p-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data

Julie Yixuan Zhu[1,3,*]   Chao Zhang[2,3,*]   Huichu Zhang[2,4]   Shi Zhi[2]   Victor O.K. Li[1]
Jiawei Han[2]   Yu Zheng[3,+]

[1]Department of Electrical and Electronic Engineering, the University of Hong Kong, HK
[2]Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[3]Microsoft Research Asia, Beijing, China
[4]Apex Data & Knowledge Management Lab, Shanghai Jiao Tong University, Shanghai.
[1]{yxzhu,vli}@eee.hku.hk   [2]{czhang82,shizhi2, hanj}@illinois.edu   [3]yuzheng@microsoft.com
[4]zhc@apex.sjtu.edu.cn

ABSTRACT

Many countries are suffering from severe air pollution. Understanding how different air pollutants accumulate and propagate is critical to making relevant public policies. In this paper, we use urban big data (air quality data and meteorological data) to identify the *spatiotemporal (ST) causal pathways* for air pollutants. This problem is challenging because: (1) there are numerous noisy and low-pollution periods in the raw air quality data, which may lead to unreliable causality analysis; (2) for large-scale data in the ST space, the computational complexity of constructing a causal structure is very high; and (3) the *ST causal pathways* are complex due to the interactions of multiple pollutants and the influence of environmental factors. Therefore, we present *pg-Causality*, a novel pattern-aided graphical causality analysis approach that combines the strengths of *pattern mining* and *Bayesian learning* to efficiently identify the *ST causal pathways*. First, *pattern mining* helps suppress the noise by capturing frequent evolving patterns (FEPs) of each monitoring sensor, and greatly reduce the complexity by selecting the pattern-matched sensors as "causers". Then, *Bayesian learning* carefully encodes the local and ST causal relations with a Gaussian Bayesian Network (GBN)-based graphical model, which also integrates environmental influences to minimize biases in the final results. We evaluate our approach with three real-world data sets containing 982 air quality sensors in 128 cities, in three regions of China from 01-Jun-2013 to 31-Dec-2016. Results show that our approach outperforms the traditional causal structure learning methods in time efficiency, inference accuracy and interpretability.

Keywords

Causality; pattern mining, Bayesian learning; spatiotemporal (ST) big data; urban computing.

1. INTRODUCTION

Recent years have witnessed the air pollution problem becoming a severe environmental and societal issue around the world. For example, in 2015, the average concentration of PM2.5 in Beijing is greater than 150, classified as hazardous to human health by the World Health Organization, on more than 46 days. On Dec 7th 2015, the Chinese government issues the first red alert because of the extremely heavy air pollution, leading to suspended schools, closed construction sites, and traffic restrictions. Though many ways have been deployed to reduce the air pollution, the severe air pollution in Beijing has not been significantly alleviated.

Identifying the causalities has become an urgent problem for mitigating the air pollution and suggesting relevant public policy making. Previous research on the air pollution cause identification mostly relies on chemical receptor [1] or dispersion models [2]. However, these approaches often involve domain-specific data collection which is labor-intensive, or require theoretical assumptions that real-world data may not guarantee. Recently, with the increasingly available air quality data collected by versatile sensors deployed in different regions, and pubic meteorological data, it is possible to analyze the causality of air pollution through a data-driven approach.

The goal of our research is to learn the *spatiotemporal (ST) causal pathways* among different pollutants, by mining the dependencies among air pollutants under different environmental influences. Fig. 1 shows two example causal pathways for PM10 in Beijing. Let us first consider the pathway in Fig. 1(a). When the wind speed is less than 5 m/s, the high concentration of PM10 in Beijing is mainly caused by SO_2 in Zhangjiakou and PM2.5 in Baoding. In contrast, as shown in Fig. 1(b), when the wind speed is larger than 5m/s, PM10 in Beijing is mainly due to PM2.5 in Zhangjiakou and NO_2 in Chengde. Based on this example, we can see the *spatiotemporal (ST) causal pathways* should reflect the following two aspects: 1) the *structural dependency*, which indicates the reactions and propagations of multiple pollutants in the ST space; and 2) the *global confounder*, which denotes how different environmental conditions could lead to different causal pathways.

However, identifying the *ST causal pathways* from big air quality and meteorological data is not trivial because of the following challenges. *First*, not all air pollution data are useful for causality analysis. In the raw sensor-collected air quality data, there are numerous uninteresting fluctuations and noisy variations. Including such data into the causality analysis process is expected to lead to unreliable conclusions. *Second*, the sheer size of the air quality makes the causality analysis difficult. In most air quality moni-
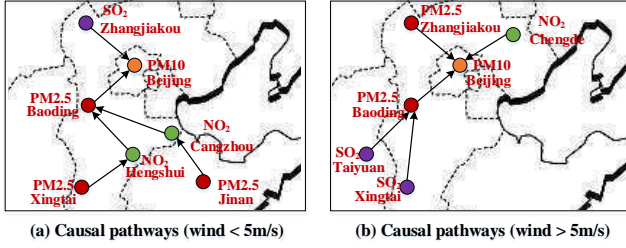
(a) Causal pathways (wind < 5m/s)   **(b) Causal pathways (wind > 5m/s)**

Figure 1: An illustration of identifying causal pathways.

toring applications, thousands of sensors are deployed at different locations to record the air quality hourly for years. Discovering the ST causal relationships from such a large scale is challenging. *Third*, air pollution causal pathways are complex in nature. The air polluting process typically involves multiple types of pollutants that are mutually interacting, and is subject to local reactions, ST propagations and confounding factors, such as wind and humidity.

Existing data mining techniques for learning the causal pathways have been proposed from two perspectives: pattern-based [3][4] and Bayesian-based [5][6]. Pattern-based approaches aim to extract frequently occurring phenomena from historical data by applying pattern mining techniques; while Bayesian-based techniques use directed acyclic graphs (DAGs) to encode the causality and then learn the probabilistic dependencies from historical data. Though inspiring results have been obtained by pattern-based and Bayesian-based techniques, both approaches have their merits and downsides. Pattern-based approaches can fast extract a set of patterns (e.g., frequent patterns, contrast patterns) from historical air quality data. Such patterns can capture the intrinsic regularity present in historical air quality data. However, they only provide shallow understanding of the air polluting process, and there are usually a huge number of frequent patterns, which largely limits the usability of the pattern set. On the other hand, Bayesian-based approaches depict the causal dependencies between multiple air pollutants in a principled way. However, the performance of Bayesian-based models is highly dependent on the quality of the training data. When there exist massive noise and data sparsity, as the case of the air quality data, the performance of the Bayesian-based models is limited. Besides, Bayesian-based approaches are limited by high computational cost [7] and the impact of confounding [8].

We propose *pg-Causality*, which combines *pattern mining* with *Bayesian learning* to unleash the strengths of both. We claim *pg-Causality* is essential for *ST causal pathway* identification, with the contributions listed as below:

• First, we propose a framework that combines frequent pattern mining with Bayesian-based graphical model to identify the spatiotemporal (ST) causal relationship between air pollutants in the ST space. The frequent pattern mining [9] can accurately estimate the correlation between the air quality of each pair of locations, capturing the meaningful fluctuation of two time series. Using the correlation patterns, whose scales are significantly smaller than the raw data, as an input of a Bayesian network (BN), the computational complexity of the Bayesian network causality model has been significantly reduced. The patterns also help suppress the noise for learning a Bayesian network's structure. This not only leads to a more efficient but also more effective causal pathway identification. We also integrate the environmental factors in the Bayesian-based graphical model to minimize the biases in the final results.

• Second, we have carefully evaluated our proposed approach on three real data sets with 3.5 years' air quality and meteorological data collected from hundreds of cities in China. Our results show that the proposed approach is significantly better than the existing baseline methods in time efficiency, inference accuracy and interpretability.

2. FRAMEWORK

In this section, we first describe the problem of identifying spatiotemporal causal pathways for air pollutants, and then introduce the framework of *pg-Causality*.

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n, \ldots\}$ be the location set of the air quality monitoring sensors deployed in a geographical region. Each sensor is deployed at a location $s_n \in \mathcal{S}$ to periodically measure the target condition around it. All sensors have synchronized measurements over the time domain $\mathcal{T} = \{1, 2, \ldots, T\}$, where each $t \in \mathcal{T}$ is a timestamp. We also consider a set $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ of pollutants. Given $c_m \in \mathcal{C}$, $s_n \in \mathcal{S}$, and $t \in \mathcal{T}$ ($1 \le m \le M, 1 \le n \le N, 1 \le t \le T$), we use $P_{c_m s_n t}$ to denote the measurement of pollutant c_m at location s_n and timestamp t. In addition, we also have the meteorological data at timestamp t for the entire geographical region, denoted as $\boldsymbol{E_t}$, as a vector of environmental factors. Using the air pollutant measurements and meteorological data, we aim to identify faithful causal relationships among different pollutants at different locations. We integrated the environmental facotors $\boldsymbol{E_t}$ to the causal pathways through a graphical model, setting the number of clusters as K and time lag constraint as L. We list the notations in TABLE 1.

Table 1: Notation Table.

$\mathcal{S}$	The location set of the air quality monitoring sensors. $\mathcal{S} = \{s_1, s_2, \ldots, s_n, \ldots\}$
$s_n \in \mathcal{S}$	The location of the n-th neighborhood sensor.
s_0	The location of the target sensor.
N	Number of "causers" in the neighborhood.
$\mathcal{T}$	Timestamps domain $\mathcal{T} = \{1, 2, \ldots, T\}$.
$t \in \mathcal{T}$	The current timestamp.
T	Number of timestamps.
$\mathcal{C}$	Category set of pollutants $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$.
M	Number of pollutants measured by each sensor.
$c_m \in \mathcal{C}$	The pollutant of the m-th category.
c_{m_n}	The most likely category of "causer" pollutant at s_n.
$P_{c_m s_n t}$	Pollutant c_m at location s_n and timestamp t. $1 \le m \le M, 1 \le n \le N, 1 \le t \le T$.
K	Number of clusters in the graphical causality model.
$l \in [1, L]$	Time lag in the graphical causality model.
$\boldsymbol{E_t}$	The environmental factors. $\boldsymbol{E_t} = \{E_t^{(1)}, E_t^{(2)}, \ldots\}$.

Fig. 2 shows the framework of our proposed approach *pg-Causality*. It consists of two main modules: pattern mining and Bayesian Network Learning, detailed as follows.
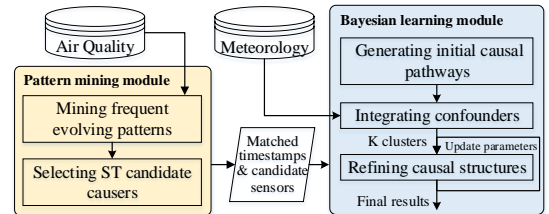

Figure 2: The framework of our approach.

Pattern Mining Module: This module first extracts the *frequent evolving patterns* (FEPs) [9] for each sensor. The FEPs essentially

capture the air quality changing behaviors that frequently appear on the target sensor. By mining all FEPs from the historical air quality data, this module efficiently captures the regularity in raw data and largely reduces the noise (Section 3.1 and 3.2). Afterwards, we examine the pattern-based similarities between locations to select candidate causers for each target sensor. By comparing the FEPs occurring on different sensors, we can obtain a shallow understanding of the causal relationships between different sensors, which can be further utilized to simplify learning the causal structures (Section 3.3).

Bayesian Learning Module: By using the matched timestamps of the extracted FEPs at different sensors, together with the selected candidate sensors in the pattern mining module, this module further trains high-quality causal pathways from the large-scale air quality and context data in an effective and scalable way. We first generate the initial causal pathways from the selected candidate causers, taking into account both the local interactions of multiple air pollutants and the ST propagations (Section 4.1). Then to minimize the impact of confounding (Section 4.2), we integrate the confounders (e.g., wind, humidity) into the a Gaussian Bayesian Network (GBN)-based graphical model. Last, we refine the parameters and structures of the Bayesian network to generate the final causal pathways (Section 4.3).

We argue that the combination of two modules helps efficiently identify the causal pathways of the air pollutants. First, the meaningful behaviors of each time series selected by the pattern mining module could significantly reduce the noise in calculating the causal relationships. For example, Fig. 3(a) shows an illustration of three time series at sensors 1, 2, and 3, in North China, with sensor 1 as the target sensor. When simply using statistical models to identify the dependencies among the three time series, the causal pathway $2 \to 1$ and $3 \to 1$ cannot be faithfully justified, since the fluctuations and low pollution periods will make the dependency metric for sensors $2 \to 1$ and $3 \to 1$ very similar. By using the pattern mining module, we found that the increasing behaviors of sensor 2 frequently happen before sensor 1, and thus can select sensor 2 as the candidate "causer" for target sensor 1. Second, the selected "causers" by the pattern mining module will greatly reduce the complexity of the Bayesian structure learning. For example, Fig. 3(b) illlustrates a scenario of learning the 1-hop Bayesian structure from 100 sensors to a target pollutant. We use the pattern mining module to select top "N = 2" candidate causers, thus reducing the searching space from $O(100)$ to $O(2)$ for Bayesian structure construcion. Third, we verify the effectiveness of causal pathway learning with pg-Causality, compared with only using Bayesian learning without pattern mining. Combining pattern mining with Bayesian learning demonstrates better inference accuracy, time efficiency, and interpretability.
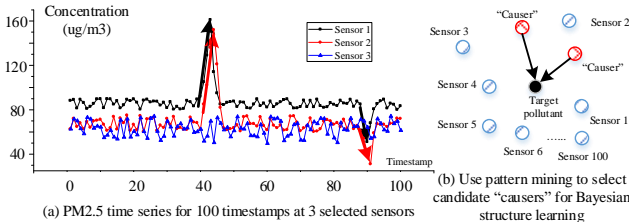

(a) PM2.5 time series for 100 timestamps at 3 selected sensors

(b) Use pattern mining to select candidate "causers" for Bayesian structure learning

Figure 3: Illustration of how pattern mining helps to reduce the effect of fluctuations in causal structure learning.

3. THE PATTERN MINING MODULE

3.1 Frequent Evolving Pattern

To capture frequent evolving behaviors of each sensor, we define *frequent evolving pattern* (FEP), an adaption of the classic sequential pattern concept [29]. As the sequential patterns are defined on transactional sequences, we first discretize the raw air quality data. Given a pollutant c_m at sensor s_n, the measurements of c_m at s_n over the time domain $\mathcal{T}$ form a time series. We discretize the time series as follows: (1) partition it by day to obtain a collection of daily time series, denoted as $P_{c_m s_n}$; and (2) for each daily time series $\langle (p_1, t_1), (p_2, t_2), \ldots, (p_l, t_l) \rangle$, map every real-value measure p_i $(1 \leq i \leq l)$ to a discrete level $\hat{p}_i$ using *symbolic approximation aggregation* [30]. After discretization, we obtain a database of symbolic sequences, as defined in Definition 1.

DEFINITION 1 (SYMBOLIC POLLUTION DATABASE). *For pollutant c_m and sensor s_n, the symbolic pollution database $\hat{P}_{c_m s_n}$ is a collection of daily sequences. Each sequence $d \in \hat{P}_{c_m s_n}$ has the form $\langle (\hat{p}_1, t_1), (\hat{p}_2, t_2), \ldots, (\hat{p}_l, t_l) \rangle$ where an element $(\hat{p}_i, t_i)$ means the pollution level of c_m at sensor s_n and time t_i is $\hat{p}_i$.*

Given the database $\hat{P}_{c_m s_n}$, our goal is to find frequent evolving behaviors of s_n regarding c_m. Below, we introduce the concepts of *evolving sequence* and *occurrence*.

DEFINITION 2 (EVOLVING SEQUENCE). *A length-k evolving sequence T has the form $T = \hat{p}_1 \xrightarrow{\Delta t} \hat{p}_2 \xrightarrow{\Delta t} \cdots \xrightarrow{\Delta t} \hat{p}_k$, where (1) $\forall i > 1, \hat{p}_{i-1} \neq \hat{p}_i$ and (2) Δt is the maximum transition time between consecutive records.*

DEFINITION 3 (OCCURRENCE). *Given a daily sequence $d = \langle (\hat{p}_1, t_1), (\hat{p}_2, t_2), \ldots, (\hat{p}_l, t_l) \rangle$ and an evolving sequence $T = \hat{p}_1 \xrightarrow{\Delta t} \hat{p}_2 \cdots \xrightarrow{\Delta t} \hat{p}_k$ $(k \leq l)$, T occurs in d (denoted as $T \sqsubseteq d$) if there exist integers $1 \leq j_1 < j_2 < \cdots < j_k \leq l$ such that: (1) $\forall 1 \leq i \leq k, \hat{p}_{j_i} = \hat{p}_i$; and (2) $\forall 1 \leq i \leq k-1, 0 < t_{j_{i+1}} - t_{j_i} \leq \Delta t$.*

For clarity, we denote an evolving sequence $\hat{p}_1 \xrightarrow{\Delta t} \hat{p}_2 \cdots \xrightarrow{\Delta t} \hat{p}_k$ as $\hat{p}_1 \to \hat{p}_2 \cdots \to \hat{p}_k$ when the context is clear. Now, we proceed to define *support* and *frequent evolving pattern*.

DEFINITION 4 (SUPPORT). *Given $\hat{P}_{c_m s_n}$ and an evolving sequence T, the support of T is the number of days that T occurs, i.e., $Sup(T) = |\{o | o \in \hat{P}_{c_m s_n} \wedge T \sqsubseteq o\}|$.*

DEFINITION 5 (FREQUENT EVOLVING PATTERN). *Given a support threshold σ, an evolving sequence T is a frequent evolving pattern in database $\hat{P}_{c_m s_n}$ if $Sup(T) \geq \sigma$.*

3.2 The FEP Mining Algorithm

Now we proceed to discuss how to mine all FEPs in any symbolic pollution database. It is closely related to the classic sequential pattern mining problem. However, recall that there are two constraints in the definition of FEP: (1) the consecutive symbols must be different; and (2) the time gap between consecutive records should be no greater than the temporal constraint Δt. A sequential pattern mining algorithm needs to be tailored to ensure these two constraints are satisfied.

We adapt PrefixSpan [29] as it has proved to be one of the most efficient sequential pattern mining algorithms. The basic idea of PrefixSpan is to use short patterns as the prefix to project the database and progressively grow the short patterns by searching for local frequent items. For a short pattern β, the β-projected database $\mathcal{D}_\beta$ includes the postfix from the sequences that contain β. Local frequent

items in $\mathcal{D}_\beta$ are then identified and appended to β to form longer patterns. Such a process is repeated recursively until no more local frequent items exist. One can refer to [29] for more details.

Given a sequence α and a frequent item $\hat{p}$, when creating $\hat{p}$-projected database, the standard PrefixSpan procedure generates one postfix based on the first occurrence of $\hat{p}$ in α. This strategy, unfortunately, can miss FEPs in our problem.

Table 2: An example symbolic pollution database.

Day	Daily sequence
d_1	$\langle(\hat{p}_2, 0), (\hat{p}_1, 10), (\hat{p}_2, 30), (\hat{p}_3, 40)\rangle$
d_2	$\langle(\hat{p}_1, 0), (\hat{p}_2, 30), (\hat{p}_1, 360), (\hat{p}_2, 400), (\hat{p}_3, 420)\rangle$
d_3	$\langle(\hat{p}_2, 0), (\hat{p}_3, 30)\rangle$
d_4	$\langle(\hat{p}_1, 0), (\hat{p}_1, 120), (\hat{p}_3, 140), (\hat{p}_2, 150), (\hat{p}_3, 180)\rangle$
d_5	$\langle(\hat{p}_2, 50), (\hat{p}_2, 80), (\hat{p}_3, 120), (\hat{p}_1, 210)\rangle$

EXAMPLE 1. *Let $\Delta t = 60$ and $\sigma = 3$. In the database shown in TABLE 2, item $\hat{p}_1$ is frequent. The $\hat{p}_1$-projected database generated by PrefixSpan is:*

   (1) $d_1/\hat{p}_1 = \langle(\hat{p}_2, 20), (\hat{p}_3, 30)\rangle$

   (2) $d_2/\hat{p}_1 = \langle(\hat{p}_2, 30), (\hat{p}_1, 360), (\hat{p}_2, 400), (\hat{p}_3, 420)\rangle$

   (3) $d_4/\hat{p}_1 = \langle(\hat{p}_1, 120), (\hat{p}_3, 140), (\hat{p}_2, 150), (\hat{p}_3, 180)\rangle$

The elements satisfying $t \le 60$ are $(\hat{p}_2, 20)$, $(\hat{p}_3, 30)$ and $(\hat{p}_2, 30)$. No local item is frequent, hence $\hat{p}_1$ cannot be grown any more.

To overcome this, given a sequence α and a frequent item $\hat{p}$, we generate a postfix for every occurrence of $\hat{p}$.

EXAMPLE 2. *Also for Example 1, if we generate a postfix for every occurrence of $\hat{p}_1$, the $\hat{p}_1$-projected database is:*

   (1) $d_1/\hat{p}_1 = \langle(\hat{p}_2, 20), (\hat{p}_3, 30)\rangle$

   (2) $d_2/\hat{p}_1 = \langle(\hat{p}_2, 30), (\hat{p}_1, 360), (\hat{p}_2, 400), (\hat{p}_3, 420)\rangle$

   (3) $d_2/\hat{p}_1 = \langle(\hat{p}_2, 40), (\hat{p}_3, 60)\rangle$

   (4) $d_4/\hat{p}_1 = \langle(\hat{p}_1, 120), (\hat{p}_3, 140), (\hat{p}_2, 150), (\hat{p}_3, 180)\rangle$

   (5) $d_4/\hat{p}_1 = \langle(\hat{p}_3, 20), (\hat{p}_2, 30), (\hat{p}_3, 60)\rangle$

The items $\hat{p}_2$ and $\hat{p}_3$ are frequent and meanwhile satisfy the temporal constraint, thus longer patterns $\hat{p}_1 \xrightarrow{60} \hat{p}_2$ and $\hat{p}_1 \xrightarrow{60} \hat{p}_3$ are found in the projected database.

Using the above projection principle, the projected database includes all postfixes to avoid missing patterns under the time constraint. Algorithm 1 sketches our algorithm for mining FEPs. The procedure is similar to the standard PrefixSpan algorithm in [29], except that the aforementioned full projection principle is adopted, and the time constraint Δt is checked when searching for local frequent items.
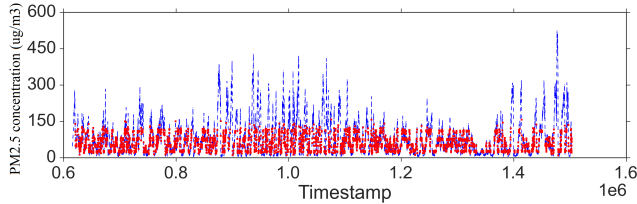

Figure 4: An illustration of the pattern-matched timestamps. The blue dashed lines represents the PM2.5 time series in Beijing during a two-year period, and the red points denote the timestamps at which a certain FEP has occurred ($\sigma = 0.1$).

Algorithm 1: Mining frequent evolving patterns.

Input: support threshold σ, temporal constraint Δt, symbolic pollution database $\hat{P}$

1 **Procedure** *InitialProjection($\hat{P}$, σ, Δt)*
2    $\leftarrow$ frequent items in $\mathcal{D}$;
3   **foreach** *item i in* **do**
4      $S \leftarrow \phi$;
5   **foreach** *sequence o in $\hat{P}$* **do**
6        $R \leftarrow$ postfixes for all occurrences of i in o;
7        $S \leftarrow S \cup R$;
8   PrefixSpan(i, i, 1, S, Δt);

9 **Function** *PrefixSpan(α, i_{prev}, l, $S|_\alpha$, Δt)*
10    $\leftarrow$ frequent items in $S|_\alpha$ meeting time constraint Δt;
11   **foreach** *item i in* **do**
12   **if** $i \ne i_{prev}$ **then**
13        $\alpha' \leftarrow$ append i to α;
14   Build $S|_{\alpha'}$ using full projection;
15   Output α';
16   PrefixSpan(α', i, $l + 1$, $S|_{\alpha'}$, Δt);

The output of Algorithm 1 is the set of all FEPs for the given database, along with the occurring timestamps for each FEP. As an example, Fig. 4 shows the raw PM2.5 time series in Beijing during a two-year period. After mining FEPs on the symbolic pollution database, we mark the timestamps at which the FEPs occur. One can observe that, the FEPs can effectively capture the regularly appearing evolvements of PM2.5 in Beijing. Because of the support threshold and the evolving constraint, infrequent sudden changes and uninteresting fluctuations are all suppressed.

3.3 Finding Candidate Causers

After discovering the FEPs, next step is leverage them to extract the candidate causers for each sensor. Consider two sensors s and s', let us use $TS(s)$ and $TS(s')$ to denote the sets of pattern starting timestamps for s and s', respectively. Below, we introduce the *pattern match* relationship.

DEFINITION 6 (PATTERN MATCH). *Let $t_{s'} \in TS(s')$ be a timestamp at which a pattern happens on s'. For a pattern starting timestamp $t_s \in TS(s)$, we say $t_{s'}$ matches t_s if $0 \le t_s - t_{s'} \le L$, where L is a pre-specified time lag threshold.*

Informally, the pattern match relation states that when there is a pattern occurring on s', then within some time interval, there is another pattern happening on s. Naturally, if s' has a strong causal effect on s, then most timestamps in $TS_{s'}$ will be matched by TS_s, and vice versa. Based on TS_s and $TS_{s'}$, we proceed to introduce *match precision* and *match recall* to quantify the correlation between s and s'.

DEFINITION 7 (MATCH PRECISION). *Given TS_s and $TS_{s'}$, we define the matched timestamp set of $TS_{s'}$ as $M_{s'} = \{t_{s'}|t_{s'} \in TS_{s'} \wedge \exists t_s \in TS_s, match(t_s, t_{s'}) = True\}$. With $M_{s'}$ and $TS_{s'}$, we define the precision of s' matching s as:*

$$P(s, s') = |M_{s'}|/|TS_{s'}|$$

DEFINITION 8 (MATCH RECALL). *Given TS_s and $TS_{s'}$, we define the matched timestamp set of TS_s as $M_s = \{t_s|t_s \in TS_s \wedge \exists t_{s'} \in TS_{s'}, match(t_s, t_{s'}) = True\}$. With M_s and TS_s, we define the recall of s' matching s as:*

$$R(s, s') = |M_s|/|TS_s|$$

Relying on the concepts of *match precision* and *match recall*, we compute the pattern-based correlation between s and s' as:

$$Corr(s, s') = \frac{2 \times P(s, s')}{P(s, s') + R(s, s')}.$$

Now we are ready to describe the process of finding candidate causers for each sensor. Given the set of all sensors and their pattern-starting timestamps, our goal is to find the candidate causers for each sensor. Consider a target sensor s, we say another sensor s' is a candidate causer for s if s' satisfies two constraints: (1) the distance between s and s' is no larger than a distance threshold δ_g; and (2) the pattern correlation between s and s' is no less than a correlation threshold δ_p. Given the pattern-starting timestamps that are ordered chronologically, the retrieval of the candidate causers can be easily done by sequentially scanning the two timestamp lists to find pattern-matched pairs.

Fig. 5 illustrates eight examples of selected candidate causers. For PM2.5 in Beijing, we reduce the number of candidate sensors to $X = 4 \sim 7$ from overall $|\mathcal{S}| = 61$ sensors in North China. Note that China is a country with monsoon climate, the candidate sensors show quite similar geo-locations in four seasons. We therefore separate the training data into four groups based on seasons, to better diagnose causalities for the air pollutants in China.
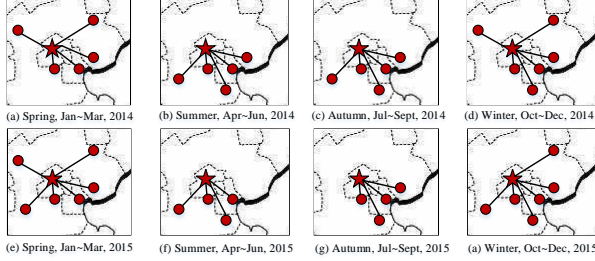

(a) Spring, Jan~Mar, 2014 (b) Summer, Apr~Jun, 2014 (c) Autumn, Jul~Sept, 2014 (d) Winter, Oct~Dec, 2014

(e) Spring, Jan~Mar, 2015 (f) Summer, Apr~Jun, 2015 (g) Autumn, Jul~Sept, 2015 (a) Winter, Oct~Dec, 2015

Figure 5:   Candidate sensors for Beijing PM2.5 in four seasons. Star: PM2.5 in Beijing. Circles: pollutants at candidate sensors.

4.   THE BAYESIAN LEARNING MODULE

In this section we first discuss how the causality learning benefits from the pattern-matched data extracted by the *pattern mining module*. Then we dive into the methodology with the *Bayesian learning module*.

Identifying the ST causality (causal pathways) for air pollutants is a problem of learning the causal structures for multiple variables, which has been well discussed with the graphical causality [5] based on Bayesian network (BN) [23]. Specifically, BN encodes the cause-and-effect relations in a directed acyclic graphs (DAG) via probabilistic dependencies. Learning BN structure from data is NP-complete [7], in the worst case requiring $2^{O(n^2)}$ searches among all the possible (DAGs). Thus when the number of variables becomes very large, the computational complexity will be unbearable. Therefore, we add the *pattern mining module* before the *Bayesian learning module* to combine the strengths of both. Pattern mining helps Bayesian learning by reducing the whole data to the selected candidate sensors and the periods matched by patterns, which greatly reduce the computational complexity as well as the noise in causality calculation. However, since the selected frequent patterns essentially demonstrates the "correlation", which is not "causality" [31], the *Bayesian learning module* helps represent and learn the causality.

Another benefit of conducting frequent pattern mining before Bayesian learning is that the selected frequent patterns could reflect the meaningful changes of the air pollutants, such as increase, decrease, sharp increase, sharp decrease, etc, thus significantly reducing the noises in Bayesian learning. When simply using Bayesian learning to identify the causality among different air pollutants time series, unreliable causal relations may be captured since there are many fluctuations and long-period low pollution cases which lead to unexpected correlation between two time series.

There are two major challenges to learn the causality among different pollutants in the ST space. The first one is to define a comprehensive representation of the causal pathways and diagnose the complex reactions and dispersions of different air pollutants. For example, the PM2.5 time series in Beijing can be strongly dependent on the NO2 time series locally, while it can also be influenced by the PM10 in another city. Therefore, both the local and ST dependencies need to be fairly considered in the model. We propose a Gaussian Bayesian network (GBN)-based graphical model, which captures the dependencies both locally and in the ST space. We elaborate how to generate initial causal pathways by GBN in Section 4.1. The second challenge is to learn faithful causal pathways given different weather conditions. As the example shown in Fig. 1, there could be different causal pathways under different wind speeds. We thus propose a method that integrates the meteorological data in the graphical model via a hidden factor representing the weather status (Section 4.2). In this way we can minimize the biases in the learning, and refine the final causal pathways (Section 4.3).

Here we give an example of combining the *pattern mining module* with the *Bayesian learning module*. Consider there are $|\mathcal{S}|$ monitoring sensors, with each sensor monitoring M categories of pollutants, there will be $|\mathcal{S}| \times M$ variables in total for the Bayesian causal structure learning and the corresponding computational complexity will be $2^{O((|\mathcal{S}| \times M)^2)}$. When combining the *pattern mining module*, we first extract the FEPs for each pollutant $P_{c_m s_n}$, i.e., the pollutant of category $m \in [1, M]$ collected at sensor $s_n \in \mathcal{S}$. Afterwards, for each target pollutant we select the pattern-matched periods (the timestamps that patterns at the neighborhood sensors happen ahead of the target sensor within some time interval, see Definition 6), as well as its top $|X|$ candidate causers (the $|X|$ neighborhood sensors that have the highest pattern-based correlation, see Definition 7 and 8). We then feed the pattern-matched periods selected and the candidate causers into the *Bayesian learning module*. In this way the computational complexity is reduced to $O(|X| \times M)$, and the noises and fluctuations in the raw data are greatly suppressed.

4.1   Generating Initial Causal Pathways

This subsection first introduces the representation of causal pathways in the ST space, and then elaborates how to generate initial causal pathways.

DEFINITION 9   (GAUSSIAN BAYESIAN NETWORK (GBN)). *GBN is a special form of Bayesian network for probabilistic inference with continuous Gaussian variables in a DAG, in which each variable is assumed as linear function of its parents [32].*

As shown in Fig. 6, the ST causal relations of air pollutants are encoded in a GBN-based graphical model, to represent both local and ST dependencies. Here we choose GBN to model the causalities because: 1) GBN provides a simple way to represent the dependencies among multiple pollutants variables, both locally and in the ST space. 2) GBN models continuous variables rather

than discrete values. Due to the sensors monitor the concentration of pollutants per hour, GBN could help better capture the fine-grained knowledge through the dependencies of these continuous values. In this subsection, based on the extracted matched patterns and candidate sensors from the *pattern mining module* for each pollutant $\hat{P}_{c_m s_n}$, we use $P_{c_m s_n}$ to represent continuous values in the graphical model. 3) The characteristics of urban data fit the GBN model well. As shown in Fig. 7, the distribution of 1-hour difference (current value minus the value 1-hour ago) of air pollutants and meteorological data obey Gaussian distribution (verified by $D'Agostino - Pearson$ test [33][34]). In the following sections, normalized 1-hour differences of time series data will be used as inputs for the model.
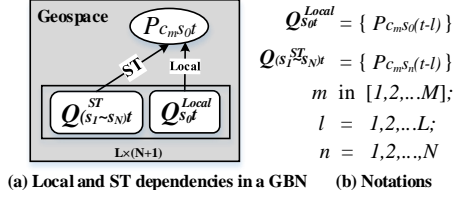

$$Q^{Local}_{S_0 t} = \{ P_{c_m s_0(t-l)} \}$$
$$Q^{ST}_{(s_1 \sim s_N)t} = \{ P_{c_m s_n(t-l)} \}$$
$$m \text{ in } [1,2,...M];$$
$$l = 1,2,...L;$$
$$n = 1,2,...,N$$

(a) Local and ST dependencies in a GBN   (b) Notations

Figure 6: GBN-based causal pathway representation and its notations.

(a) Original values normalized by standard deviation   (b) Value of 1-hour difference normalized by standard deviation
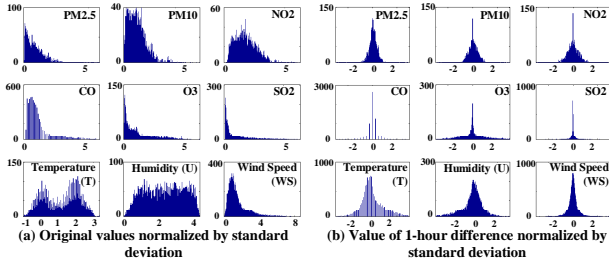
Figure 7: Histograms of urban data (original vs. 1-hour difference)

Specifically, for the target pollutant c_m at sensor s_0-th sensor and timestamp t, denoted as $P_{c_m s_0 t}, m \in [1, M]$, we capture the dependencies from both the local causal pollutants $Q^{Local}_{s_0 t}$ and the ST causal pollutants $Q^{ST}_{(s_1 \sim s_N)t}$. Here $Q^{ST}_{(s_1 \sim s_N)t}$ refer to a $1 \times NL$ vector of pollutants at N neighborhood sensors $s_1 \sim s_N$ and previous L timestamps that most probably cause the target pollutant in the ST space, i.e. $Q^{ST}_{(s_1 \sim s_N)t} = \{P_{c_{m_n} s_n(t-l)}\}, m \in [1, \ldots, M]; n = 1, \ldots, N; l = 1, \ldots, L$. In order to better trace the most likely "causers" spatially, we just preserve the one category of pollutant at each neighborhood sensor that most influences the target pollutant. We use c_{m_n} to represent the category for the most likely "causers" at sensor n. Similarly, $Q^{Local}_{s_0 t}$ is a $1 \times ML$ vector of pollutants locally at s_0. For example, when we set $L = 2, M = 6, Q^{Local}_{s_0 t}$ may take values of 12 normalized 1-hour difference time series data, i.e. $Q^{Local}_{s_0 t} = (2, -0.5, 0.8, 0.3, 1, -2, 2.2, 1, 1, 0, -0.5, 0.2)$.

The parents of $P_{c_m s_0 t}$ are denoted as $\boldsymbol{PA}(P_{c_m s_0 t}) = Q^{Local}_{s_0 t} \oplus Q^{ST}_{(s_1 \sim s_N)t}$, where $\oplus$ denotes the concatenation operator for two vectors. Based on the definition of GBN, the distribution of $P_{c_m s_0 t}$ conditioned on $\boldsymbol{PA}(P_{c_m s_0 t})$ obeys Gaussian distribution:

$$Pr(P_{c_m s_0 t} = p_{c_m s_0 t} | \boldsymbol{PA}(P_{c_m s_0 t})) \sim \mathcal{N}(\mu_{c_m s_0 t} +$$
$$\Sigma^N_{n=0} \Sigma^L_{l=1} a_{m_n}(nL+l)(p_{c_m s_n(t-l)} - \mu_{c_m s_n(t-l)}), \Sigma(\epsilon_{c_m s_0 t}))$$
(1)

$\mu_{c_m s_0 t}$ is the marginal mean for $P_{c_m s_0 t}$. Σ denotes the covariance operator. $\boldsymbol{A} = \{a_{m_n}(nL + l)\}, (m_n \in [1, \ldots, M]; n = 0, 1, \ldots, N; l = 1, \ldots, L)$ is the coefficient for the linear regression in GBN [32]:

To minimize the uncertainty of $P_{c_m s_0 t}$ given its parents, we need to find N sensors $s_1 \sim s_N$ from the ST space and the parameters $\boldsymbol{A}$ that minimize the error:

$$\Sigma(\epsilon_{c_m s_0 t}) = \Sigma(P_{c_m s_0 t}) - \boldsymbol{A}\Sigma(\boldsymbol{PA}(P_{c_m s_0 t}))^{-1}\boldsymbol{A}^T \quad (2)$$

Generating the initial causal pathways requires locating N most influential sensors from $|\mathcal{S}|$ sensors with up to $\binom{|\mathcal{S}|}{N}$ trials. Yet given the candidate sensors selected by Section 3.3, we manage to search from a subset $(X \leq |\mathcal{S}|)$ sensors with time efficiency and scalability. We further propose a Granger causality score GC_{score} to generate initial causal pathways, which is defined as:

$$GC_{score}(m, s_0, s_n) = max_{m_n \in [1,M]} max_{l \in [1,L]}$$
$$\{|match(t_{(c_m, s_0)}, t_{(c_{m_n}, s_n)})| \cdot \frac{|\Sigma(\epsilon_{c_m s_0(t-l)})_1| - |\Sigma(\epsilon_{c_m s_0(t-l)})_2|}{|\Sigma(\epsilon_{c_m s_0(t-l)})_2| \chi^2_L(0.05)}\}$$
(3)

where GC_{score} is a χ^2-test score [21] for the predictive causality, with higher score indicating more probable "Granger" causes from M pollutants at sensor s_n to the target pollutant c_m at sensor s_0 [17] ($GC_{score} \leq 1$ means none causality). For variables obeying Gaussian distribution, Granger causality is in the same form as conditional mutual information [20], which has been used successfully for constructing structures for Bayesian networks. Here $|match(t_{(c_m, s_0)}, t_{(c_{m_n}, s_n)})|$ is the number of matched timestamps of FEPs between two time series (pollutant c_{m_n} at sensor s_n and pollutant c_m at sensor s_0, see Section 3.3). And $\Sigma(\epsilon_{c_m s_0(t-l)})_1$ and $\Sigma(\epsilon_{c_m s_0(t-l)})_2$ correspond to the variances of the target pollutant $P_{c_m s_0 t}$ conditioned on lagged sequences $Q^{Local}_{s_0(t-l)}$ and $Q^{Local}_{s_0(t-l)} \oplus Q^{ST}_{s_n(t-l)}$.

4.2 Integrating Confounders

Recall the example in Fig. 1. A target pollutant is likely to have several different causal pathways under different environmental conditions, which indicate the causal pathways we learn may be biased and may not reflect the real reactions or propagations of pollutants. To overcome this, it is necessary to model the environmental factors (humidity, wind, etc.) as extraneous variables in the causality model, which simultaneously influence the cause and effect. For example, when the wind speed is less than 5m/s, city A's PM2.5 could be the "cause" of city B's PM10. However, when the wind speed is more than 5m/s, there may not be causal relations between the two pollutants in the two cities. In this subsection, we will elaborate how to integrate the environmental factors into the GBN-based graphical model, to minimize the biases in causality analysis and guarantee the causal pathways are faithful for the government's decision making. We first introduce the definition of confounder and then elaborate the integration.

DEFINITION 10   (CONFOUNDER). *A confounder is defined as a third variable that simultaneously correlates with the cause and effect, e.g. gender K may affect the effect of recovery P given a medicine Q, as shown in Fig. 8(a). Ignoring the confounders will lead to biased causality analysis. To guarantee an unbiased causal inference, the cause-and-effect is usually adjusted by averaging all the sub-classification cases of K [5], i.e. $Pr(P|do(Q)) = \Sigma^K_{k=1} Pr(P|Q, k)Pr(k)$.*
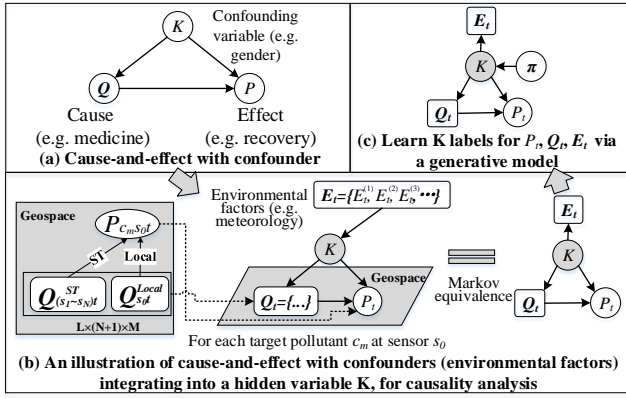
Figure 8: The GBN-based graphical model, integrating confounders to the causal pathway, and converting the model into a generative model
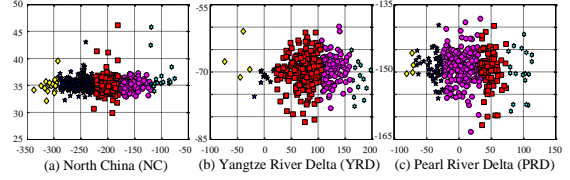

Figure 9: 2-D PCA projections of 5 clusters of meteorological data in NC, YRD and PRD. The original meteorological data contains five types, i.e. temperature (T), pressure (P), humidity (U), wind speed (WS), and wind direction (WD), with each region divided into 9 grids, thus 45-dimensional.

For integrating environmental factors as confounders, denoted as $\boldsymbol{E_t} = \{E_t^{(1)}, E_t^{(2)}, \dots\}$, into the GBN-based causal pathways, one challenge is there can be too many sub-classifications of environmental statuses. For example, if there are 5 environmental factors and each factor has 4 statuses, there will exist $4^5 = 1024$ causal pathways for each sub-classification case. Directly integrating $\boldsymbol{E_t}$ as confounders to the cause and effect will result in unreliable causality analysis due to very few sample data conditioned on each sub-classification case. Therefore, we introduce a discrete hidden confounding variable K, which determines the probabilities of different causal pathways from $\boldsymbol{Q_t}$ to $\boldsymbol{P_t}$, as shown in Fig. 8(b). The environmental factors $\boldsymbol{E_t}$ are further integrated into K, where $K = 1, 2, ...K$. In this ways, the large number of sub-classification cases of confounders will be greatly reduced to a small number K, as K clusters of the environmental factors.

Based on Markov equivalence (DAGs which share the same joint probability distribution [35]), we can reverse the arrow $\boldsymbol{E_t} \to K$ to $K \to \boldsymbol{E_t}$, as shown in the right part of Fig. 8(b). K determines the distributions of $P, \boldsymbol{Q_t}, \boldsymbol{E_t}$, thus enabling us to learn the distribution of the graphical model from a generative process. To help us learn the hidden variable K, the generative process further introduces a hyper-parameter $\boldsymbol{\pi}$ (as shown in Fig. 8(c)) that determines the distribution of K. Thus the graphical model can be understood as a mixture model under K clusters. We learn the parameters of the graphical model by maximizing the new log likelihood:

$$LL^{gen} = \Sigma_t \Sigma_{k=1}^{K} ln(Pr(p_t|\boldsymbol{q_t}, k) Pr(\boldsymbol{e_t}|k) Pr(k|\boldsymbol{\pi})) \quad (4)$$

In determining the number of the hidden variable K, we do not consider too large K values since that will induce much complexity for causality analysis. Also a too small K may not characterize the information contained in the confounders (i.e. meteorology). We observe the 2-D PCA projections of meteorological data (as shown in Fig. 9). In three regions, five clusters can characterize the data sufficiently well. Thus we choose K = 3 ∼ 7 for learning in practice.

4.3 Refining Causal Structures

This subsection tries to refine the causal structures and obtain the final causal structures under K clusters. The refining process includes two phases in each iteration: 1) an EM learning (EML) phase to infer the parameters of the model, and 2) a structure reconstruction (SR) phase to re-select the top N neighborhood sensors

based on the newly learnt parameters and GC_{score}, as illustrated in Algorithm 2.

EML (line 6-18) is an approximation method to learn the parameters $\boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{A_k}, \boldsymbol{B_k}$ of the graphical model, by maximizing the log likelihood (Equation 4) of the observed data sets via an E-step and a M-step. Here $\boldsymbol{\pi}$ contains the hyper parameters which determine the distribution of K (T × K-dimensional). $\boldsymbol{\gamma}$ are posterior probabilities for each monitoring record (T × K-dimensional). $\boldsymbol{A_k}, \boldsymbol{B_k}$ are parameters for measuring the dependencies among pollutants and meteorology (K-dimensional). Note that $\boldsymbol{A_k}, \boldsymbol{B_k}$ come in different formats. $\boldsymbol{A_k}$ is the regression parameter for:

$$P_{c_m s_0 t} = \mu_0 + (\boldsymbol{Q_{s_0 t}^{Local}} \oplus \boldsymbol{Q_{(s_1 \sim s_N)t}^{ST}})\boldsymbol{A_k} + \epsilon_{c_m s_0 t} \quad (5)$$

and $\boldsymbol{B_k} = (\boldsymbol{\mu_{B_k}}, \boldsymbol{\Sigma_{B_k}}) = (mean(\boldsymbol{E_t}), std(\boldsymbol{E_t}))$ includes the parameters for the multivariate Gaussian distribution of environmental factors $\boldsymbol{E_t}$. In the E-step, we calculate the expectation of log likelihood (Equation 6) with the current parameters, and the M-step re-computes the parameters.

E-step: Given the parameters $\boldsymbol{\pi}$, K, N, $\boldsymbol{A_k}, \boldsymbol{B_k}$, EM assumes the membership probability γ_{tk}, i.e., the probability of $p_t, \boldsymbol{q_t}, \boldsymbol{e_t}$ belonging to the k-th cluster as:

$$\begin{aligned} \gamma_{tk} = Pr(k|p_t, \boldsymbol{q_t}, \boldsymbol{e_t}) &= \frac{Pr(k)Pr(p_t, \boldsymbol{q_t}, \boldsymbol{e_t}|k)}{Pr(p_t, \boldsymbol{q_t}, \boldsymbol{e_t})} \\ &= \frac{\pi_{tk}\mathcal{N}(p_t|\boldsymbol{q_t}, \boldsymbol{A_k})\mathcal{N}(\boldsymbol{e_t}|\boldsymbol{B_k})}{\Sigma_{j=1}^{K}\pi_{tj}\mathcal{N}(p_t|\boldsymbol{q_t}, \boldsymbol{A_j})\mathcal{N}(\boldsymbol{e_t}|\boldsymbol{B_j})} \end{aligned} \quad (6)$$

M-step: The membership probability γ_{tk} in E-step can be used to calculate new parameter values $\boldsymbol{\pi^{new}}, \boldsymbol{A_k^{new}}, \boldsymbol{B_k^{new}}$. We first determine the most likely assignment tag of timestamp t to cluster k, i.e.

$$Tag_t = max_{k \in [1,K]}\pi_{tk} \quad (7)$$

By integrating the timestamps belonging to each cluster k, we can update $\boldsymbol{A_k^{new}}$ by Equation 5. Then we update $\boldsymbol{B_k}$ by:

$$\begin{aligned} \boldsymbol{\mu_{B_k}^{new}} &= \frac{1}{T_k}\Sigma_{t=1}^{T}\gamma_{tk}\boldsymbol{e_t}, T_k = \Sigma_{t=1}^{T}\gamma_{tk} \\ \boldsymbol{\Sigma_{B_k}^{new}} &= \frac{1}{T_k}\Sigma_{t=1}^{T}\gamma_{tk}(\boldsymbol{e_t} - \boldsymbol{\mu_{B_k}^{new}})(\boldsymbol{e_t} - \boldsymbol{\mu_{B_k}^{new}})^T \end{aligned} \quad (8)$$

In addition, we update π_{tk}^{new} by:

$$\pi_{tk}^{new} = \frac{\gamma_{tk}}{T_k} \quad (9)$$

The SR phase (line 19-24) utilizes the parameters provided by the EM learning phase, and re-select the top N neighborhood sensors based on the newly generated GC_{score} for each cluster k. We present a training example (as shown in Fig. 10(a)) of learning the causal pathways for Beijing PM2.5 during Jan−Mar. After 20 training iterations of the EM learning phase and structure reconstruction, we finally obtain K = 4 causal structures under each cluster, with the log likelihood shown in Fig. 10(b). We find the log likelihood does not increase much after 10 iterations, thus we set the iteration number to 10 in our experiments. For the last iteration, we calculate the percentage of labeled timestamps belonging to each cluster k. In this example, we find that Beijing's PM2.5 is more likely to be influenced by NO2 in Baoding and PM10 in Cangzhou.

Algorithm 2: Refining the causal structures for each target pollutant c_m at location s_0.

Input: T, K, N, and raining data sets $p_t, \boldsymbol{q_t}, \boldsymbol{e_t}, t \in [1, T]$
Output: Refined causal structures for K clusters

1 Initial neighborhood sensors $s_1 \sim s_N$ based on top N GC_{score};
2 **repeat**
3   EML$(P_t, \boldsymbol{Q_t}, \boldsymbol{E_t}, s_1 \sim s_N, \text{K})$
        $\rightarrow Log\_likelihood, \pi_{tk}, \gamma_{tk}, \boldsymbol{A_k}, \boldsymbol{B_k}$;
4   SR$(\boldsymbol{A_k}, s_1 \sim s_N, \text{K}) \rightarrow s_1' \sim s_N', Q'$;
5 **until** *Log_likeoihood converges*;
6 **Function** *EM_Learning(EML)$(P_t, \boldsymbol{Q_t}, \boldsymbol{E_t}, s_1 \sim s_N, \text{K})$*
7   **repeat**
8   InitialAssign: K clusters via K-means$(\boldsymbol{E_t})$
9   **foreach** *item $t = 1$ to T* **do**
10   **foreach** *item $k = 1$ to K* **do**
11   Update π_{tk} by Equation (9);
12   **foreach** *item $k = 1$ to K* **do**
13   Update $\boldsymbol{A_k}, \boldsymbol{B_k}$ by Equation (5),(8);
14   **foreach** *item $t = 1$ to T* **do**
15   **foreach** *item $k = 1$ to K* **do**
16   Update γ_{tk} by Equation (6);
17   **until** *Log likelihood converges*;
18   return: $Log\_likelihood$ and $\pi_{tk}, \gamma_{tk}, \boldsymbol{A_k}, \boldsymbol{B_k}$;
19 **Function** *Structure_Reconstruction(SR)$(\boldsymbol{A_k}, s_1 \sim s_N, \text{K})$*
20   **foreach** *item s_n in All candidate sensors* **do**
21   Compute $GC_{score}(m, s_0, s_n)$ for $s_1 \sim s_N$;
22   Rank GC_{score} and re-select the top N neighborhood sensors $s_1' \sim s_N'$;
23   Update $Q \rightarrow Q'$ corresponding to $s_1' \sim s_N'$;
24   return: $s_1' \sim s_N', Q'$;

---

5. EXPERIMENTS

We evaluate the empirical performance of our method in this section. All the experiments were conducted on a computer with Intel Core i5 3.3Ghz CPU and 16GB memory. We use MATLAB for our Bayesian learning module, and the open-source MATLAB BNT toolbox [36] for baseline methods.

5.1 Experimental Setup

5.1.1 Data Sets

We use three data sets that contain the records of 6 air pollutants and 5 meteorological measurements:

• North China (NC), with 61 cities, 544 air quality monitoring sensors and 404 meteorological sensors in North China. The latitude and longitude ranges are 34N-43N, 110E-123E.

• Yangtze River Delta (YRD), with 49 cities, 330 air quality monitoring sensors and 48 meteorology sensors. The latitude and longitude ranges are 28N-35N, 115E-123E, respectively.

• Pearl River Delta (PRD), with 18 cities, 124 air quality monitoring sensors and 406 meteorology sensors. The latitude and longitude ranges are 22N-25N, 110E-116E.

The 6 air pollutants are PM2.5, PM10, NO_2, CO, O_3, SO_2, and the 5 meteorological measurements are temperature (T), pressure (P), humidity (H), wind speed (WS), and wind direction (WD), which are updated hourly. The time span for all data sets is from 01/06/2013 to 31/12/2016. We separate each data set into four groups based on four seasons, and use the last 15 days in each season in year 2014, 2015, 2016 for testing, and the remaining data for model training. The total numbers of training timestamps are 5424, 6193, 7753, 7752 in the four seasons, respectively, and the number of the corresponding testing timestamps is $15 \times 24 \times 3 = 1080$ in each season. To get the environmental factors $\boldsymbol{E_t}$ for the coupled model, we divide each region into 3×3 grids and average the meteorology values within each grid.

We conduct experiments at both city level (Section 5.2.2, 5.2.1, 5.2.5) and sensor level (Section 5.2.3). The city-level experiments average value of the sensors in the city to form a pseudo sensor, and discover the pathways among all the cities in three data sets. The sensor-level experiments analyze the causal relationships among sensors in each data set.

5.1.2 Baselines

Since Bayesian-based methods have been well used to learn causal Bayesian structures [23], we choose the most commonly used BN structure learning approaches as baselines to compare with our method. To identify the dependencies among different pollutants, the baselines are deployed to learn the causal structures for each target pollutant.

1. MWST. Maximum Weighted Spanning Tree (MWST) generates an undirected tree structure based on the MWST algorithm [37]. Each time it connects one edge between two nodes with the maximum mutual information. Furthermore, [38] proposed an independency test method to assign a direction to each edge in the tree structure.

2. MCMC. Markov-chain Monte Carlo (MCMC) is a statistical method that also samples from the Directed Acyclic Graph (DAG) space [39]. The method maximizes the score from a set of similar DAGs that add, delete, or reverse connections, and updates the structure in the next iteration.

3. K2+PS. $K2$ is a widely used greedy method for Bayesian structure learning, which selects at most N parents based on the $K2$ score [40] for each variable given the updating order of all the variables. In our case, we use pattern search algorithm [41] to optimize the updating order, thus reducing the search space of casual pathways of different pollutants. Note that the original $K2$ score is defined for discrete variables. Here we use GC_{score} instead for the continuous variables.

4. CGBN. Coupled Gaussian Bayesian network [6] is a data-driven causality model considering the dependencies between both the air pollutants and meteorology. CGBN assumes there is a third variable (confounder, such as gender as a confounder to evaluate the effect of a medicine on a disease) which simultaneously influences the dependences among pollutants and among environmental factors, coupling pollutants and environmental factors together. The difference between CGBN and our approach is that 1) our approach integrates the environmental factors directly into the graphical model, instead of through coupling, and 2) our approach has a pattern mining module and a refining algorithm to optimize the
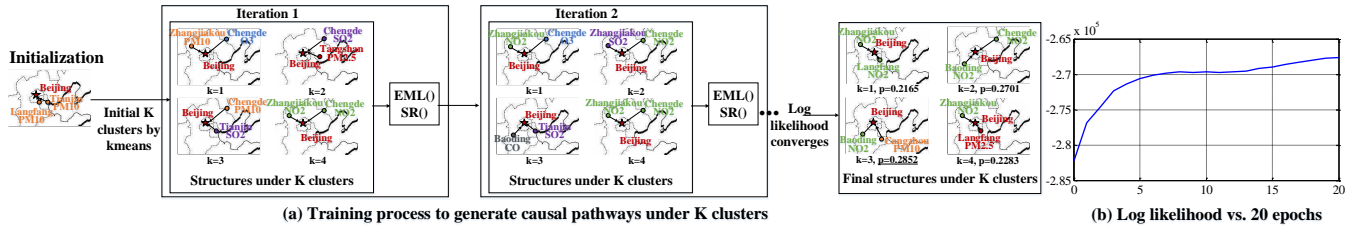
(a) Training process to generate causal pathways under K clusters

(b) Log likelihood vs. 20 epochs

Figure 10: An example of learning the causal pathway for PM2.5, Jan−Mar in Beijing under $K = 4$ clusters.

learning process.

5.1.3 Parameter Setting

The parameters of *pg-Causality* include: (1) the support threshold σ; (2) the temporal constraint Δt; (3) the distance threshold δ_g for finding candidate causers; and (4) the correlation threshold δ_p for finding candidate causers; (5) the number of time lags L = 3; (6) and the number of pollutant categories M = 6. When finding causal pathways at city level, we set $\sigma = 0.1$, $\Delta t = 1$ hour, $\delta_g = 200$ km, and $\delta_p = 0.5$. At the station level, all the the parameters are set the same except that $\delta_g = 15$ km to impose a finger granularity for finding candidate causers. K and N are evaluated within the range K = 3 ∼ 7, and N= 1 ∼ 5.

5.2 Experimental Results

The verification of causality is a very critical part in causal modelling. The simplest method for evaluating causal dependence is to intervene in a system and determine if the model is accurate under intervention. However, substantial and direct intervention in air pollution is impossible. By investigating the verification methods in previous causality works, we propose five tasks to evaluate the effectiveness of our approach, namely, 1) inference accuracy for a 1-hour prediction task, 2) time efficiency, 3) scalability, 4) verification on synthetic data, and 5) visualizing the causal pathways. Tasks 1-3 target to evaluate whether the model fits the dependences among the datasets well. Task 4 tries to learn the causal pathways for a predefined causal structure generated by synthetic datasets. And Task 5 targets at the interpretability of the causal pathways we learn.

5.2.1 Inference Accuracy

We first evaluate the effectiveness of our approach via the causal inference accuracy through the causal pathways at city level, which is a 1-hour prediction task based on our proposed GBN-based graphical model. Note this prediction task is not general for all the timestamps, it only predicts the future 1-hour based on the extracted pattern-matched periods, indicating the causal inference for the frequent evolving behaviors. Specifically, we first infer the probability $Pr(k)$ of the testing data belonging to cluster k. Then, we use the structure and parameters from the trained causal pathways regarding this cluster to estimate the future pollutant concentration by Eq. 10.

$$P_{c_m s_0 t}^{est} = \Sigma_{k=1}^{K}(\mu_{0k} + \boldsymbol{PA}(P_{c_m s_0 t})\boldsymbol{A_k})Pr(k) \quad (10)$$

The accuracy is defined as $\Sigma_{t=1}^{T_{test}}(P_{c_m s_0 t}^{est} - P_{c_m s_0 t}^{*})/P_{c_m s_0 t}^{*}T_{test}$, where $P_{c_m s_0 t}^{*}$ is the ground truth value and T_{test} is the number of test cases. TABLE 3 shows the 1-hour prediction accuracy for PM2.5 and PM10 with our approaches *pg-Causality*, *pg-Causality-n*, *pg-Causality-p*, and the three baseline methods in Beijing (Region NC), Shanghai (Region YRD), and Shenzhen (Region PRD). Here *pg-Causality-n* represents *pg-Causality* without the pattern

mining module, and *pg-Causality-p* represents *pg-Causality* without integrating confounders. The accuracy shown in TABLE 3 is the accuracy for spring for three cities. The *pg-Causality* gets the highest accuracy (92.5%, 93.78%, 95.39% for PM2.5 in Beijing, Shanghai, and Shenzhen, respectively; 91.36%, 92.39%, 93.18% for PM10, repectively.), compared to *pg-Causality-n* and *pg-Causality-p*, as well as the three baseline methods WMST, K2+PS, and CGBN. We did not include the accuracy of MCMC in TABLE 3 due to its unbearably high computational time. The accuracy for MCMC is lower than 60%, which is not competitive with the other methods mentioned. The highest inference accuracy for the three cities are marked with three different colors (orange for Beijing, blue for Shanghai, and green for Shenzhen) given different parameters K and N. K and N are obtained based on the maximum inference accuracy for each city. We note N = 2, K = 4 provides the best performance for Beijing, while N = 0, K = 5 or 6 generate the best accuracy for Shanghai and N = 0, K = 1 for Shenzhen. The optimal number N = 2 for Beijing also suggests that the air pollution is mainly influenced by the most influential sensors in the ST space. While the optimal number N = 0 for Shanghai and Shenzhen suggests that the PM2.5 in these two cities are mainly influenced by historical pollutants locally.

We also evaluate the 1-hour prediction accuracy with three well-used time series model, i.e., auto-regression moving average (ARMA) model, linear regression model (LR), and support vector machine for regression with a Gaussian radial basis function (rbf) kernel (represented as SVM-R). Generally, *pg-Causality* demonstrates higher inference accuracy compared with these time series models, except for the PM2.5 in Shanghai.

5.2.2 Time efficiency

We also compare the training time of *pg-Causality* with baseline methods, as shown in TABLE 4. Since our approach consists of both pattern mining and Bayesian learning modules, we present the averaged time consumption of training all the three data sets, for each step in the two modules. We also evaluate the overall time consumption of *pg-Causality* and *pg-Causality-n* without the pattern mining module (Section 5.1 (p+g) refers to the time cost of causal structure initialization with both pattern mining and Granger causality score. Section 5.1 (g) refers to only using Granger causality score). Results show that our approach is very efficient, with the second minimum computation time among all the methods. MWST consumes the minimal time, however, it does not generate satisfactory accuracy for prediction (as in Section 5.2.1). We thus consider our approach provides the best trade-off regarding accuracy and time efficiency.

5.2.3 Scalability

Another superior characteristic of our approach is the scalability. We further identify the causal pathways for air pollutants at sensor level, which is more than ten times as large as in the city-

Table 3: Accuracy of PM2.5/PM10 1-hour prediction vs. baselines, Beijing, Shanghai, and Shenzhen.

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.9174	0.9075	0.9067	0.9149	0.9134	0.9132									
K = 2	0.9164	0.9059	0.8987	0.9168	0.9180	0.9211									
K = 3	0.9162	0.9089	0.9216	0.9177	0.9236	0.9179	0.925 (K=4, N=2)	0.9174 (K=1, N=0)	0.9105	0.691	0.9236	0.801	0.8756	0.9048	0.9157
K = 4	0.9148	0.9127	0.9250	0.9155	0.9209	0.9216									
K = 5	0.9123	0.9214	0.9244	0.9081	0.9198	0.9153									
K = 6	0.9144	0.9190	0.9238	0.9195	0.9193	0.9189									
K = 7	0.9129	0.9162	0.9157	0.9229	0.9201	0.9201									

Beijing PM2.5, 1-hour prediction accuracy

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.8958	0.8936	0.8951	0.9066	0.9035	0.9060									
K = 2	0.8989	0.8981	0.8990	0.9070	0.9123	0.9118									
K = 3	0.8996	0.8985	0.8992	0.9016	0.9107	0.9094	0.9136 (K=4, N=3)	0.9003 (K=7, N=1)	0.8857	0.653	0.9131	0.842	0.8561	0.8932	0.8977
K = 4	0.8990	0.8984	0.8980	0.9136	0.9111	0.9119									
K = 5	0.8995	0.9008	0.9015	0.9059	0.9017	0.9108									
K = 6	0.8985	0.9061	0.9061	0.9079	0.9097	0.9095									
K = 7	0.8998	0.8991	0.9012	0.9134	0.9096	0.9127									

Beijing PM10, 1-hour prediction accuracy

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.9375	0.9356	0.9356	0.9356	0.9355	0.9355									
K = 2	0.9372	0.9349	0.9359	0.9363	0.9358	0.9332									
K = 3	0.9375	0.9373	0.9314	0.9323	0.9327	0.9301	0.9378 (K=5, N=0)	0.9378 (K=1, N=0)	0.8928	0.667	0.9376	0.751	0.9209	0.9378	0.9381
K = 4	0.9377	0.9345	0.9328	0.9296	0.9330	0.9325									
K = 5	0.9328	0.9328	0.9339	0.9337	0.9335	0.9263									
K = 6	0.9378	0.9328	0.9342	0.9341	0.9315	0.9323									
K = 7	0.9372	0.9335	0.9303	0.9316	0.9304	0.9289									

Shanghai PM2.5, 1-hour prediction accuracy

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.9239	0.9229	0.9232	0.9226	0.9226	0.9227									
K = 2	0.9231	0.9212	0.9226	0.9149	0.9204	0.9201									
K = 3	0.9228	0.9198	0.9203	0.9194	0.9188	0.9186	0.9239 (K=1, N=0)	0.9239 (K=1, N=0)	0.9173	0.631	0.9239	0.835	0.9042	0.9238	0.9239
K = 4	0.9227	0.9201	0.9172	0.9183	0.9157	0.9174									
K = 5	0.9215	0.9199	0.9199	0.9147	0.9183	0.9167									
K = 6	0.9226	0.9183	0.9171	0.9187	0.9178	0.9180									
K = 7	0.9236	0.9177	0.9170	0.9167	0.9186	0.9154									

Shanghai PM10, 1-hour prediction accuracy

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.9539	--	--	--	--	--									
K = 2	0.9533	--	--	--	--	--									
K = 3	0.9535	--	--	--	--	--	0.9539 (K=1, N=0)	0.954 (K=1, N=0)	0.9006	0.658	0.9539	0.713	0.9097	0.9482	0.9484
K = 4	0.9534	--	--	--	--	--									
K = 5	0.9532	--	--	--	--	--									
K = 6	0.9524	--	--	--	--	--									
K = 7	0.9530	--	--	--	--	--									

Shenzhen PM2.5, 1-hour prediction accuracy

Accuracy	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	pg-Causality (Optimal K, N)	pg-Causality-n	pg-Causality-p	MWST	CGBN	K2+PS	ARMA	LR	SVM-R
K = 1	0.9315	0.9307	0.9318	0.9299	0.9296	0.9300									
K = 2	0.9308	0.9291	0.9294	0.9294	0.9290	0.9286									
K = 3	0.9309	0.9289	0.9304	0.9289	0.9292	0.9297	0.9318 (K=1, N=2)	0.9315 (K=1, N=0)	0.9226	0.694	0.9315	0.853	0.95	0.9250	0.9275
K = 4	0.9307	0.9303	0.9311	0.9301	0.9279	0.9286									
K = 5	0.9313	0.9307	0.9263	0.9297	0.9302	0.9283									
K = 6	0.9303	0.9311	0.9284	0.9282	0.9263	0.9268									
K = 7	0.9301	0.9295	0.9305	0.9299	0.9261	0.9254									

Shenzhen PM10, 1-hour prediction accuracy

Table 4: Computation time for training data sets at city level.

Time (s)	m = 1	m = 2	m = 3	m = 4	m = 5	m = 6
Section 4.1 - 4.2	2.74	3.49	3.74	3.74	3.94	3.71
Section 4.3	29.88	43.28	55.15	39.07	45.44	43.13
Section 5.1 (p + g)	73.43	111.33	151.63	94.56	136.64	128.47
Section 5.1 (g)	1125.51	1076.97	1068.13	1074.94	1057.67	1082.85
Section 5.2	--	--	--	--	--	--
Section 5.3	38421.53	42094.47	39137.81	44162.31	49192.68	44601.73
pg-Causality	38527.58	42252.57	39348.57	44299.68	49378.7	44777.04
pg-Causality-n	39547.04	43171.44	40205.94	45237.25	50250.35	45684.58
MWST	6357.9	6529.88	6605.31	7033.58	7216.45	7374.13
CGBN	72785.54	79165.28	80356.3	75578.74	79623.57	78191.32
MCMC	524731.63	562835.19	--	--	--	--
K2 + PS	286592.52	324851.47	--	--	--	--

level analysis. Our approach provides linear scalability in time with 11.6 hours training time at city level for 128 cities, and 126 hours at sensor level for 982 stations. We here claim linear scalability since we did not try to find the optimal causal structure by searching the DAG space, which is an NP hard problem and in the worst case requires $2^{O(n^2)}$ searches [7]. In this paper, the causal pathways we learnt are based on greedy-based approximation. For the structure learning algorithm, we assume the number of parameters of the Bayesian-based graphical model to be (#), and the training iterations to be N_{iter}. For totally N sensors in the geospace and T timestamps in the training records, the time cost for the EM learning (EML) phase is $O(N_{iter} \times (\#) \times N \times T)$, assuming every parameter is updated once for every record. In addition, the time cost for the structure reconstruction (SR) phase is

$O(N_{iter} \times X \times L \times N \times T + N_{iter} \times K \times (\#) \times N \times T)$, where X is the candidate "causers" selected by pattern mining and L is the number of time lags. Thus the overall training time is $N_{iter}O(XL + (1 + K)(\#))NT$. If the number of the graphical model (#) is fixed, the computation time will approximately be at linear scalability with the sensor number N and timestamps number T. We verified the linear scalability in Fig. 11(b)(c). For the baseline methods, $MCMC$ even cannot compute such large data sets. $CGBN$ and $K2 + PS$ are unable to compute within 10 days and we leave their time cost as blank, as shown in 11(c). Meanwhile, the accuracy is guaranteed when extending city-level data to sensor-level data, as shown in Fig. 11(a).

5.2.4 Verification with Synthetic Data

Since the verification of causality via prediction task may not fully reflect the cause-and-effect relationships learned by the model, we further conduct experiments with synthetic data to judge whether the causality identification is correct or not.

As shown in Fig. 12, we generate N = 20 time series, with the pre-defined causal structure as in Fig. 12(a). This is done by randomly choosing the lag k for any edge x → y in the feature causal graph [22]. To imitate the confounding effect, one time series is selected to influence all other time series. We reconstruct the causal structures through Granger causality (as shown in Fig. 12(b)), lasso Granger causality (as shown in Fig. 12(c) [22]), and pg-Causality (as shown in Fig. 12(d)). To fit pg-Causality in this "toy" model, we simplified the model by randomly assigning locations to N time series. In the meanwhile, we set the distance constraint for selecting candidate "causers" to infinity, in order to consider every pair
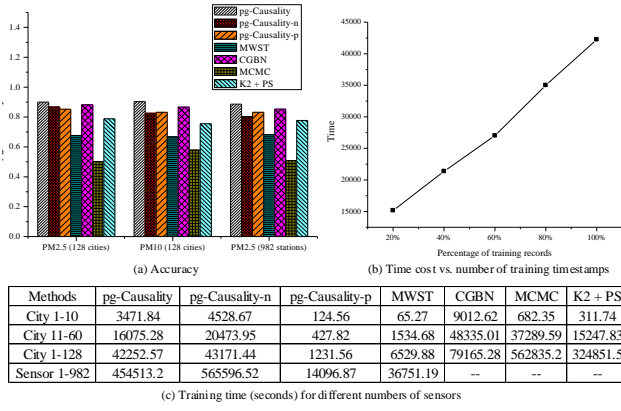
Methods	pg-Causality	pg-Causality-n	pg-Causality-p	MWST	CGBN	MCMC	K2 + PS
City 1-10	3471.84	4528.67	124.56	65.27	9012.62	682.35	311.74
City 11-60	16075.28	20473.95	427.82	1534.68	48335.01	37289.59	15247.83
City 1-128	42252.57	43171.44	1231.56	6529.88	79165.28	562835.2	324851.5
Sensor 1-982	454513.2	565596.52	14096.82	36751.19	--	--	--

(c) Training time (seconds) for different numbers of sensors

Figure 11: Accuracy and time efficiency at city and station level.

of causal relations between N time series. We mark the incorrect constructed edges in red. Result shows that pg-Causality generates the most likely structure compared with the baseline structure.
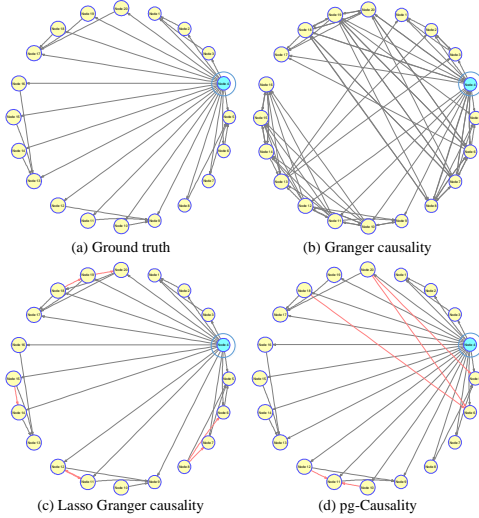

Figure 12: Causal structures generated by 20 synthetic time series. (a) Original structure with Node 4 (blue node, surrounded by a circle outside) as confounder, (b) Reconstructed by Granger causality, (c) Reconstructed by Lasso Granger causality, (d) Reconstructed by simplified pg-Causality. (Since the causa structure reconstructed by Granger causality in (b) significantly differs from the original one in (a), we only mark the incorrect connections for Lasso Granger causality and pg-Causality in red in (c) and (d).)

5.2.5 Case Study

To analyze the causal pathways for air pollutants, we study two cases corresponding to PM2.5 in specific cities. First we analyze the causal pathways for PM2.5 in the spring of Beijing and in the winter of Shanghai, the period of which are considered as the most heavily polluted season. Then we analyze Beijing PM2.5 before and during the APEC period (1st − 14th, Nov, 2014) as a case study for human intervention in causal systems.

1. Beijing and Shanghai. Fig. 10 is a real example for the

causal pathways for Beijing PM2.5 during Jan−Mar. We provide the probability for each causal pathway for each cluster, defined as the proportion of labeled timestamps that belong to each cluster. As shown in Fig. 10(a), Cluster 3 takes a relatively higher proportion (28.52%) of time for Beijing PM2.5, indicating the causal pathway during Jan−Mar more probably come from southern sensors, i.e. Baoding and Cangzhou. Actions can be taken to control these pollutants in these cities. We then present the causal pathways for PM2.5 in Shanghai, during Oct−Dec, which statistically has the highest air pollution concentration. As shown in Fig. 13, for PM2.5 in Shanghai, the N = 3 neighborhood cities generally come from the northwest and the southwest. Cluster 2 takes a relatively higher proportion (29.89%) of time for Shanghai PM2.5, suggesting the pollutants may be dispersed from PM2.5 in Suzhou and Wuxi, and SO2 in Nantong.

2. Beijing during APEC period. Traditionally, causality is verified via interventions in a causal system. For example, we can verify the effect of a medicine by setting two groups of patients and only giving medicine to the treatment group. However, it is impossible to conduct intervention for air pollutants in the real environment. APEC period is a good opportunity to verify the causality, since the Chinese government shuttered factories in NC, and implemented traffic bans in and around Beijing [42]. Therefore, we compare the causal pathways for PM2.5 in Beijing before and during the APEC period. To illustrate the propagation of pollutants along the causal pathway, we connect the one-hop pathway to 3-hop as shown in Fig. 14(a)(b). The connection originates from the target pollutant, i.e., Beijing PM2.5, and connect its causal pollutants at neighbor cities. Then for each new connected pollutant, we repeat the same procedure for the next hop. The connection stops if in inference accuracy of one target pollutant based on its historical data is higher than based on the historical data of its ST "causers", indicating the pollutant is more likely to be generated locally. Fig. 14(a) shows Beijing's PM2.5 is likely to be caused by NO2 in Baoding (City 14), and PM10 in Cangzhou (City 18), during Jan − Mar. Further, for example, Cangzhou's PM10 is mostly influenced by PM10 in Dingzhou (City 15) and Binzhou (City 71), as well as PM2.5 in Dezhou (City 64). We list the information of all 128 cities in Fig. 15, as well as their corresponding optimal K and N for pollutant PM2.5 in Spring. Note that the causal pathways forms "circles" in the southwestern cities to Beijing, which is identical to the locations of the major plants in NC shown in Fig. 14(c). However, we notice that the causal pathway cannot be connected into 3-hops during the APEC period, since each "causer" pollutant to Beijing PM2.5 (i.e. NO2 in Chengde and Zhangjiakou, and Tianjin) is more likely to be inferred by its own historical data over its ST "causers" in this period. This may suggest the PM2.5 in Beijing during the APEC period are mostly affected by pollutants locally and nearby. The 3-hop causal pathways learnt by three baselines are quite similar, thus we only present the result learned by CGBN, pg-Causality-p, pg-Causality-n, MWST, and MCMC in Fig. 14(d-h). Our approach has better interpretability. It is noted that without pattern mining module, the candidate "causers" for Beijing tend to be at irrelevant locations. While without integrating confounders, the causal pathways tend to have too many paths to be distinguished. We summarize the discovery for Beijing's air pollution as follows.

• Among all the cities within a region, a target pollutant can be mainly affected by only several cities in the ST space. The locations of most influential cities to a target pollutant demonstrate seasonal similarities.

• The causal pathways for PM2.5 in Beijing may come in multiple hops that form "circle" in the southwest of Beijing, suggest-
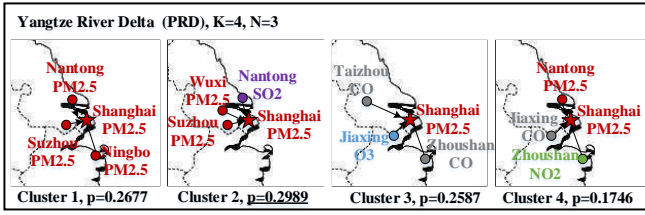
Figure 13: Visualization of final causal pathways for PM2.5 in Shanghai.

ing superposition or reaction of air pollutants in the corresponding area. While during the APEC period with low pollution level, we did not see multi-hop causal pathways, suggesting the PM2.5 are more likely to be generated locally or nearby within this period.

6. RELATED WORK

Data-driven Air Pollution Analysis: In recent years, air pollution analysis has drawn a lot of attention from the data mining community [10][11]. [12][13][14] propose data-driven approaches to infer and forecast fine-grained air quality using heterogeneous urban data. [15] estimates the gas consumption and pollutants emission of vehicles, based on the vehicles' GPS trajectories in the road network. Our paper differs from these works in that, we target at understanding the underlying causal pathways of air pollution. We identified the most likely "causers" in the geospace by learning the most likely graphical structures of an ST causality network, rather than predicting air quality or estimating pollutant emission with a black-box neural network.

Causality Modelling for Time Series: Causal modelling has been systematically studied for over half a century [16][17], from the statistical and mathematical perspectives. For time series data, existing works on modelling causality can be classified into three categories. The first category is based on Rubin's unit-level causality [16], which is the statistical analysis on the potential outcome between two groups, given "treatment" and "control", respectively [18]. With the increase of computation power, variations of unit-level causality were conducted, such as the cause-and-effect of advertising on behaviour change [8], genes on phenotype [19], etc. The second category considers a pair of time series, and aims to quantify the strength of causal influence from one time series to another. Researchers have developed different measures for this purpose, such as transfer entropy [20], and Granger's causality [17][21]. The third category aims to extract graphical causal relations from multiple time series. [22] combines graphical techniques with the classic Granger causality, and proposes a model to infer causality strengths for a large number of time series variables. Pearl's causality model [5] encodes the causal relationships in a directed acyclic graph (DAG) [23] for probabilistic inference. The most well used graphical representation of DAG is Bayesian network (BN) [23]. Temporal dependencies can be incorporated in the DAG by using Murphy's dynamic Bayesian network (DBN) [24]. There are also various extensions that incorporate spatiotemporal dependencies in the domain of traffic [4], climate [25][26][27] and flood prediction [28].

Our proposed approach *pg-Causality* belongs to the third category, i.e., using graphical model to detect causalities from multiple time series, where "p" refers to "pattern-aided" and "g" refers to graphical causality. The terms "causality" or "causalities" used later in this article are actually graphical causality.

The approach differs from the above works in three aspects: (1)

As a data-driven causality learning method, we combine pattern mining and Bayesian learning to make the causality analysis more efficient and robust to the noise present in the input data. (2) Besides the multi-variate time series data, we also consider the impact of confounding given different environmental factors for unbiased causality analysis. (3) Since we cannot conduct human intervention on air pollution at the nation-wide scale, this article identifies the causality from historical data. We proposed a Bayesian-based graphical causality model to capture the dependencies among different air pollution in the spatiotemporal (ST) space. Verification is based on the training accuracy, synthetic results, as well as observation.

7. CONCLUSION

In this paper, we identified the *ST causal pathways* for air pollutants using large-scale air quality data and meteorological data. We have proposed a novel causal pathway learning approach named *pg-Causality* that tightly combines pattern mining and Bayesian learning. Specifically, by extending existing sequential pattern mining techniques, *pg-Causality* first extracts a set of FEPs for each sensor, which captures most regularities in the air polluting process, largely suppresses data noise and reduces the complexity in the ST space. In the Bayesian learning module, *pg-Causality* leverages the pattern-matched data to train a graphical structure, which carefully models multi-faceted causality and environmental factors. We performed extensive experiments on three real-word data sets. Experimental results demonstrate that the causal pathways detected by *pg-Causality* are highly interpretable and meaningful. Moreover, it outperforms baseline methods in both efficiency and inference accuracy. For future work, we plan to apply this pattern-aided causality analysis framework for other tasks in the ST space, such as traffic congestion analysis and human mobility modelling [43].

8. REFERENCES

[1] S. Lee, W. Liu, Y. Wang, A. G. Russell, and E. S. Edgerton, "Source apportionment of PM 2.5: Comparing PMF and CMB results for four ambient monitoring sites in the southeastern united states," *Atmospheric Environment*, vol. 42, no. 18, pp. 4126–4137, 2008.

[2] A. Keats, E. Yee, and F.-S. Lien, "Bayesian inference for source determination with applications to a complex urban environment," *Atmospheric environment*, vol. 41, no. 3, pp. 465–479, 2007.

[3] C. Zhang, Y. Zheng, X. Ma, and J. Han, "Assembler: Efficient discovery of spatial co-evolving patterns in massive geo-sensory data," in *KDD*. ACM, 2015, pp. 1415–1424.

[4] H. Nguyen, W. Liu, and F. Chen, "Discovering congestion propagation patterns in spatio-temporal traffic data," vol. 3, no. 2. IEEE, 2017, pp. 169–180.

[5] J. Pearl, "Causality: models, reasoning and inference," *Economet. Theor*, vol. 19, pp. 675–685, 2003.

[6] J. Y. Zhu, Y. Zheng, X. Yi, and V. O. Li, "A gaussian bayesian model to identify spatio-temporal causalities for air pollution based on urban big data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on*. IEEE, 2016.

[7] D. M. Chickering, "Learning bayesian networks is np-complete," in *Learning from data*. Springer, 1996, pp. 121–130.

[8] W. Sun, P. Wang, D. Yin, J. Yang, and Y. Chang, "Causal inference via sparse additive models with application to online advertising," in *AAAI*, 2015, pp. 297–303.
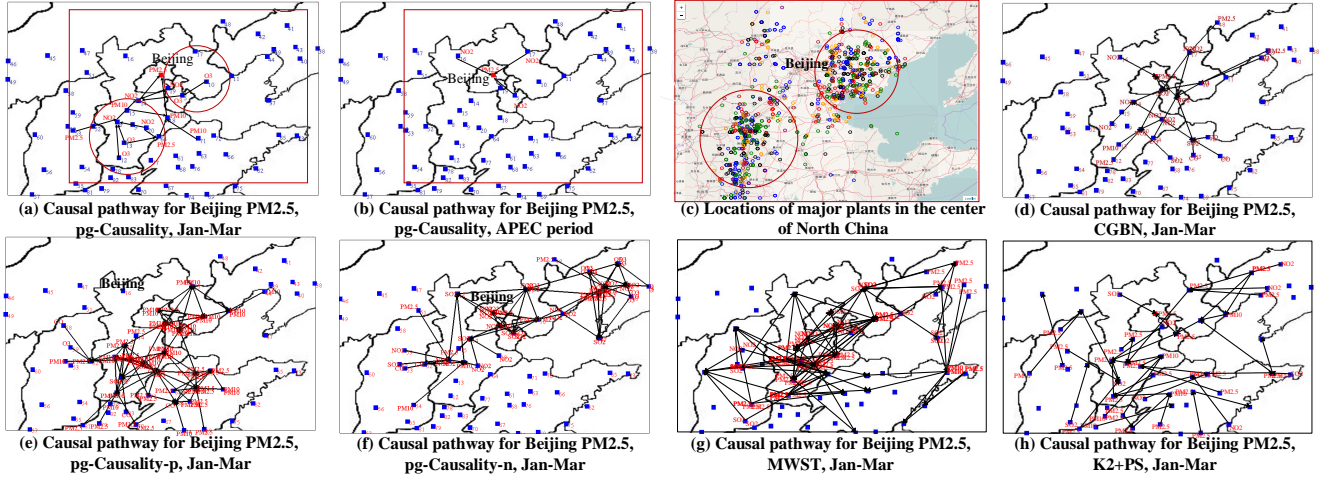
(a) Causal pathway for Beijing PM2.5, pg-Causality, Jan-Mar

(b) Causal pathway for Beijing PM2.5, pg-Causality, APEC period

(c) Locations of major plants in the center of North China

(d) Causal pathway for Beijing PM2.5, CGBN, Jan-Mar

(e) Causal pathway for Beijing PM2.5, pg-Causality-p, Jan-Mar

(f) Causal pathway for Beijing PM2.5, pg-Causality-n, Jan-Mar

(g) Causal pathway for Beijing PM2.5, MWST, Jan-Mar

(h) Causal pathway for Beijing PM2.5, K2+PS, Jan-Mar

Figure 14: The causal pathways for Beijing PM2.5 before (a) and during APEC period (b), compared with the locations of major plants in Hebei Province, China (c), and the causal pathways learned by baseline method CGBN (d), pg-Causality-p (e), pg-Causality-n (f), MWST (g), MCMC (h).

City No.	City_Name	Latitude	Longitude	K	N	Accuracy	Region	City No.	City_Name	Latitude	Longitude	K	N	Accuracy	Region	City No.	City_Name	Latitude	Longitude	K	N	Accuracy	Region
1	Beijing	39.993	116.413	4	2	0.925	NC	44	Huludao	40.751	120.851	1	1	0.811	NC	87	Yancheng	33.391	120.157	1	1	0.891	YRD
2	Shanghai	31.184	121.456	5	0	0.938	YRD	45	Huhehaote	40.801	111.665	1	0	0.834	NC	88	Xuzhou	34.315	117.359	2	0	0.930	YRD
3	Shenzhen	22.635	114.121	1	0	0.954	PRD	46	Baotou	40.573	110.022	1	0	0.860	NC	89	Huaian	33.582	119.036	4	0	0.912	YRD
4	Ningbo	29.832	121.509	2	0	0.933	YRD	47	Wulanchabu	41.015	113.114	2	1	0.799	NC	90	Lianyungang	34.657	119.258	1	0	0.891	YRD
5	Tianjin	39.156	117.306	1	5	0.931	NC	48	Chifeng	42.210	119.008	1	5	0.869	NC	91	Changzhou	31.787	119.962	5	2	0.916	YRD
6	Guangzhou	23.159	113.377	4	0	0.957	PRD	49	Erduosi	39.813	110.002	6	0	0.786	NC	92	Taizhou	32.367	120.031	2	1	0.939	YRD
7	Hong Kong	22.343	114.163	4	0	0.948	PRD	50	Taiyuan	37.863	112.517	4	1	0.898	NC	93	Suqian	33.956	118.281	2	5	0.916	YRD
8	Shijiazhuang	38.045	114.588	4	3	0.921	NC	51	Datong	40.094	113.303	1	2	0.890	NC	94	Huangshi	30.216	115.055	1	0	0.954	YRD
9	Xinji	37.949	115.224	1	0	0.872	NC	52	Yangquan	37.861	113.566	2	5	0.931	NC	95	Hangzhou	30.076	119.893	3	1	0.936	YRD
10	Tangshan	39.720	118.311	1	4	0.908	NC	53	Jinzhong	37.696	112.734	6	1	0.910	NC	96	Huzhou	30.787	119.951	1	0	0.957	YRD
11	Qinhuangdao	39.955	119.367	6	1	0.887	NC	54	Changzhi	36.190	113.109	1	3	0.940	NC	97	Jiaxing	30.655	120.809	1	0	0.907	YRD
12	Handan	36.568	114.659	2	1	0.937	NC	55	Jincheng	35.498	112.849	4	2	0.838	NC	98	Shaoxing	29.869	120.613	1	5	0.890	YRD
13	Xingtai	37.185	114.879	1	0	0.921	NC	56	Linfen	36.078	111.514	1	0	0.938	NC	99	Taizhou	28.683	121.197	1	0	0.910	YRD
14	Baoding	38.933	115.474	5	0	0.926	NC	57	Yuncheng	35.041	111.015	1	0	0.903	NC	100	Wenzhou	28.061	120.753	1	1	0.916	YRD
15	Dingzhou	38.522	114.997	1	1	0.855	NC	58	Shuozhou	39.344	112.431	3	0	0.790	NC	101	Lishui	28.349	119.704	1	5	0.896	YRD
16	Zhangjiakou	40.787	114.925	2	2	0.846	NC	59	Yizhou	38.443	112.726	1	3	0.859	NC	102	Jinhua	29.160	119.902	5	0	0.896	YRD
17	Chengde	40.974	117.833	2	0	0.861	NC	60	Lvliang	37.522	111.136	1	0	0.863	NC	103	Quzhou	28.942	118.777	2	0	0.901	YRD
18	Cangzhou	38.224	116.688	2	0	0.921	NC	61	Jinan	36.644	117.030	2	5	0.922	NC	104	Zhoushan	30.034	122.238	1	1	0.894	YRD
19	Langfang	39.444	116.694	1	0	0.893	NC	62	Qingdao	36.123	120.384	1	4	0.922	NC	105	Hefei	31.848	117.248	7	0	0.910	YRD
20	Hengshui	37.809	115.800	1	5	0.917	NC	63	Zibo	36.744	118.005	1	0	0.921	NC	106	Bengbu	32.929	117.357	1	5	0.892	YRD
21	Dongguan	23.024	113.762	2	1	0.932	PRD	64	Dezhou	37.459	116.328	1	3	0.886	NC	107	Wuhu	31.366	118.375	1	1	0.896	YRD
22	Foshan	22.988	113.063	2	1	0.933	PRD	65	Yantai	37.511	121.336	1	0	0.902	NC	108	Whuainan	32.655	116.874	6	3	0.880	YRD
23	Heyuan	23.746	114.687	1	0	--	PRD	66	Weifang	36.709	119.124	2	1	0.898	NC	109	Maanshan	31.697	118.525	1	1	0.934	YRD
24	Huizhou	23.012	114.368	1	0	0.935	PRD	67	Jining	35.409	116.622	2	0	0.935	NC	110	Anqing	30.547	117.031	1	0	0.914	YRD
25	Jiangmen	22.516	112.912	5	0	0.894	PRD	68	Taian	36.180	117.122	1	0	0.899	NC	111	Suzhou	33.639	116.971	1	1	0.879	YRD
26	Jieyang	22.593	113.082	4	0	0.861	PRD	69	Linyi	35.053	118.329	1	0	0.919	NC	112	Fuyang	32.881	115.831	1	0	0.883	YRD
27	Qingyuan	23.677	113.042	7	0	0.917	PRD	70	Heze	35.248	115.468	2	0	0.913	NC	113	Bozhou	33.848	115.795	1	1	0.898	YRD
28	Shanwei	22.783	115.371	1	1	0.906	PRD	71	Binzhou	37.374	117.975	2	2	0.882	NC	114	Huangshan	29.903	118.255	3	0	0.804	YRD
29	Shaoguan	24.772	113.593	3	0	0.930	PRD	72	Dongying	37.488	118.614	1	1	0.896	NC	115	Chuzhou	32.300	118.317	1	0	0.883	YRD
30	Yunfu	22.937	112.043	1	0	0.925	PRD	73	Weihai	37.475	122.092	7	1	0.903	NC	116	Huaibei	33.940	116.797	1	0	0.895	YRD
31	Zhaoqing	23.091	112.484	3	1	0.898	PRD	74	Zaozhuang	34.815	117.481	1	1	0.921	NC	117	Tongling	30.936	117.820	1	0	0.889	YRD
32	Zhongshan	22.516	113.392	1	0	0.943	PRD	75	Rizhao	35.393	119.501	1	0	0.888	NC	118	Xuancheng	30.954	118.738	1	0	0.888	YRD
33	Zhuhai	22.285	113.501	1	0	0.922	PRD	76	Laiwu	36.209	117.726	1	1	0.932	NC	119	Liuan	31.762	116.515	3	5	0.907	YRD
34	Nanjing	31.985	118.816	2	2	0.917	YRD	77	Liaocheng	36.457	115.982	2	2	0.922	NC	120	Chizhou	30.652	117.483	1	0	0.831	YRD
35	Suzhou	31.438	120.716	5	0	0.938	YRD	78	Anyang	36.096	114.392	1	0	0.864	NC	121	Nanchang	28.690	115.879	1	5	0.883	YRD
36	Wuxi	31.616	120.209	1	2	0.930	YRD	79	Xinxiang	35.293	113.923	5	3	0.890	NC	122	Jiujiang	29.672	116.002	1	1	0.944	YRD
37	Dalian	38.950	121.628	3	0	0.900	NC	80	Shangqiu	34.417	115.655	1	0	0.883	YRD	123	Shangrao	28.449	117.958	6	0	0.951	YRD
38	Anshan	41.096	122.968	2	2	0.867	NC	81	Jiaozuo	35.223	113.235	1	4	0.917	NC	124	Fuzhou	28.040	116.291	7	0	0.942	YRD
39	Jinzhou	41.059	121.128	1	0	0.804	NC	82	Hebi	35.744	114.301	6	0	0.932	NC	125	Jingdezhen	29.304	117.224	1	0	0.946	YRD
40	Yingkou	40.676	122.222	4	0	0.839	NC	83	Puyang	35.772	115.053	1	3	0.909	NC	126	Yingtan	28.209	117.013	1	0	0.910	YRD
41	Fuxin	42.042	121.685	1	0	0.837	NC	84	Zhenjiang	32.108	119.477	3	2	0.920	YRD	127	Wuzhou	23.462	111.276	2	0	0.941	PRD
42	Chaoyang	41.692	120.461	1	0	0.730	NC	85	Nantong	31.990	120.879	1	2	0.903	YRD	128	Hezhou	24.413	111.544	1	0	0.903	PRD
43	Panjin	41.151	122.032	2	2	0.847	NC	86	Yangzhou	32.537	119.397	5	0	0.938	YRD								

Figure 15: Optimal K and N for 128 cities, in Region NC, YRD and PRD, for PM2.5 during Jan − Mar.

[9] C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta, "Splitter: Mining fine-grained sequential patterns in semantic trajectories," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 769–780, 2014.

[10] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[11] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE transactions on big data*, vol. 1, no. 1, pp. 16–34, 2015.

[12] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *KDD*. ACM, 2013, pp. 1436–1444.

[13] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in

KDD, 2015.

[14] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 437–446.

[15] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *KDD*, 2014.

[16] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[17] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[18] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, pp. 41–55, 1983.

[19] D. S. Wald, M. Law, and J. K. Morris, "Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis," *Bmj*, vol. 325, no. 7374, p. 1202, 2002.

[20] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical Review Letters*, 2009.

[21] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[22] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *KDD*. ACM, 2007, pp. 66–75.

[23] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[24] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, University of California, Berkeley, 2002.

[25] I. Ebert-Uphoff and Y. Deng, "Causal discovery from spatio-temporal data with applications to climate science," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 606–613.

[26] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, "Escaping the curse of dimensionality in estimating multivariate transfer entropy," *Physical review letters*, vol. 108, no. 25, p. 258701, 2012.

[27] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe, "Spatial-temporal causal modeling for climate change attribution," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 587–596.

[28] P. Jangyodsuk, D.-J. Seo, R. Elmasri, and J. Gao, "Flood prediction and mining influential spatial features on future flood with causal discovery," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1462–1469.

[29] J. Pel, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining sequential patterns by prefix-projected growth," in *Proc. 17th IEEE International Conference on Data Engineering (ICDE).*

[30] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *SIGMOD*, 2003.

[31] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[32] M. A. Gómez, P. M. Villegasa, H. Navarrob, and R. Susia, "Dealing with uncertainty in gaussian bayesian networks from a regression perspective," *on Probabilistic Graphical Models*, p. 145, 2010.

[33] R. D'Agostino and E. S. Pearson, "Tests for departure from normality. empirical results for the distributions of b2 and b1," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.

[34] J. Y. Zhu, C. Sun, and V. O. Li, "Granger-causality-based air quality estimation with spatio-temporal (st) heterogeneous big data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2015 IEEE Conference on*. IEEE, 2015, pp. 612–617.

[35] I. Flesch and P. J. Lucas, "Markov equivalence in bayesian networks," in *Advances in Probabilistic Graphical Models*. Springer, 2007, pp. 3–38.

[36] K. Murphy *et al.*, "The bayes net toolbox for matlab," *Computing science and statistics*, vol. 33, no. 2, pp. 1024–1034, 2001.

[37] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *Information Theory, IEEE Transactions on*, 1968.

[38] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," pp. 222–228, 1987.

[39] J. L. Beck and S.-K. Au, "Bayesian updating of structural models and reliability using markov chain monte carlo simulation," *Journal of Engineering Mechanics*, vol. 128, no. 4, pp. 380–391, 2002.

[40] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[41] R. M. Lewis and V. Torczon, "A globally convergent augmented lagrangian pattern search algorithm for optimization with general constraints and simple bounds," *SIAM Journal on Optimization*, vol. 12, no. 4, pp. 1075–1089, 2002.

[42] K. Huang, X. Zhang, and Y. Lin, "The "apec blue" phenomenon: Regional emission control effects observed from space," *Atmospheric Research*, vol. 164, pp. 65–75, 2015.

[43] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "Gmove: Group-level mobility modeling using geo-tagged social media," in *KDD*, 2016, pp. 1305–1314.

Heidelberg, Germany, 2001, pp. 215–224.