X) do
    for all B ∈ AHC(X) do
        if A = B then
            Continue.
        end if
        T ← min(‖A‖, ‖B‖)
        if T ≥ Supp_min then
            Append ⟨A, B⟩ to C
        end if
    end for
end for
# Step 2: Evaluation
D ← ∅                                  ▷ Set of preliminary sequence patterns
for S_{A→B} ∈ C do
    L = ℒ(S_{A→B})                                 ▷ Eq. (17)
    if L ≥ ℒ_min then
        Append S_{A→B} to D
    end if
end for
# Step 3: Elimination of Inclusive Relation
Sort D by evaluation score, high to low.
for i from 1 to ‖D‖ do
    for j from i + 1 to ‖D‖ do
        if PatternInclusive(D[i], D[j]) then
            Remove D[j] from D
        end if
    end for
end for
return D
```

that has the time interval between the corresponding event in prior and posterior event clusters following an exponential distribution. The synthetic data also have noise added. This process is similar to the synthetic data used in the CSM paper [6].

The embedded relation is generated as:

1. Generate N data points from a normal distribution for two clusters: $\mathbf{x^{(i)}} \in X \sim \mathcal{N}(m_A, \Sigma_A)$ and $\mathbf{y^{(i)}} \in Y \sim \mathcal{N}(m_B, \Sigma_B)$.
2. For each pair of $\langle \mathbf{x^{(i)}}, \mathbf{y^{(i)}} \rangle$, generate a $t^{(i)} \sim \text{Exp}(\lambda)$.
3. Set $t_{Gap} = ((t_1 - t_0) - \sum t^{(i)})/(N-1)$, the gap between each event. The input parameter t_0, t_1, and λ should be set such that $t^{(i)} \ll t_{Gap}$.
4. Each pair of $\langle \mathbf{x^{(i)}}, \mathbf{y^{(i)}} \rangle$ are allocated a timestamp such that $t(\mathbf{x^{(i)}}) = t_0 + (i-1)t_{Gap} + \sum_{j=0}^{i-1} t^{(j)}$ and $t(\mathbf{y^{(i)}}) = t_0 + (i-1)t_{Gap} + \sum_{j=0}^{i} t^{(j)}$.

The noise is generated as a uniform spatial noise uniformly over the timeline:

1. Generate N data points from a uniform distribution: $\mathbf{x^{(i)}} \sim \mathcal{U}[\mathbf{a}, \mathbf{b}]$.
2. For an event $\mathbf{x^{(i)}}$, set $t(\mathbf{x^{(i)}}) \sim \mathcal{U}[t_0, t_1]$

In this experiment, the embedded relation was generated using the parameters $N = 300$, $m_A = (-2, 0)$, $\Sigma_A = (0.5, 0.5)$, $m_B = (2, 0)$, $\Sigma_B = (0.5, 0.5)$, $t_0 = 0$, and $t_1 = 100,000$. Noises were generated using the parameters: $\mathbf{a} = (-4, -4)$, $\mathbf{b} = (4, 4)$, $t_0 = 0$, $t_1 = 100,000$. This created a single cluster sequence pattern, with noise directly over the event cluster, to test the basic accuracy of the algorithm. The number of noise, N_{noise}, is varied by each experiment.

The example spatial view of the data with $N_{noise} = 1000$ and the histogram of interval between each relation with $\lambda = 2$ (average interval = 0.5) are shown in Fig. 3.

4.2 Evaluation measure

We measured the precision, recall, and F-score of the identified prior and posterior clusters, and of the relation itself. The equation for these scores is as follows:

$$Prec(C) = \frac{\|C \cap X\|}{\|C\|}, \tag{18}$$

$$Rec(C) = \frac{\|C \cap X\|}{N}, \tag{19}$$

where $Prec(C)$ and $Rec(C)$ are the precision and recall score of the cluster C given the ground-truth cluster X. $\|\cdot\|$ denoted the cardinal of the event set. F-scores were calculated as the harmonic mean between the precision and recall score.

We also define relation-based precision and recall as follows:

$$Prec(S_{A\to B}) = \frac{\|\{i | \mathbf{x}^{(i)} \in A \wedge \mathbf{y}^{(i)} \in B\}\|}{0.5 \times (\|A\| + \|B\|)}, \tag{20}$$

$$Rec(S_{A\to B}) = \frac{\|\{i | \mathbf{x}^{(i)} \in A \wedge \mathbf{y}^{(i)} \in B\}\|}{N}, \tag{21}$$

where $\mathbf{x}^{(i)} \in X$ and $\mathbf{y}^{(i)} \in Y$ are ground-truth prior and posterior event cluster, with $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ being the associated event pair as generated in Sect. 4.1, and N is the number of event pairs in the ground-truth relations. The precision and recall score of each relation was calculated using the number of pairs of events that were actually related to each other in that relation.

We also counted how many relations were outputted. The number of relations in our generated data is one. However, the algorithm may output more than one relation, whether because it detected more than one relation in the input, or because multiple subsets of the same relation were detected. In such cases, the prior clusters of all relations were merged,

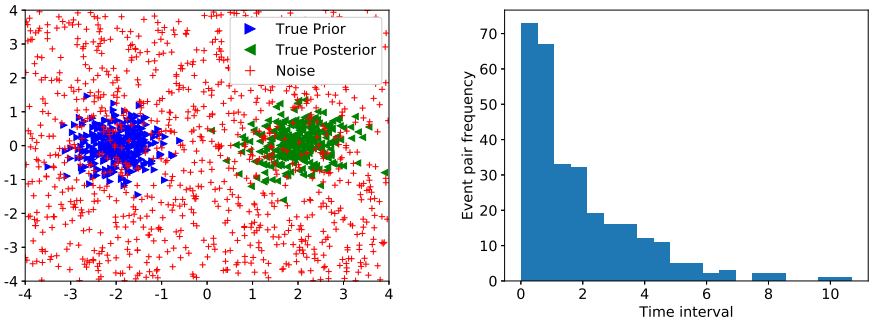


Fig. 3 Example of the input data

(a) Spatial view of data with $N_{noise} = 1000$ (b) Histogram of the intervals between relation, $\lambda = 0.5$

Table 1 AIC value at different history lengths

Noise	History 2	3	5	8	15
100	2928	2856	2781	**2760**	2771
500	3333	3269	3206	**3191**	3202
1000	3722	3675	3632	**3625**	3637
2000	4644	4621	**4603**	**4603**	4616
3000	5344	5333	**5326**	5330	5344

Bold is the lowest

and the posterior clusters were also similarly merged for the purpose of evaluation only.

4.3 Performance validation

In this section, we compared the accuracy of the original CSM algorithm with our proposed G-CSM algorithm with and without FDR.

4.3.1 Parameter settings

The hyperparameters were set with $\sigma = 0.5$, $\gamma = 0.5$, $\mathcal{L}_{\min} = 0.8$. These parameters were set to balance the effect of the spatial and temporal scores.

For the Granger causality, we set the window size $W = 1$. The history length, M_b, was set using Akaike information criteria (AIC) [34] on the Granger causality model according to Table 1. α for significant testing is set to 0.05, which means we accepted 5% error rate for causality detection.

4.3.2 Result with varying noise level

First, we tested the algorithms with various amount of noise (N_{noise}) and $\lambda = 0.5$. The G-CSM using $\mathcal{F}_{TH}$ (Eq. (14)) is denoted with **FDR-TH**, while the one using $\mathcal{F}_{SC}$ (Eq. (15)) is denoted with **FDR-SC**. The result is shown in Table 2. All results were an average of 20 runs.

The example of relations extracted by each algorithm is shown in Fig. 4. The G-CSM without FDR and G-CSM with FDR-SC result are similar, but G-CSM with FDR-SC result has a slightly bigger posterior cluster, which better matches the generated data. The G-CSM with FDR-TH has a lower precision cluster and sometimes identified an erroneous relation as shown.

Meanwhile, the original CSM failed to extract any relations at all even with noise = 100. We believe the main reason is that the original CSM tried to match each event pair together, so by having noise without matching pair in the spatial cluster, it failed to detect any meaningful relations. In contrast, G-CSM is using window-based event counting in the time-space; therefore, G-CSM is more robust against noise in the time domain than CSM.

In Table 2, the FDR-TH algorithm is also uniformly bad, having worse results than the G-CSM without FDR and FDR-SC in almost all measurements. We believe that because the temporal score becomes either 0 or 1, the score relies entirely on the spatial evaluation. The spatial evaluation prefers compact clusters; thus, FDR-TH has high precision but low recall. FDR-TH also has a tendency to detect erroneous relations, especially at a higher noise level.

G-CSM with FDR-SC has better recall score than G-CSM without FDR, especially as the noise increased. This result shows the robustness of G-CSM with FDR-SC against spatial noise. G-CSM with FDR-SC also maintains the relational precision score better than G-CSM without FDR.

Note that for all algorithms, the number of extracted relation (Cnt. in Table 2) also increased along with the noise. This is because the algorithm finds multiple relations that are a subset of the ground-truth relation, resulting in more coverage of the ground-truth data. In turn, this increases the recall score.

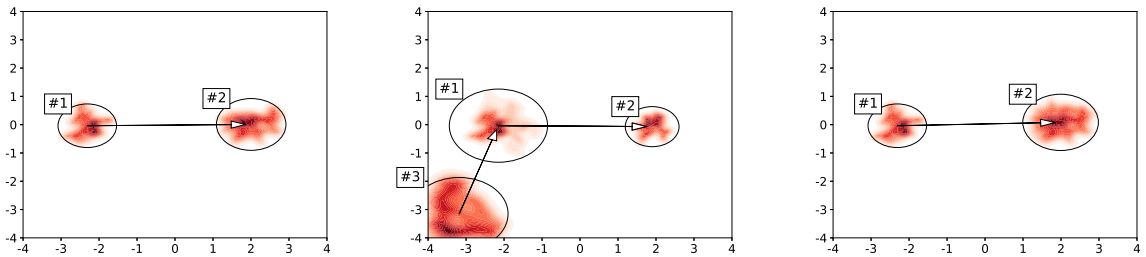
4.3.3 Analysis of the spatial and temporal score

We plotted the temporal evaluation score against the spatial evaluation score of all candidates as a scatter plot in Fig. 5. We

Table 2 CSM and G-CSM at various noise levels

Algo.	P (1)	R (1)	F (1)	P (2)	R (2)	F (2)	P (R)	R (R)	F (R)	Cnt.
Noise = 100										
G-CSM	0.987	**0.540**	**0.693**	0.987	**0.548**	**0.701**	**0.534**	**0.297**	**0.381**	**1.050**
FDR-TH	0.987	0.355	0.512	**0.989**	0.327	0.485	0.324	0.116	0.169	1.250
FDR-SC	**0.989**	0.495	0.654	0.988	0.492	0.649	0.473	0.242	0.319	**1.050**
Noise = 500										
G-CSM	0.940	**0.556**	**0.695**	0.936	**0.537**	**0.678**	**0.510**	**0.301**	**0.377**	**1.100**
FDR-TH	0.921	0.389	0.507	0.912	0.362	0.504	0.318	0.143	0.187	1.450
FDR-SC	**0.941**	0.536	0.678	**0.937**	0.513	0.658	0.490	0.277	0.352	**1.100**
Noise = 1000										
G-CSM	**0.885**	0.551	0.674	**0.888**	0.549	0.673	0.484	0.305	0.373	**1.100**
FDR-TH	0.771	0.428	0.519	0.772	0.416	0.503	0.318	0.192	0.227	1.900
FDR-SC	0.883	**0.560**	**0.681**	0.887	**0.555**	**0.678**	**0.491**	**0.314**	**0.382**	**1.100**
Noise = 2000										
G-CSM	0.783	0.550	0.637	**0.791**	0.571	0.655	0.436	0.320	0.365	**1.300**
FDR-TH	0.662	0.476	0.516	0.675	0.473	0.522	0.309	0.245	0.257	2.450
FDR-SC	**0.784**	**0.560**	**0.643**	0.788	**0.604**	**0.679**	**0.449**	**0.341**	**0.385**	**1.300**
Noise = 3000										
G-CSM	0.679	0.595	0.623	**0.705**	0.609	0.643	0.407	0.366	0.381	1.850
FDR-TH	0.500	0.497	0.467	0.530	**0.634**	0.542	0.272	0.322	0.281	4.050
FDR-SC	**0.683**	**0.611**	**0.635**	0.701	0.631	**0.657**	**0.423**	**0.389**	**0.402**	**1.700**

P = Precision, R = Recall, F = F-score, (1) = Prior cluster, (2) = Posterior cluster, (R) = Relation, Cnt. = Number of relations identified. Bold indicated the best result



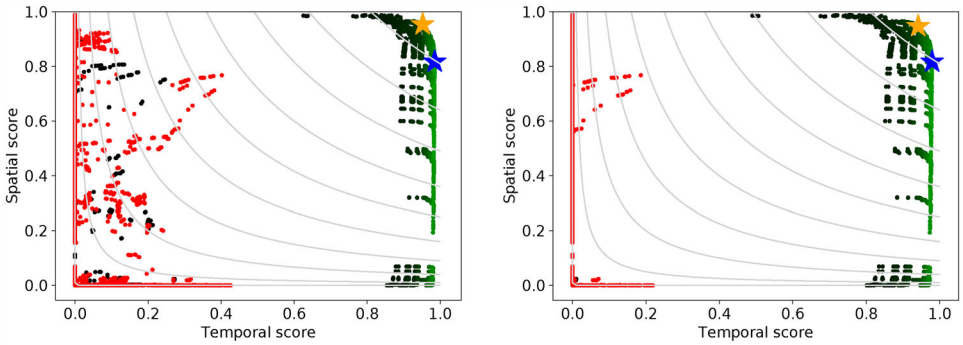
(a) G-CSM w/o FDR (b) G-CSM w/ FDR-TH (c) G-CSM w/ FDR-SC

Fig. 4 Example of relation(s) extracted by each algorithm with noise = 3000. CSM cannot identify any relation

define *possibly correct relation* as a relation that the both prior and posterior event clusters is a subset of the ground-truth prior/posterior event clusters, excluding noise. The black and green dots indicated possibly correct relation, shaded by the F-score of the relation from black (low F-score) to green (high F-score). The red dots indicated wrong relations. The scores of the orange mark, which is the relation with the highest evaluation score, are shown in Table 3.

Figure 5 indicates that FDR-SC algorithm helped to clean up wrong relations and possibly correct relations with low F-scores, while still keeping the possibly correct relations with high F-scores intact. This results in a wider range of the temporal score over the pattern candidates. As shown in Fig. 5, the FDR-SC version has a usable range from around 0.6 to 1.0, while the without FDR is around 0.7 to 1.0. This can also be seen in Table 3, as the relation with the highest evaluation score also has higher recall and F-score for the G-CSM with FDR-SC compared to G-CSM without FDR.

Therefore, the spatial score ended up having less effect on the final evaluation score. With the noisy data tested here—the higher the noise, the more noise were included in the candidate clusters—having less influence from the spatial evaluation allow the clusters to be bigger, thus a higher recall score. This can also be seen in the table, where FDR-SC has higher recall score than the one without FDR.



(a) G-CSM w/o FDR (b) G-CSM w/ FDR-SC

Fig. 5 Scatter plot of the temporal and spatial score. Red indicates the wrong relationship. Black to green indicated possibly correct relations, shaded by F-score. The blue mark is the relation with the highest F-score, with orange being the highest evaluation score. Noise = 3000, $\lambda = 2$. Grey lines represented positions with equal final evaluation scores (color figure online)

Table 3 Evaluation scores of the relation with the highest evaluation score from G-CSM

Algo.	Relation Prec.	Rec.	F-Score	Score Temporal[1]	Spatial
G-CSM w/o FDR	0.334	0.243	0.282	0.952	0.954
G-CSM w/ FDR-SC	0.331	0.270	0.298	0.942	0.947

[1]The temporal evaluation score of G-CSM without FDR and with FDR-SC are calculated differently and cannot be compared directly

4.4 Parameter analysis

4.4.1 Minimum sequence threshold $\mathcal{L}_{\text{min}}$

To analyze the effect of the minimum sequence threshold $\mathcal{L}_{\text{min}}$, we plotted the histogram of the final evaluation score of all valid cluster sequence patterns, and whether they are considered to be correct or wrong relations. The data used in this experiment were Noise = 100 and $\lambda = 2$. The other hyperparameters were the same as in Sect. 4.3.1.

The resulting histogram is shown in Fig. 6, and it is clear that G-CSM has very good separation between the evaluation score for correct and wrong relations, unlike the original CSM. We can conclude that G-CSM is less sensitive to the $\mathcal{L}_{\text{min}}$ parameter. Note that $\mathcal{L}_{\text{min}} = 0.8$ and all the generated patterns by the CSM algorithm have scores less than 0.8, so CSM cannot output any patterns. According to Fig. 6, the detected relations have final evaluation score of 0.7 or less.

4.4.2 Significant threshold α

We also investigated how changing the α significant threshold affected the result. Here, we used the same data as in the first experiment (Noise = 3000, $\lambda = 2$) with different values of α. The result is shown in Table 4. The G-CSM uses the specified alpha value directly to calculate the threshold in Eq. (13).

With different significant threshold settings from 0.001 (0.1%) to 0.2 (20%), G-CSM with FDR-SC can maintain the both the cluster and relation F-score better than the other algorithms. The result shows that G-CSM with FDR-SC is less sensitive to the alpha setting and also has the number of extracted relations (Cnt.) close to one, which is better in this case.

4.5 Other type of patterns

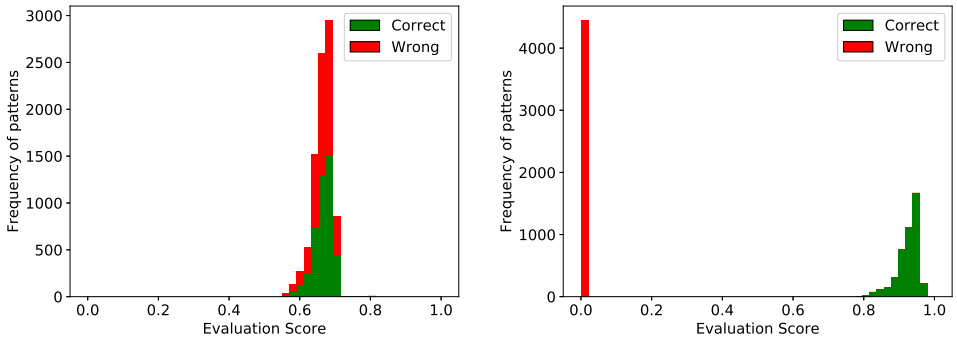
The experiment in the previous section uses a single pair of events cluster. In this section, we showed that our algorithm also worked with other types of data as well.

We tested four different types of patterns as shown in Fig. 7. First, two relations were generated using the method described in Sect. 4.1, with Noise = 1000 and $\lambda = 2$, shown in Fig. 7a. The second data also have two relations, but the prior event locations were shared between two relations, shown in Fig. 7c. Third, the posterior cluster of one relation shared the location with the prior cluster of the second relation, shown in Fig 7e. And fourth, the variance of the prior and the posterior cluster were varied, shown in Fig. 7g. The timestamps of events are shifted by a uniform random number between 0 and 10,000, different for each relation.

The hyperparameters were the same as in Sect. 4.3.1. The extracted relations are shown in Fig. 7. The proposed G-CSM with FDR-SC algorithm can correctly extract relations in all cases.

4.6 Semi-real data

In this section, we experimented using semi-real-world data. That is, real-world data are used as the spatial component, while the temporal component utilized the same method as the synthetic generation.



(a) CSM (b) G-CSM w/o FDR

Fig. 6 Histogram of $\mathcal{L}$, CSM versus G-CSM. Noise = 100, λ = 2

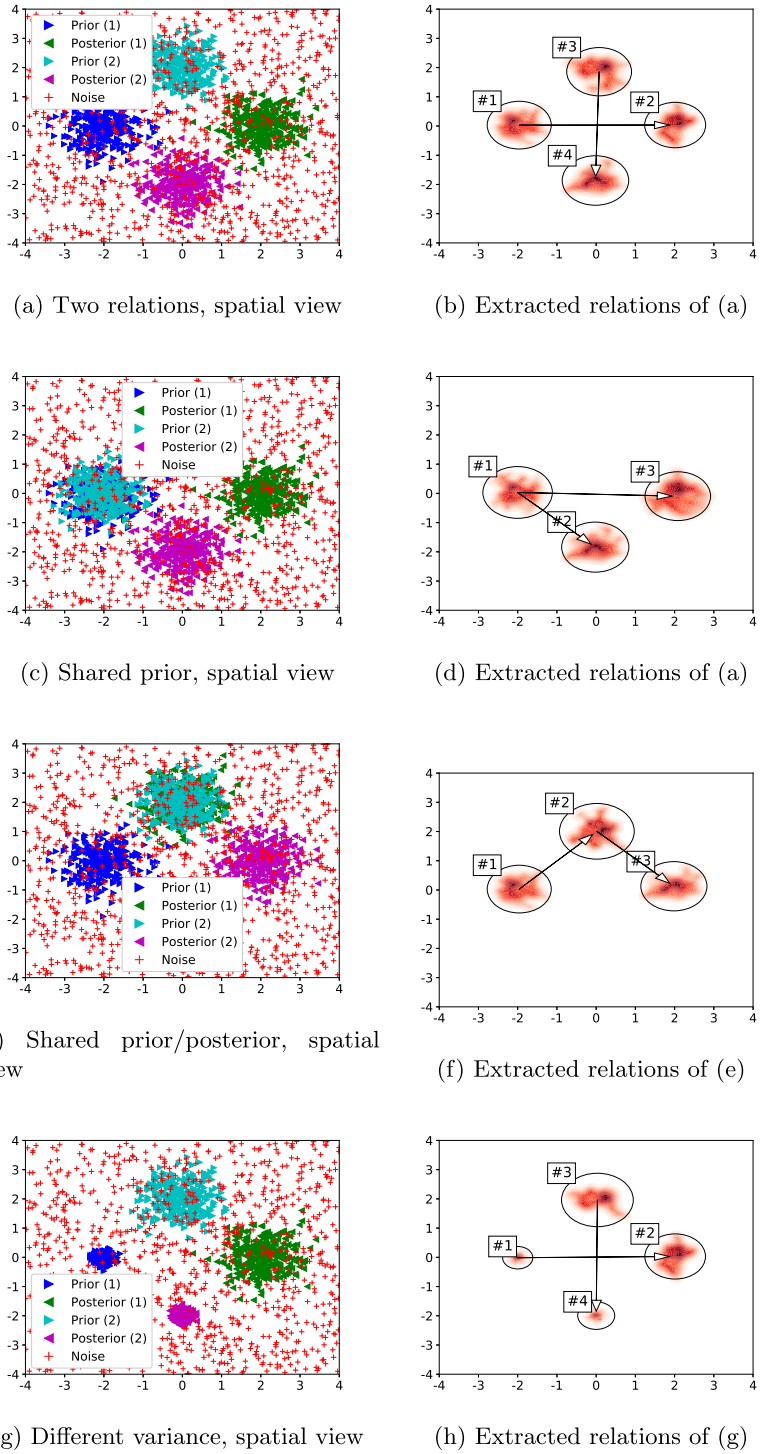
Table 4 CSM and G-CSM at various alpha settings

Algo.	P (1)	R (1)	F (1)	P (2)	R (2)	F (2)	P (R)	R (R)	F (R)	Cnt.
alpha = 0.001										
G-CSM	0.684	**0.632**	**0.650**	0.696	**0.627**	0.652	**0.428**	**0.399**	**0.410**	1.400
FDR-TH	0.657	0.499	0.537	0.679	0.555	0.594	0.329	0.279	0.295	2.950
FDR-SC	**0.695**	0.603	0.642	**0.704**	0.620	**0.654**	0.422	0.373	0.395	**1.300**
alpha = 0.005										
G-CSM	0.683	**0.633**	**0.650**	**0.700**	0.632	**0.659**	**0.431**	**0.401**	**0.413**	1.550
FDR-TH	0.663	0.492	0.539	0.676	0.552	0.592	0.330	0.277	0.295	3.000
FDR-SC	**0.685**	0.625	0.645	0.698	**0.632**	0.657	0.428	0.398	0.410	**1.450**
alpha = 0.01										
G-CSM	**0.685**	0.624	0.645	**0.700**	0.624	0.655	0.428	0.393	0.407	1.600
FDR-TH	0.606	0.509	0.528	0.612	0.545	0.563	0.304	0.274	0.283	3.250
FDR-SC	0.685	**0.627**	**0.647**	0.696	**0.641**	**0.662**	**0.431**	**0.404**	**0.414**	**1.500**
alpha = 0.05										
G-CSM	0.682	0.611	0.635	**0.703**	0.625	0.654	0.422	0.386	0.400	1.700
FDR-TH	0.452	0.519	0.457	0.462	0.560	0.481	0.234	0.284	0.245	4.050
FDR-SC	**0.685**	**0.625**	**0.646**	0.700	**0.630**	**0.658**	**0.428**	**0.394**	**0.408**	**1.600**
alpha = 0.1										
G-CSM	0.675	0.614	0.629	**0.702**	0.613	0.646	0.413	0.380	0.391	1.850
FDR-TH	0.375	0.612	0.426	0.400	**0.639**	0.455	0.222	0.392	0.264	4.950
FDR-SC	**0.682**	**0.619**	**0.640**	0.698	0.639	**0.662**	**0.428**	**0.398**	**0.410**	**1.700**
alpha = 0.2										
G-CSM	0.677	0.604	0.624	0.697	0.603	0.636	0.404	0.368	0.379	1.950
FDR-TH	0.251	**0.649**	0.348	0.296	**0.664**	0.389	0.169	**0.434**	0.235	7.050
FDR-SC	**0.682**	0.611	**0.635**	**0.701**	0.629	**0.657**	**0.423**	0.388	**0.401**	**1.700**

P = Precision, R = Recall, F = F-score, (1) = Prior cluster, (2) = Posterior cluster, (R) = Relation, Cnt. = Number of relations identified. Bold indicated the best result

We use the test data of UCI Machine Learning Optical Recognition of Handwritten Digits Data Set [35] as the spatial data. The data contain 1797 samples divided into 10 class. Each class has approximately 180 samples. The data were preprocessed to normalize the mean and variance of each dimension (z-mean normalization). The input data with 64 dimensions were reduced to 10 dimensions using neighborhood components analysis [36]. We randomly selected 2 pairs of digits as the embedded relations, while the other 6 digits data were used as noise with uniform distribution over the temporal component. The interval between each embedded event pair of both relations followed exponential distribution with $\lambda = 2$. The parameters were the same as the previous experiment in Sect. 4.3.

The result is shown in Table 5. In this data set, as there was no extra random noise added, even the original CSM, which is quite weak to spatial noise, can



(a) Two relations, spatial view

(b) Extracted relations of (a)

(c) Shared prior, spatial view

(d) Extracted relations of (a)

(e) Shared prior/posterior, spatial view

(f) Extracted relations of (e)

(g) Different variance, spatial view

(h) Extracted relations of (g)

Fig. 7 Generated data and extracted relations using G-CSM with FDR-SC for other types of pattern

extract some relations, but it still performed worse than the other algorithms. FDR-SC performed the best in all our evaluation metrics. Since the spatial dimension is reasonably well-separated and no random noise was added, the precision score was very high as the algorithm can easily extract the proper cluster. The recall score is limited by the spatial evaluation score, which prefers a compact cluster over a larger cluster. An adjustment to the hyperparameters of the evaluation function may be needed to get a higher recall score, but otherwise, both the prior and posterior cluster of the extracted relations were the subset of those of the ground-truth relation. The scores on the relation evaluation are also low for the similar reason.

Table 5 CSM and G-CSM result from digits dataset

Algo.	P (1)	R (1)	F (1)	P (2)	R (2)	F (2)	P (R)	R (R)	F (R)	Cnt.
CSM	0.349	0.063	0.106	0.350	0.173	0.231	0.185	0.063	0.094	0.350
G-CSM	**0.997**	0.519	0.675	**1.000**	0.487	0.650	0.505	0.255	0.338	**2.000**
FDR-TH	**0.997**	0.495	0.658	0.989	0.457	0.622	0.476	0.228	0.308	2.050
FDR-SC	**0.997**	**0.532**	**0.687**	**1.000**	**0.497**	**0.658**	**0.517**	**0.267**	**0.351**	**2.000**

P = Precision, R = Recall, F = F-score, (1) = Prior cluster, (2) = Posterior cluster, (R) = Relation, Cnt. = Number of relations identified. Bold indicated the best result

4.7 Complexity analysis of the G-CSM algorithm

The original CSM algorithm has a run-time complexity of $\mathcal{O}(N^2 \log N)$ in the average case, where N is the number of data points. Within the algorithm, the time proximity of temporal evaluation is $\mathcal{O}(|A| + |B|)$ where $|A|$ and $|B|$ are the number of events in the prior and posterior cluster of each pattern, respectively.

For our proposed G-CSM algorithm, the time proximity algorithm uses GLM model fitting to calculate Granger causality strength. GLM model-fitting has runtime complexity of $\mathcal{O}(p^3 + Rp^2)$ where p is the number of predictors and R is the number of samples. In our case, $p = 2 \times M_i + 1$, which is a constant, and R is at most $(2M_i + 1) \times (|A| + |B|) \sim (|A| + |B|)$; thus, the GLM model fitting takes $\mathcal{O}(|A| + |B|)$, which is the same as original CSM. The FDR procedure takes $\mathcal{O}(N^2 \log N)$ in the worst case. Thus, G-CSM also has the runtime complexity of $\mathcal{O}(N^2 \log N)$, which is also the same as the original CSM.

4.8 Limitation of the G-CSM algorithm

The G-CSM algorithm inherited all the limitations of the original Granger causality. The major point is that though Granger causality is one of the accepted methods to detect causality, we still cannot guarantee whether it is an actual causality or not, just that it is causality under Granger's definition.

Since Granger measured the causality based on predictability, it is also limited by the predictor. In the case of G-CSM, the limitation of GLM models used as a predictor is the same as the traditional multivariate vector autoregressive (MVAR) model, mainly: linearity, stationarity, and dependency on observed variables. Moreover, since we were doing pairwise causality, more data from the environment might be missed, such as when two events are the cause another event.

5 Conclusion

We proposed an extended version of the CSM algorithm with Granger causality called the G-CSM algorithm. The experiments with the synthetic data and semi-real data show that the method could identify correct relations, even in a noisy environment.

The proposed G-CSM algorithm is much better than the original CSM algorithm in almost all cases. G-CSM has been shown that it is also much less sensitive to the hyperparameters than the original CSM. The significant testing using FDR method can improve the accuracy, especially with more noisy data, and also even less sensitive to the hyperparameters.

The next steps of this research include an experiment with more semi-real data or actual real-world data, including earthquake data and climate data. Discovering interesting causal relations in these real-world data can lead to a new insight for understanding and predicting natural phenomena. There is also a need to extend the current causality detection beyond the linearity assumption we used in this work, possibly by using nonlinear Granger causality techniques [37]. Additionally, a different distribution of the interval between the event pairs other than exponential distributions may also be experimented.

Author Contributions Pavasant conceptualized the idea, implemented, performed the experiments and wrote the manuscript. Fukui also conceptualized and revised the manuscript. All authors discussed and reviewed the manuscript.

Funding Open access funding provided by Osaka University.

Declarations

Conflict of interest Not applicable.

References

1. Atluri, G., Karpatne, A., Kumar, V.: Spatio-temporal data mining: a survey of problems and methods. ACM Comput. Surv. **51**(4), 1–41 (2018)
2. González, J.A., Rodríguez-Cortés, F.J., Cronie, O., Mateu, J.: Spatio-temporal point process statistics: a review. Spat. Stat. **18**, 505–544 (2016)
3. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969)
4. Le, T.D., Hoang, T., Li, J., Liu, L., Liu, H., Hu, S.: A fast PC algorithm for high dimensional causal discovery with multi-core PCs. IEEE/ACM Trans. Comput. Biol. Bioinf. **16**(5), 1483–1495 (2019)
5. Fukui, K., Inaba, D., Numao, M.: Discovery of damage patterns in fuel cell and earthquake occurrence patterns by co-occurring cluster mining. In: Proceedings of the 2014 AAAI Workshop for Discovery Informatics, pp. 19–26 (2014)
6. Fukui, K., Okada, Y., Satoh, K., Numao, M.: Cluster sequence mining from event sequence data and its application to damage correlation analysis. Knowl.-Based Syst. **179**, 136–144 (2019)
7. Bressler, S.L., Seth, A.K.: Wiener–Granger causality: a well established methodology. Neuroimage **58**(2), 323–329 (2011)
8. Pavasant, N., Numao, M., Fukui, K.: Spatio-temporal change detection with Granger causality based cluster sequence mining. In: Proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 551–558 (2020)
9. Cox, D.R., Isham, V.: Point Processes. Monographs on Applied Probability and Statistics, Chapman and Hall, London (1980)
10. Truccolo, W., Eden, U.T., Fellows, M.R., Donoghue, J.P., Brown, E.N.: A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J. Neurophysiol. **93**(2), 1074–1089 (2005)
11. Ogata, Y.: Space-time point-process models for earthquake occurrences. Ann. Inst. Stat. Math. **50**(2), 379–402 (1998)
12. Ogata, Y., Zhuang, J.: Space-time etas models and an improved extension. Tectonophysics **413**(1), 13–23 (2006)
13. Musmeci, F., Vere-Jones, D.: A space-time clustering model for historical earthquakes. Ann. Inst. Stat. Math. **44**(1), 1–11 (1992)
14. Chen, R.T.Q., Amos, B., Nickel, M.: Neural spatio-temporal point processes. In: International Conference on Learning Representations (2021)
15. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1555–1564 (2016)
16. Zhou, Z., Matteson, D.S., Woodard, D.B., Henderson, S.G., Micheas, A.C.: A spatio-temporal point process model for ambulance demand. J. Am. Stat. Assoc. **110**(509), 6–15 (2015)
17. Okawa, M., Iwata, T., Kurashima, T., Tanaka, Y., Toda, H., Ueda, N.: Deep mixture point processes: spatio-temporal event prediction with rich contextual information. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 373–383 (2019)
18. Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., Song, L.: Learning temporal point processes via reinforcement learning. Adv. Neural Inf. Process. Syst. **31**, 10804–10814 (2018)
19. Higuchi, M., Matsutani, K., Kumano, M., Kimura, M.: Discovering spatio-temporal latent influence in geographical attention dynamics. In: Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2018), vol. 11052, pp. 517–534 (2019)
20. Zhu, S., Li, S., Peng, Z., Xie, Y.: Interpretable deep generative spatio-temporal point processes. In: AI for Earth Sciences Workshop at NeurIPS (2020)
21. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**(3), 358–386 (2005)
22. Davidson, I., Gilpin, S., Carmichael, O., Walker, P.: Network discovery via constrained tensor analysis of FMRI data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 194–202 (2013)
23. Ebert-Uphoff, I., Deng, Y.: Causal discovery from spatio-temporal data with applications to climate science. In: Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA 2014), pp. 606–613 (2015)
24. Stokes, P.A., Purdon, P.L.: A study of problems encountered in granger causality analysis from a neuroscience perspective. Proc. Natl. Acad. Sci. **114**(34), 7063–7072 (2017)
25. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**(2), 461–464 (2000)
26. Barnett, L., Barrett, A.B., Seth, A.K.: Granger causality and transfer entropy are equivalent for Gaussian variables. Phys. Rev. Lett. **103**, 238701 (2009)
27. Sun, J., Bollt, E.M.: Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. Physica D **267**, 49–57 (2014)
28. Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., Munch, S.: Detecting causality in complex ecosystems. Science **338**(6106), 496–500 (2012)
29. Kim, S., Putrino, D., Ghosh, S., Brown, E.N.: A granger causality measure for point process models of ensemble neural spiking activity. PLoS Comput. Biol. **7**(3), 1001110 (2011)
30. Casile, A., Faghih, R.T., Brown, E.N.: Robust point-process granger causality analysis in presence of exogenous temporal modulations and trial-by-trial variability in spike trains. PLoS Comput. Biol. **17**(1), 1007675 (2021)
31. Ansari, M.Y., Ahmad, A., Khan, S.S., Bhushan, G., Mainuddin: Spatiotemporal clustering: a review. Artif. Intell. Rev. **53**, 2381–2423 (2020)
32. King, G.: Unifying Political Methodology: The Likelihood Theory of Statistical Inference. University of Michigan Press, Ann Arbor (1998)
33. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc.: Ser. B (Methodol.) **57**(1), 289–300 (1995)
34. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
35. Dua, D., Graff, C.: UCI Machine Learning Repository (2019). http://archive.ics.uci.edu/ml
36. Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R.: Neighbourhood components analysis. Adv. Neural. Inf. Process. Syst. **17**, 513–520 (2005)
37. Rosoł, M., Młyńczak, M., Cybulski, G.: Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. Comput. Methods Progr. Biomed. **216**, 106669 (2022)