A Gaussian Bayesian Model to Identify Spatio-temporal Causalities for Air Pollution based on Urban Big Data

Julie Yixuan Zhu [1,2], Yu Zheng [2], Xiuwen Yi [2,3], Victor O.K. Li [1]
[1] The University of Hong Kong, Hong Kong
[2] Microsoft Research, Beijing, China
[3] Southwest Jiaotong University, Chengdu, Sichuan, China
{yxzhu, vli}@eee.hku.hk; {yuzheng, v-xiuyi}@microsoft.com

***Abstract*— Identifying the causalities for air pollutants and answering questions, such as, where do Beijing's air pollutants come from, are crucial to inform government decision-making. In this paper, we identify the spatio-temporal (ST) causalities among air pollutants at different locations by mining the urban big data. This is challenging for two reasons: 1) since air pollutants can be generated locally or dispersed from the neighborhood, we need to discover the causes in the ST space from many candidate locations with time efficiency; 2) the cause-and-effect relations between air pollutants are further affected by confounding variables like meteorology. To tackle these problems, we propose a coupled Gaussian Bayesian model with two components: 1) a Gaussian Bayesian Network (GBN) to represent the cause-and-effect relations among air pollutants, with an entropy-based algorithm to efficiently locate the causes in the ST space; 2) a coupled model that combines cause-and-effect relations with meteorology to better learn the parameters while eliminating the impact of confounding. The proposed model is verified using air quality and meteorological data from 52 cities over the period Jun 1st 2013 to May 1st 2015. Results show superiority of our model beyond baseline causality learning methods, in both time efficiency and prediction accuracy.**

Keywords—Causality analysis; Bayesian network (BN); urban big data; spatio-temporal (ST); air pollution.

I. INTRODUCTION

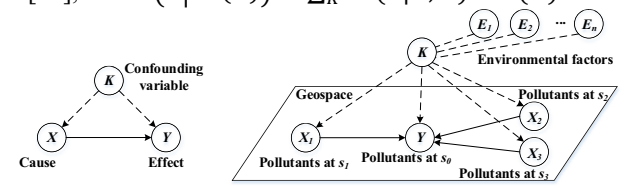
Air pollution has become a global concern, impacting the health of billions of people and the sustainable development of the world [1]. Many cities have built on-the-ground air quality monitoring stations to measure hourly concentration of air pollutants, such as PM2.5 and NO2. Besides monitoring and forecasting [2] air quality, there is a strong demand on diagnosing the causalities of air pollutants. For example:

- *Where do a city's air pollutants come from in the spatio-temporal (ST) space?*
- *How is one pollutant affected by other pollutants, given different weather conditions?*

Identifying the causalities for air pollution will help inform governments' decision-making on mitigating air pollution. As there are many uncertain factors affecting air pollution, a comprehensive causality analysis is needed to 1) discover where these factors are in the ST space and 2) how much they influence the target air pollutant.

In practice, the cause-and-effect relation *"X to Y"* is usually a non-deterministic problem, and requires probabilistic languages to represent the uncertainty, e.g., using $Pr(Y|X)$ to represent the treatment effect of *Y* given *X*. There are two major streams of causality modelling. One is Pearl's causality model [3] based on the Bayesian network [4] which encodes the cause-and-effect relations in a graphical structure and conducts causal inference via a *"do"* operator $Pr\big(Y\big|do(X)\big)$. The other is unit-level causality [5-6], which estimates the potential outcome (effects) given different treatments (causes). Applications extended from the two causality frameworks have been developed to learn the cause-and-effect of medicine on recovery, advertising on behavior change, genes on phenotype, etc. [7-9]. Recently, with advances in computation, causality modelling is greatly driven by learning the patterns between events, which is more practical and operable to predict future events than modelling the strict cause-and-effect relations [10-13].

Note that there may be a biased estimation of effect *Y* given cause *X* when a confounding variable *K* exists [14], as shown in Fig. 1(a). The confounding variable here refers to a third variable that simultaneously correlates the cause *X* and effect *Y*, e.g., gender *K* may affect the effect of recovery *Y* given a medicine *X*. To guarantee an unbiased causal inference, the treatment effect is usually adjusted by averaging all the sub-classification cases of *K* [15], i.e. $Pr\big(Y\big|do(X)\big) = \sum_K Pr(Y|X,K)\cdot Pr(K)$.



(a) Cause-and-effect with confounding **(b) An illustration of cause-and-effect with confounding for air pollutants in the ST space**

Fig. 1. Examples of cause-and-effect with confounding.

However, to identify the causalities among air pollutants with the existing causal modelling frameworks, one must overcome two challenges. *First*, an ST representation of cause-and-effect relations is needed, e.g., Fig. 1(b) shows an example where air pollutants X_1, X_2, X_3 at locations s_1, s_2, s_3 cause the pollutant Y at location s_0. Since the air pollution could be generated locally or dispersed from the neighborhood, we need to locate the causes in the ST space from many candidate locations efficiently. *Second*, the cause-and-effect relations between air pollutants can be further affected by confounding variables. As shown in Fig. 1(b), the confounding variable *K*

may contain the information from various environmental factors $E_1, E_2, \ldots E_n$ (wind, humidity, etc.).

To tackle the challenges, we propose a coupled Gaussian Bayesian model with two major components: a Gaussian Bayesian Network (GBN)-based representation for cause-and-effect relations in the ST space, and a model that couples the cause-and-effect relations with meteorology. The contribution of this paper is two-fold:

- *First, the GBN-based representation captures both the local and neighborhood cause-and-effect relations, with an entropy-based algorithm to efficiently locate the causes in the ST space.*
- *Second, we propose a coupled model that integrates the influence of the meteorology into the cause-and-effect relations, to help better estimate the parameters while eliminating the impacts of confounding.*

The causality analysis for air pollution is different from basic causality inference [3]. The latter is verified via intervention of specific treatments in a predefined causal structure [16]. But for air pollution, it is impossible to conduct interventions, besides, the causal structure is unknown in the complex ST space. Thus the major part of this paper will discuss the ST causality representation for air pollutants, and the parameter learning of the cause-and-effect relations. Verification is conducted via a prediction task instead of intervention-based causal inference. We evaluate the model with air quality and meteorological data from 52 cities over the period Jun 1st 2013 to May 1st 2015, and demonstrate both time efficiency and prediction accuracy. This suggests the model is capable of identifying the causalities for air pollutants in the ST space, finding where the air pollutants come from and understanding how they interact with each other.

II. Component 1: Gaussian Bayesian Network (GBN) for Cause-and-Effect Representation

We first describe the general idea of GBN and its benefits for modelling the causality. Second, we introduce the GBN-based representation of cause-and-effect in the ST space. Thirdly, we propose an algorithm to locate the causes for a target air pollutant based on transfer entropy. An optimization of the algorithm is presented at the end of this section.

A. Gaussian Bayesian Network (GBN)

GBN is a special form of Bayesian network [4], capable of encoding the causal relations in a directed acyclic graph (DAG) and providing a graphical representation of conditional dependencies among variables.

In GBN, all the variables are assumed Gaussian and all the dependencies linear Gaussian [17-18]. For example, If $\boldsymbol{Y}$ is a linear Gaussian of its parents $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, then:

$$\Pr(\boldsymbol{Y}|\boldsymbol{X}) \sim N(\boldsymbol{\mu_Y} + (\boldsymbol{X} - \boldsymbol{\mu_X}) \cdot A, \Sigma(\boldsymbol{\varepsilon})) \quad (1)$$

where A is the corresponding linear regression coefficients matrix for the regression function $\boldsymbol{Y} = \boldsymbol{\mu_0} + \boldsymbol{X} \cdot A + \boldsymbol{\varepsilon}$, and $\Sigma(\boldsymbol{\varepsilon})$ is the covariance of $\boldsymbol{Y}$ conditioned on $\boldsymbol{X}$ with $\boldsymbol{\Sigma}$ denoting the covariance operator. By using ordinary least square (OLS) to find A that minimizes the regression error, i.e. $trace(\Sigma(\boldsymbol{\varepsilon}))$, the covariance $\Sigma(\boldsymbol{\varepsilon})$ can be rewritten as:

$$\begin{aligned}\Sigma(\boldsymbol{\varepsilon}) = \Sigma(\boldsymbol{Y}|\boldsymbol{X})) &= \Sigma(\boldsymbol{Y}) - A^{\mathrm{T}} \cdot \Sigma(\boldsymbol{X}) \cdot A \\ &= \Sigma(\boldsymbol{Y}) - \Sigma(\boldsymbol{X}, \boldsymbol{Y}) \cdot \Sigma(\boldsymbol{X})^{-1} \cdot \Sigma(\boldsymbol{X}, \boldsymbol{Y})^{\mathrm{T}} \quad (2)\end{aligned}$$

In this way, GBN provides a simple way to capture many dependencies among multiple variables. Furthermore, the characteristics of urban data fit the GBN model well. As shown in Fig. 2, the 1-hour difference (current value minus the value 1-hour ago) air pollutants time series at two cities, Beijing PM2.5 and Shijiazhuang PM10, after being normalized by their corresponding standard deviations, obey Gaussian distribution. In addition, the distribution of Beijing PM2.5 normalized 1-hour difference when conditioned on the value of Shijiazhuang PM10 normalized 1-hour difference at the same timestamp to be between (0.5, 1) is observed to be Gaussian. This suggests the 1-hour difference of air pollutants are Gaussian and their dependencies are also Gaussian.

For other air quality and meteorological data, [19] found that all the temporal differences obey Gaussian distribution. Thus in this paper, we use 1-hour differences as inputs of multi-source time series data for the GBN, to represent the cause-and-effect relations in the ST space.

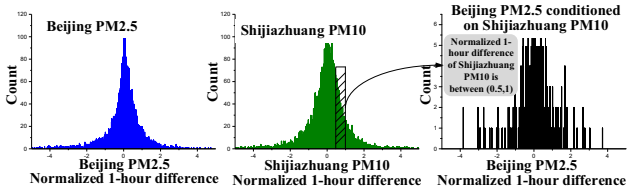


Fig. 2. Histograms of normalized 1-hour difference of Beijing PM2.5, Shijiazhuang PM10, and their dependency.

B. Cause-and-Effect Representation in the ST Space

We first denote air pollutants at each location s and time t, as a random vector $\boldsymbol{P_{st}}$, which is the effect caused by the historical air pollutants at the same location, as well as neighborhood locations. For each instance of $\boldsymbol{P_{st}}$, its value $\boldsymbol{p_{st}} = (p_{1st}, p_{2st}, \ldots, p_{n_pst})$ is the value vector of n_p types of pollutants. For example, $\boldsymbol{p_{st}} = (1, -1, 0.2, 0.5, -2.7, 3)$ represents the normalized 1-hour temporal difference values at location s and time t for $n_p = 6$ types of pollutants, i.e. PM2.5, PM10, NO_2, CO, O3, SO_2. Fig. 3(a) gives an example of the GBN representation for air pollution data. The child node is the air pollutants $\boldsymbol{P_{s_0t}}$ at location s_0 and time t, and the parent nodes are the historical air pollutants (1 and 2 hours ago) $\boldsymbol{P_{s_0(t-1)}}$, $\boldsymbol{P_{s_0(t-2)}}$ at s_0 and $\boldsymbol{P_{s_1(t-1)}}$, $\boldsymbol{P_{s_1(t-2)}}$, $\boldsymbol{P_{s_2(t-1)}}$, $\boldsymbol{P_{s_2(t-2)}}$ at two neighborhood locations s_1, s_2.

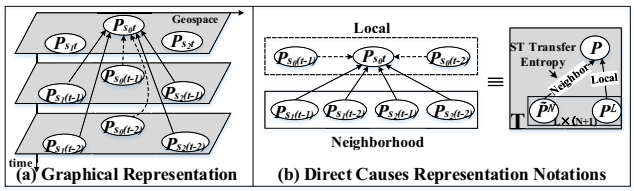


Fig. 3. Representation of the GBN-based ST cause-and-effect.

For the general representation (as shown in Fig. 3(b)), suppose the air pollutants at location s_0 are influenced by L historical time lags locally and at N neighbor locations. The local pollutants in the past time are denoted as $\boldsymbol{P_{s_0t}^{(L)}} = \boldsymbol{P_{s_0(t-1)}} \oplus \boldsymbol{P_{s_0(t-2)}} \oplus \cdots \oplus \boldsymbol{P_{s_0(t-L)}}$, which is a $1 \times n_p L$ vector with L lags of $\boldsymbol{P_{s_0t}}$. $\oplus$ is the notion for vector concatenation.

The neighborhood pollutants in the past time are denoted by

$\widetilde{\boldsymbol{P}}_{\boldsymbol{s_0t}}^{[N]} = \boldsymbol{P}_{\boldsymbol{s_1t}}^{(L)} \oplus \boldsymbol{P}_{\boldsymbol{s_2t}}^{(L)} \oplus \cdots \oplus \boldsymbol{P}_{\boldsymbol{s_Nt}}^{(L)}$, which is a $1 \times n_p NL$ random vector for each t and represents the historical pollutants at N neighbor locations. The parents of $\boldsymbol{P}_{\boldsymbol{s_0t}}$ are $\boldsymbol{PA}(\boldsymbol{P}_{\boldsymbol{s_0t}}) = \boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)} \oplus \widetilde{\boldsymbol{P}}_{\boldsymbol{s_0t}}^{[N]}$. Based on Equation (1), the distribution of $\boldsymbol{P}_{\boldsymbol{s_0t}}$ conditioned on its parents $\boldsymbol{PA}(\boldsymbol{P}_{\boldsymbol{s_0t}})$ can be represented by a Gaussian distribution:

$$\Pr\left(\boldsymbol{P}_{\boldsymbol{s_0t}} = \boldsymbol{p}_{\boldsymbol{s_0t}} \middle| \boldsymbol{PA}(\boldsymbol{P}_{\boldsymbol{s_0t}})\right) \sim N(\boldsymbol{\mu}_{\boldsymbol{s_0t}} + \sum_{n=0}^{N}\sum_{l=1}^{L} \boldsymbol{a}_{\boldsymbol{nL+l}}(\boldsymbol{p}_{\boldsymbol{s}_n(t-L)} - \boldsymbol{\mu}_{\boldsymbol{s}_n(t-L)}), \Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{s_0t}})) \quad (3)$$

$\boldsymbol{\mu}_{\boldsymbol{s_0t}}$ is the marginal mean for $\boldsymbol{P}_{\boldsymbol{s_0t}}$, $\boldsymbol{a}_{\boldsymbol{nL+l}}$ is the regression coefficient of $\boldsymbol{P}_{\boldsymbol{s_0t}}$ given its parents. The regression function is as follows:

$$\boldsymbol{P}_{\boldsymbol{s_0t}} = \boldsymbol{\mu}_0 + \left(\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)} \oplus \widetilde{\boldsymbol{P}}_{\boldsymbol{s_0t}}^{[N]}\right) A + \boldsymbol{\varepsilon}_{\boldsymbol{s_0t}} \quad (4)$$

where $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_{\boldsymbol{s_0t}} - \sum_{n=0}^{N}\sum_{l=1}^{L} \boldsymbol{a}_{\boldsymbol{nL+l}}\boldsymbol{\mu}_{\boldsymbol{s}_n(t-L)}$.

By finding a $n_p(N+1)L \times n_p$ matrix A that minimizes the uncertainty of $\boldsymbol{P}_{\boldsymbol{s_0t}}$ given its parents, we obtain:

$$\Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{s_0t}}) = \Sigma(\boldsymbol{P}_{\boldsymbol{s_0t}}) - A \cdot \boldsymbol{\Sigma}\left(\boldsymbol{PA}(\boldsymbol{P}_{\boldsymbol{s_0t}})\right)^{-1} \cdot \mathrm{A}^{\mathrm{T}} \quad (5)$$

However, to minimize $\Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{s_0t}})$ is very time consuming as there are many combinations of N neighborhood locations.

C. Locating the N Most Influential Neighbors

This subsection introduces how to select a combination of N neighborhood locations that could minimize $\Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{s_0t}})$ in equation (5), and guarantee the time efficiency. For air pollution $\boldsymbol{P}_{\boldsymbol{s_0t}}$ at location s_0, assuming there are totally S neighbor locations within a given distance d, the complexity for selecting its N most influential neighborhood locations would be $O(S^N)$. To optimize the selection process, we first use ST transfer entropy to measure the pairwise information transfer for air pollution from one neighbor location to the target location s_0, and then we propose an ST hierarchical pruning algorithm.

Transfer entropy is a metric to measure the amount of directed (time-asymmetric) transfer of information between two random processes X and Y:

$$T_{X\to Y} = H(Y_t|Y_{t-1:t-L}) - H(Y_t|Y_{t-1:t-L}, X_{t-1:t-L}) \quad (6)$$

The transfer entropy from a process X to another process Y is the amount of uncertainty reduced in the future values of Y by knowing the past values of X given past values of Y. In our case, the transfer entropy from the nth neighbor location to the target location s_0 is defined as:

$$T_{\boldsymbol{s_n}\to\boldsymbol{s_0}} = H\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)}\right) - H\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)} \oplus \widetilde{\boldsymbol{P}}_{\boldsymbol{s_nt}}^{(L)}\right) \quad (7)$$

When the variables are Gaussian, Equation (7) could be transformed to:

$$T_{\boldsymbol{s_n}\to\boldsymbol{s_0}} \xrightarrow{\boldsymbol{Gaussian}} \frac{1}{2} ln\left(\frac{\Sigma\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)}\right)}{\Sigma\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)} \oplus \widetilde{\boldsymbol{P}}_{\boldsymbol{s_nt}}^{(L)}\right)}\right) = \frac{1}{2} ln\left(\frac{|\Sigma(\varepsilon'_{\boldsymbol{s_0t}})|}{|\Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{s_0t}})|}\right) \quad (8)$$

Details of the proof can be found in [20]. Here $T_{\boldsymbol{s_n}\to\boldsymbol{s_0}}$ is used to represent the pairwise transfer entropy, i.e. the reduced uncertainty of air pollution at s_0 reduced by the history of its nth neighbor given the past value of the air pollution.

Similarly, we calculate the transfer entropy from N locations to the target location s_0, which characterizes how the air pollution at N neighbor locations could together reduce the uncertainty of the target pollution at s_0:

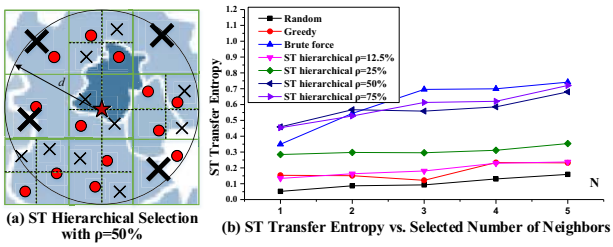
$$T_{\boldsymbol{s_{1:N}}\to\boldsymbol{s_0}} \xrightarrow{\boldsymbol{Gaussian}} \frac{1}{2} ln\left(\frac{\Sigma\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)}\right)}{\Sigma\left(\boldsymbol{P}_{\boldsymbol{s_0t}}\middle|\boldsymbol{P}_{\boldsymbol{s_0t}}^{(L)} \oplus \widetilde{\boldsymbol{P}}_{\boldsymbol{s_0t}}^{[N]}\right)}\right) = \frac{1}{2} ln\left(\frac{|\Sigma(\varepsilon'_{\boldsymbol{s_0t}})|}{|\Sigma(\varepsilon''_{\boldsymbol{s_0t}})|}\right) \quad (9)$$

Besides selecting N locations from the neighbor locations by brute force, there are two common ways to maximize the ST transfer entropy $T_{\boldsymbol{s_{1:N}}\to\boldsymbol{s_0}}$, such as *randomly* selecting N locations or *greedily* selecting the top N locations based on pairwise transfer entropy $T_{\boldsymbol{s_n}\to\boldsymbol{s_0}}$. However, these two approaches fail to achieve high performance regarding to the ST transfer entropy, either due to the randomness or missing data. Thus we propose an ST hierarchical algorithm integrating the ST characteristics into the spatial selection.

Fig. 4(a) illustrates the idea of the ST hierarchical algorithm. Given a spatial radius d, the region is initially divided into 9 squares with s_0 at the center. Then we do the following steps:

1) Average the pollution concentration in each region, and select a percentage ρ of the neighborhood regions with minimum pairwise transfer entropy in order.
2) Quad-tree based division for the remaining regions.
3) Again remove a percentage ρ of the neighborhood regions in each remaining sub-region with minimum transfer entropy in order, until the remaining locations = $3N$ (The choice of $3N$ is due to its generating the best results).
4) Finally, perform N_{iter} ($N_{iter} = 50$ in our experiment) Monte Carlo iterations and choose the combination of N neighbor locations that maximizes ST transfer entropy.

We show an example of ST transfer entropy vs. the selected number of neighbors in Fig. 4(b). The ST hierarchical algorithm with the selection rate $\rho = 50\%$ and $\rho = 75\%$ greatly outperform the greedy and random algorithm. Compared to the brute force method, the ST hierarchical selection algorithm achieves acceptable performance with low complexity $O(S + 9(1 + \log_{\rho}\frac{3N}{S}) + N_{iter})$. $\rho = 25\%$ does not generate as good performance as $\rho = 50\%$ or $\rho = 75\%$, but still outperforms the greedy algorithm. This is because the missing data severely distort the calculation of ST transfer entropy, while the Monte Carlo process in the ST hierarchical selection could avoid the missing items by using the neighborhood value as an alternative. When $\rho = 12.5\%$, the performance of ST hierarchical algorithm is similar to greedy. To achieve the optimum ST entropy and time efficiency, we use $\rho = 50\%$ in the following experiments.



(a) ST Hierarchical Selection with ρ=50%

(b) ST Transfer Entropy vs. Selected Number of Neighbors

Fig. 4. Illustration of (a) ST hierarchical selection algorithm with $\rho = 50\%$, and (b) the peformance with different ρ compared to the greedy and random selection methods.

III. Component 2: The Coupled Model

We further propose a coupled model that integrates the meteorology data into the cause-and-effect relations between air

pollution as Component 2, to eliminate the confounding induced by meteorology. Below we will introduce the coupled model, datasets selection and parameter learning.

A. Model Description

The coupled model assumes the cause-and-effect relations between air pollutants P and the environmental factors E in the format of GBN are simultaneously controlled by a hidden confounding variable K as shown in Fig. 5. K determines the parameters A, B of the cause-and-effect relations for air pollutants and environmental factors (meteorology), which are further determined by a hyper-parameter $\boldsymbol{\pi}$.

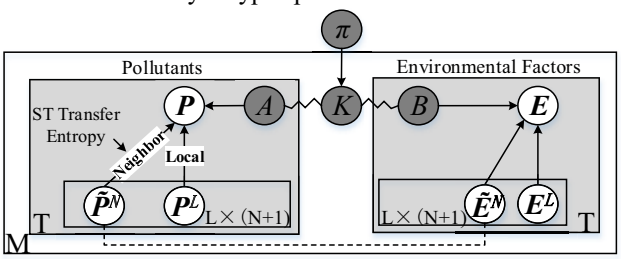


Fig. 5. The coupled Gaussian Bayesian model.

The notations for this model are listed as follows:

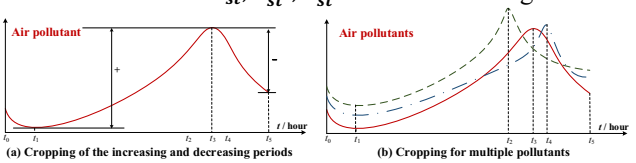
- M: Number of time windows (as training documents), each containing the air pollution and the corresponding meteorology data, within a time window $[t_1^d, t_2^d]$
- T: Number of timestamps for each document
- K: A hidden confounding variable that simultaneously determines parameters A and B
- $\boldsymbol{\pi}$: A hyper parameter which determines the distribution of K ($M \times K$-dimensional)
- A, B: Parameters for air pollutants and meteorology in the corresponding coupled model

For location s and time t and for each selected time window (document) $[t_1^d, t_2^d]$, the generation process for the current air pollutants values $\boldsymbol{P}$ and the environmental factors $\boldsymbol{E}$ is based on their histories locally and in the neighborhood:

1) Choose $k \sim Multinomial(\boldsymbol{\pi})$
2) Choose A_{dk}, B_{dk} corresponding to each k
3) Find the N most influential neighbor locations with the cause-and-effect relations for time period $[t_1^d, t_2^d]$. ($\tilde{\boldsymbol{P}}_{\boldsymbol{s_0 t}}^{[N]}$ represents the air pollutants at the most influential N neighbor locations, and we rank them based on $T_{s_n \to s_0}$ in descending order)
4) For each of the $T^d = [t_1^d, t_2^d]$ timestamps t:
   - Generate $\boldsymbol{p}_{st} \sim \Pr\left(\boldsymbol{P}_{st} \middle| \boldsymbol{P}_{st}^{(L)} \oplus \tilde{\boldsymbol{P}}_{st}^{[N]}, A_{dk}\right)$
   - Generate $\mathbf{e}_{\mathbf{st}} \sim \Pr\left(\mathbf{E}_{\mathbf{st}} \middle| \mathbf{E}_{\mathbf{st}}^{(\mathbf{L})} \oplus \tilde{\mathbf{E}}_{\mathbf{st}}^{[\mathbf{N}]}, \mathrm{B_{dk}}\right)$, with the locations in $\tilde{\boldsymbol{E}}_{st}^{[N]}$ the same with $\tilde{\boldsymbol{P}}_{st}^{[N]}$

B. Training Datasets Selection

We select the training datasets by cropping the time windows reflecting the increasing (+) and decreasing (−) periods of air pollution. There are two reasons for selecting the (+) and (−) periods instead of randomly selecting the time windows: 1) we care mostly about the variations, e.g. why the air pollution increases to an unhealthy level or what causes the air pollution to decrease. 2) This way of data selection will help the coupled model converge more efficiently, since otherwise there will be many fluctuation periods and it will be hard to recognize their patterns. Fig. 6(a) illustrates the logic of time window cropping of the air pollutants. For each time series, we detect the peaks and bottoms, and choose the adjacent bottom to peak as the start and ending of an increasing period $[t_1^d, t_3^d]$ (similar for the decreasing period selection which is $[t_3^d, t_5^d]$). Note that when there are many pollutant time series at a location s, we observe almost all the (+) or (−) periods of these time series have the same trends. Therefore, the final cropped period is a "union" of an increasing or decreasing case for n_p air pollutants ($[t_1^d, t_4^d]$, as Fig. 6(b) shows). After cropping time windows, we put the air pollutants datasets $\boldsymbol{P}_{st}, \boldsymbol{P}_{st}^{(L)}, \tilde{\boldsymbol{P}}_{st}^{[N]}$ and the meteorological data at the same locations $\boldsymbol{E}_{st}, \boldsymbol{E}_{st}^{(L)}, \tilde{\boldsymbol{E}}_{st}^{[N]}$ into each training document.



(a) Cropping of the increasing and decreasing periods (b) Cropping for multiple pollutants

Fig. 6. The logic for training datasets selection.

C. Parameter Learning

Learning the parameters A, B for the GBNs and the hidden variable $\boldsymbol{K}, \boldsymbol{\pi}$ for the Gaussian mixture clustering is an expectation maximization (EM) problem [24]. The EM algorithm iteratively estimates the parameters that maximize the log likelihood of the observed air pollutants and meteorological data. In the E-step, we calculate the expectation of log likelihood (Equation (10)) with the current parameters, and the M-step recomputes the parameters of the coupled causality model.

1) E-step: Given the existing parameters A, B and $\boldsymbol{K}, \boldsymbol{\pi}$, EM assumes the membership probability of a document d belonging to the k_{th} cluster to be:

$$\Pr(k|d) = \frac{\Pr(k)\Pr(Z^d|k)}{\Pr(Z^d)} = \frac{\pi_{\mathrm{dk}} \boldsymbol{N}(\boldsymbol{p}_{st}^{\boldsymbol{d}} | \boldsymbol{\mu}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{k}}, \Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{k}})) \boldsymbol{N}(\boldsymbol{e}_{st}^{\boldsymbol{d}} | \boldsymbol{\mu}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{k}}, \Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{k}}))}{\sum_{j=1}^{K} \pi_{\mathrm{dj}} \boldsymbol{N}(\boldsymbol{p}_{st}^{\boldsymbol{d}} | \boldsymbol{\mu}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{j}}, \Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{j}})) \boldsymbol{N}(\boldsymbol{e}_{st}^{\boldsymbol{d}} | \boldsymbol{\mu}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{j}}, \Sigma(\boldsymbol{\varepsilon}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{j}}))} \quad (11)$$

2) M-step: The membership probability is used to calculate new parameters. Determining the most likely assignment tag of each document d to cluster k, i.e. $Tag_d = \max_{k \in [1,K]} \pi_{dk}$, we union the timestamps in the documents in the k_{th} cluster to a new document $d_k^{Union} = Union(d)_{Tag_d = k}$, thus obtaining new parameters $A_{dk}^{new}, B_{dk}^{new}$ for GBNs by solving the following two regressions with new documents:

$$\boldsymbol{P}_{st} = \boldsymbol{\mu}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{0}} + \left(\boldsymbol{P}_{st}^{(L)} \oplus \tilde{\boldsymbol{P}}_{st}^{[N]}\right) A_{dk}^{new} + \boldsymbol{\varepsilon}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{k}}$$
$$\boldsymbol{E}_{st} = \boldsymbol{\mu}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{0}} + \left(\boldsymbol{E}_{st}^{(L)} \oplus \tilde{\boldsymbol{E}}_{st}^{[N]}\right) B_{dk}^{new} + \boldsymbol{\varepsilon}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{k}}$$

The conditional dependencies $\Pr(\boldsymbol{P}_{st} | \boldsymbol{P}_{st}^{(L)} \oplus \tilde{\boldsymbol{P}}_{st}^{[N]}, A_{dk})$ and $\Pr(\boldsymbol{E}_{st} | \boldsymbol{E}_{st}^{(L)} \oplus \tilde{\boldsymbol{E}}_{st}^{[N]}, B_{dk})$ obey Gaussian distributions, thus we update the means and variances of GBNs to:

$$\boldsymbol{\mu}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{dk}} = \boldsymbol{\mu}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st}}^{\boldsymbol{0}} + \left(\boldsymbol{P}_{st}^{(L)} \oplus \tilde{\boldsymbol{P}}_{st}^{[N]}\right) A_{dk}^{new}$$
$$\boldsymbol{\mu}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{dk}} = \boldsymbol{\mu}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}^{\boldsymbol{0}} + \left(\boldsymbol{E}_{st}^{(L)} \oplus \tilde{\boldsymbol{E}}_{st}^{[N]}\right) B_{dk}^{new}$$
$$\Sigma\left(\boldsymbol{\varepsilon}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})}^{\boldsymbol{dk}}\right) = \Sigma(\boldsymbol{P}_{st}) - A_{dk}^{new} \cdot \boldsymbol{\Sigma}\left(\boldsymbol{P}_{st}^{(L)} \oplus \tilde{\boldsymbol{P}}_{st}^{[N]}\right)^{-1} \cdot A_{dk}^{new\mathrm{T}}$$

$$\Sigma\left(\boldsymbol{\varepsilon}^{dk}_{\boldsymbol{E}_{st}|\boldsymbol{PA}(\boldsymbol{E}_{st})}\right) = \Sigma(\boldsymbol{E}_{st}) - B^{new}_{dk} \cdot \boldsymbol{\Sigma}\left(\boldsymbol{E}^{(L)}_{st} \oplus \tilde{\boldsymbol{E}}^{[N]}_{st}\right)^{-1} \cdot B^{new\mathrm{T}}_{dk} \quad (12)$$

And, we update π_{dk} by:

$$\pi^{new}_{\mathrm{dk}} = \sum_{d=1}^{M} \frac{\Pr(k|d)}{M} \quad (13)$$

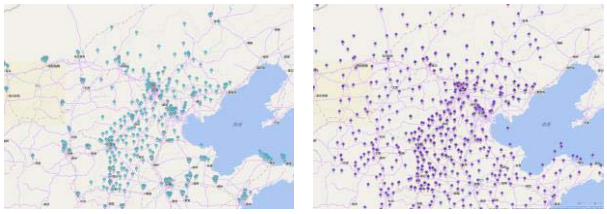
In this way the coupled model iteratively unions the documents d^{Union}_{k} and trains the parameters with new union datasets until convergence, thus overcoming the local optima problem of traditional EM algorithm. Details of EM algorithm of the coupled model are elaborated in Algorithm 1:

Algorithm 1: EM learning of the coupled model

Input: 1) Air pollutants and meteorological datasets in documents $\{\boldsymbol{P}^{d_1}, \boldsymbol{P}^{d_2}, \ldots, \boldsymbol{P}^{d_M}\}$ **&** $\{\boldsymbol{E}^{d_1}, \boldsymbol{E}^{d_2}, \ldots, \boldsymbol{E}^{d_M}\}$, 2) *K:* Number of clusters
Output: A_{dk}, B_{dk} for the he ST cause-and-effect relations in GBNs
Repeat:
  For d = 1 to M
    For k = 1 to K
      Update π_{dk} by Equation (13)
    End
  End
  For k = 1 to K
    Update A_{dk}, B_{dk}, plus the means and variances for GBNs
  End
  Update the cluster assignment $\Pr(k|d)$ based on Equation (11)
Until: convergence;

IV. Evaluation

We use 6 air pollutants and 5 meteorological data from 52 cities, 515 air quality monitoring stations, 404 meteorology stations in Huabei, the center region of China, during Jun 1st 2013 – May 1st 2015. The urban data are: PM2.5, PM10, NO_2, CO, O_3, SO_2, as well as temperature (T), pressure (P), humidity (H), wind speed (WS), and wind direction (WD). The spatial range is with 35N-43N, and 110E-123E. Fig. 7 visualizes the locations of the urban data for experiments. We average the meteorology value around each air quality monitoring station, to make the locations of meteorological data consistent with the air pollution data.



(a) Air quality monitoring stations **(b) Meteorology stations**

Fig. 7. Locations of air quality and meteorological data.

A. Evaluation of Component 1

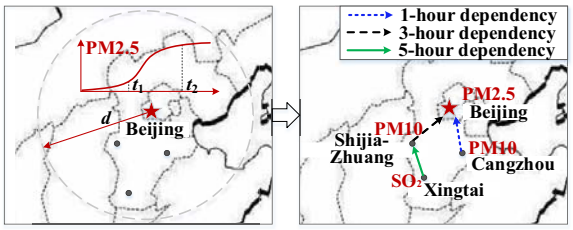
To evaluate Component 1, i.e. the GBN-based ST representation of cause-and-effect relations between air pollutants, we compare the time efficiency for learning the GBN structure in the ST space with 4 baseline Bayesian-based causal structure learning algorithms. The baselines for Bayesian structure learning are: 1) hill climbing, 2) Markov-chain Monte Carlo (MCMC), 3) K2, and conditional independency (CI) test. These methods usually encode the cause-and-effect relations into a directed acyclic graph (DAG), and learn the structure by maximizing a score, e.g. BIC score or K2 score [21]. Hill climbing and K2 are greedy-based algorithm, while MCMC samples and updates the structure by "adding", "subtracting", or "reversing" connections. CI test is a pairwise conditional probability based method which verifies the dependency by χ^2 test [22]. The global structure is generated by connecting the link with the maximum conditional dependency, and later assigns the direction of each connection based on a *d*-separation rule [3].

Our method and four baselines are realized by MATLAB BNT toolbox [23]. Table I shows the time for learning the causal structures. Our method provides nearly linear scalability in time, with about 4.6 hours for learning ST causality at 52 cities and about 47.4 hours for learning 515 air quality monitoring stations. This suggests the ST transfer entropy based algorithm can be a feasible way to locate the causes for air pollution with big urban data. To learn the causal structure with 515 stations, the first three baselines either takes very long time or fails to compute. The last baseline has superiority in time but fails to perform accurately in the prediction task for Component 2.

TABLE I. Time for Learning the Causal Structure

Time (s)	*Our Method*	*Hill climbing*	*MCMC*	*K2 + PS*	*CI test*
52 cities	16624	20204.3	82947	14102	3562
515	170802	--	--	202543	18463

Fig. 8 presents an example of the learnt cause-and-effect representation in the ST space. In Fig. 8(a), PM2.5 in Beijing increases to a very unhealthy level during a time interval $[t_1, t_2]$. We want to know the locations of causes during the selected time interval. Given a spatial range d as input, we learn and present the specific causes for the increase of Beijing's PM2.5 in Fig. 8(b). The illustrated structure suggests the increase of PM2.5 in Beijing is most likely caused by: 1) the PM10 of Cangzhou city one hour ago; 2) the PM10 of Shijiazhuang three hours ago. Further, PM10 in Shijiazhuang is likely to be caused by SO_2 in Xingtai city five hours ago. We also visualize the time dependency for each neighbor location by showing the time lag with the maximum Pearson correlation with the target pollutant.



(a) Input of the model with spatial range *d* and temporal range $[t_1,t_2]$ **(b) Cause-and-effect relations given spatio-temporal ranges**

Fig. 8. An example of the ST cause-and-effect representation.

B. Evaluation of Component 2

Component 2 is evaluated by predicting the future air pollutants with historical and neighborhood air pollutants with the parameters we learnt in the GBNs. We use two data segments, Jun 1st 2013 – April 1st 2014 and Jun 1st 2014 - April 1st 2015 for training, with the corresponding data in April 1st 2014 – May 1st 2014 and April 1st 2015 - May 1st 2015 for verification. Prediction precision is calculated and averaged by the two verification segments. To avoid biased prediction caused by meteorology, we calculate the expected air pollution values with respect to the different confounding variable *K*, i.e.

$$\boldsymbol{P}_{st} = \boldsymbol{\mu}^{\mathbf{0}}_{\boldsymbol{P}_{st}|\boldsymbol{PA}(\boldsymbol{P}_{st})} + \sum_{k=1}^{K} \Pr(k|d) \cdot \left(\boldsymbol{P}^{(L)}_{st} \oplus \tilde{\boldsymbol{P}}^{[N]}_{st}\right) A_{dk}. \quad (14)$$

We first conduct a 1-hour prediction task, to study the effects of neighbor location number *N* and cluster number *K* on the

prediction accuracy. As illustrated in Fig. 9, the general trend of the precision of 1-hour prediction task increases first and then decreases. N=3 and K=8 present the best performance. When K>15, the performance will degrade due to over-fitting. Note that when the number of influential neighborhood locations N=1, the precision fluctuates a lot as K increases. This may suggest the air pollutants may not just be caused by air pollutants at 1 neighborhood location.

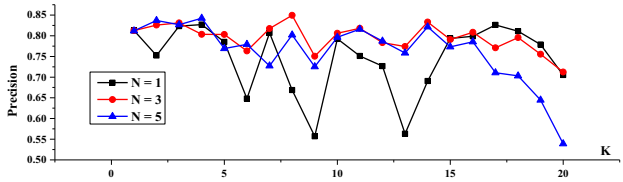
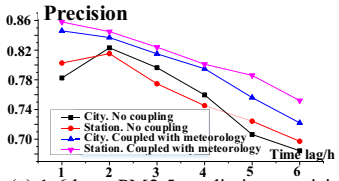


Fig. 9. Precision of 1-hour PM2.5 prediction for 515 monitoring stations, with different N and K in the coupled model.

Afterwards, we verify the coupling idea based on the prediction precision, coupled or not coupled with meteorology. Fig. 10(a) shows 1-6 hour prediction precision over all the datasets. Generally the model with coupling outperforms the model without coupling. With coupling, we can achieve over 84% precision for 1-hour prediction and over 74% precision for 6-hour prediction in station level. This suggests our proposed model is capable of identifying the causalities between air pollutants and understand how they interact. Note that without coupling, the 2-hour prediction outperforms 1-hour prediction. This may be due to the air pollutants at city level being more likely to be "caused" by neighborhood locations 2 hour ago. We also compare our model with/without coupling with four baselines for 1-hour prediction, as shown in Fig. 10(b). Results show our method obviously outperforms the baselines.



(a) 1-6 hour PM2.5 prediction precision coupled or not with meteorology

Method	1-hour prediction precision	
	City	Station
Our method	0.804	0.859
Our method without coupling	0.789	0.832
Hill climbing	0.528	--
MCMC	0.597	--
K2 + PS	0.684	0.753
CI test:	0.382	0.298

(b) 1-hour prediction precision compared with four baselines

Fig. 10. Precision of PM2.5 prediction based on our proposed model with and without meteorology coupling.

V. Conclusion

We propose a coupled Gaussian Bayesian model to identify the causalities for air pollutants in the ST space. The model comprises two components, 1) a GBN-based cause-and-effect representation to locate where air pollution comes from, and 2) a coupled model that integrates meteorology into the cause-and-effect relations, thus eliminating the impact of confounding. Evaluation with real-world urban big data shows both time efficiency and prediction accuracy of this model. We further plan to modify the model for online causality analysis and provide real-time poor air quality warnings.

References

[1] Y. Zheng, L. Furui, and H.P. Hsieh. "U-Air: When urban air quality inference meets big data." In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1436-1444. ACM, 2013.

[2] 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 Proceedings of the 21th ACM SIGKDD.

[3] J. Pearl. Causality: Models, Reasoning and Inference. Vol. 29. Cambridge: MIT press, 2000.

[4] D. Heckerman. "A tutorial on learning with Bayesian networks." In Innovations in Bayesian Networks, pp. 33-82. Springer Berlin Heidelberg, 2008.

[5] D.B. Rubin."Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of educational Psychology 66, no. 5 (1974): 688.

[6] P.W. Holland. "Statistics and causal inference." Journal of the American statistical Association 81, no. 396 (1986): 945-960.

[7] P. R. Rosenbaum, and D. B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70, no. 1 (1983): 41-55.

[8] D. C. Kulp, and M. Jagalur. "Causal inference of regulator-target pairs by gene mapping of expression phenotypes." BMC genomics 7, no. 1 (2006): 125.

[9] W. Sun, P. Wang, D. Yin, J. Yang, and Y. Chang. "Causal inference via sparse additive models with application to online advertising." In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[10] T. M. Snowsill, N. Fyson, T. D. Bie, and N. Cristianini. "Refining causality: who copied from whom?." In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 466-474. ACM, 2011.

[11] K. Radinsky, S. Davidovich, and S. Markovitch. "Learning causality for news events prediction." In Proceedings of the 21st International Conference on World Wide Web, pp. 909-918. ACM, 2012.

[12] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. "Discovering spatio-temporal causal interactions in traffic data streams." In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1010-1018. ACM, 2011.

[13] C. Hashimoto, K. Torisawa, J. Kloetzer, and J. H. Oh. "Generating Event Causality Hypotheses through Semantic Relations." In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[14] P. R. Rosenbaum, and D. B. Rubin. "Reducing bias in observational studies using subclassification on the propensity score." Journal of the American Statistical Association 79, no. 387 (1984): 516-524.

[15] K. H. Hullsiek, and T. A. Louis. "Propensity score modeling strategies for the causal analysis of observational data." Biostatistics 3, no. 2 (2002): 179-193.

[16] C. Glymour, R. Scheines, and P. Spirtes. Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. Academic Press, 2014.

[17] M. Grzegorczyk. "An introduction to Gaussian Bayesian networks." In Systems Biology in Drug Discovery and Development, pp. 121-147. Humana Press, 2010.

[18] 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 (2010): 145.

[19] J. Y. Zhu, C. Sun, and V. O. K. Li. "Granger-Causality-based air quality estimation with spatio-temporal (ST) heterogeneous big data." In Computer Communications Workshops (INFOCOM WKSHPS), 2015 IEEE Conference on, pp. 612-617. IEEE, 2015.

[20] L. Barnett, A. B. Barrett, and A. K. Seth. "Granger causality and transfer entropy are equivalent for Gaussian variables." Physical Review Letters 103, no. 23 (2009): 238701.

[21] G. F. Cooper, and E. Herskovits. "A Bayesian method for the induction of probabilistic networks from data." Machine Learning 9, no. 4 (1992): 309-347.

[22] H. Ku, and S. Kullback. "Approximating discrete probability distributions." Information Theory, IEEE Transactions on 15, no. 4 (1969): 444-447.

[23] K. Murphy. "The bayes net toolbox for matlab." Computing science and statistics 33, no. 2 (2001): 1024-1034.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society. Series B (methodological) (1977): 1-38.