IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 4, 2020, accepted November 2, 2020, date of publication November 6, 2020, date of current version November 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036354

Visual Analytics of Air Pollution Propagation Through Dynamic Network Analysis

KE REN[1,2], YIYAO WU[1,2], HUIJIE ZHANG[1,2], (Member, IEEE), JIA FU[1,2], DEZHAN QU[1,2,3], AND XIAOLI LIN[4]

[1]School of Information Science and Technology, Northeast Normal University, Changchun 130000, China
[2]Key Laboratory of Intelligent Information Processing of Jilin Universities, Changchun 130000, China
[3]Library, Northeast Normal University, Changchun 130000, China
[4]School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China

Corresponding author: Huijie Zhang (zhanghj167@nenu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 41671379.

ABSTRACT The propagation of pollutants between regions has become a noticeable factor affecting air pollution. Given the complicated propagation relationship, most of the existing works lack an effective perception mechanism of geographic correlations and time-varying features, which is crucial in exploring and understanding the propagation mechanism by integrating empirical knowledge and data inherent characteristics. In this paper, we abstract the complicated propagation relationship between regions as a dynamic network, and introduce visual analytics techniques to explore the spatiotemporal multivariate patterns of air pollution propagation. A particle tracking based model is first proposed to construct pollution propagation networks under multi-source factors. It combines numerical simulation and data characteristics simultaneously, and detects active pollution source areas based on long-term transport relationships and temporal correlations. Based on it, we extract propagation patterns and analyze the temporal evolution of diachronic propagation networks. Moreover, we design an interactive system to achieve an in-depth analysis of air pollution issues. Through elaborate multi-level glyphs and linkage views, the system facilitates users to perceive and explore propagation mechanism in spatiotemporal multivariate information, and compare propagation patterns from global and local perspectives. We present several case studies to demonstrate the usefulness of our work in air pollution propagation analysis.

INDEX TERMS Visualization, air pollution propagation, network visualization, spatiotemporal multivariate feature.

I. INTRODUCTION

Air pollution has become one of the highly focused environmental issues for the public and scientists [1]–[3]. Pollutants are absorbed in gases and propagated under different climatic conditions. Finding the sources of pollution and the propagation patterns between different regions is critical to preventing and controlling air pollution [4]. A large amount of air quality data and meteorological data are collected in real-time by widely deployed sensors, making it possible to analyze air pollution propagation problems from a data science perspective.

The previous analysis works are mainly carried out in two aspects [5]–[10]. One is to explore the co-occurrence relationships of air pollution in different regions [5]–[7], to identify the hidden laws of propagation. However, this kind of method's accuracy is heavily dependent on the co-occurrence interval and is not accepted by scientists. The other is constructing an ideal model to simulate the pollution spread in the wind field [8]–[10]. They commonly focus on a specific location and have limited usefulness in detecting the transport trend in a short period. Whereas, the independent simulating results with significant uncertainty bring great challenges for defining pollution sources. Besides, they ignore the transport difference between different pollutants and fail in analyzing the propagation path of specific pollutants. Thus, it is imperative to establish a method combining the model simulation with historical data features, comprehensively analyzing the long-term and large-scale multivariate pollution propagation mechanism. Such a technique is helpful for the formulation of time-space control policies.

Based on the pollutant transport relationship, a target area and multiple source areas form a pollution propagation

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

network, and it evolves with structures changing over time [11]. It is still an analytical difficulty of understanding the temporal evolution of the dynamic network in the context of multivariate geographic information. The most common way to solve this problem is to observe the network screenshots of different pollutants in chronological order [12]. However, it relies on human cognition ability and is tedious for environmental experts. In dynamic network visualization, some graph embedding methods have been proven successful in analyzing network features [13]–[15]. They characterized network nodes or the whole structures with representations in a vector space and then interpreted by visual analytics technology. It is essential to employ suitable vectorization strategies and design intuitive visual representations according to the data characteristics. There is currently a lack of visual analytics methods for exploring the air pollution propagation network with spatiotemporal multivariate features. Moreover, novel data-driven views or glyphs should be designed to demonstrate and summarize air pollution propagation mechanism.

To this end, we propose a novel visual analytics methodology for identifying the air pollution sources and extracting propagation patterns. To track pollutant particles, we construct an air pollution propagation model to quantify the point-to-point transport values, which takes into account the impacts of meteorological conditions and the co-occurrence of pollutants. In this way, a group of active source areas corresponding to each target area can be detected based on long-term pollution transport relationships. At the same time, we build propagation networks centering on each target area, with its source areas involved. Considering that the networks continuously change with time, we further extract different structural patterns and evaluate their importance to analyze time-varying laws of the propagation networks. Moreover, we design and develop an interactive visual analytics system. Several novel visual designs are proposed, including a two-level glyph for summarizing multivariate transport conditions of sources in different periods, and interactive views for comparing propagation patterns in the context of structure and temporal information. The contributions of this work are the following:

1) **A model integrating simulation and data features for constructing dynamic air pollution propagation network and detecting active source stations.** It takes into account not only the point-to-point meteorological propagation effects but also the time-varying similarity of pollutants, which provides a data basis for air pollution propagation analysis.
2) **A novel two-level glyph design as the visual representation of pollution propagation.** It shows multivariate, time-varying, and geographical features from the summary and detail levels to fulfill the needs of analyzing the pollution situation of the concerned station.
3) **Case studies based on real data.** Taking Baoding, China as an example, summarizing the impact of surrounding areas on air pollution from multiple aspects and analyzing the time-varying laws of the propagation patterns.

II. RELATED WORK

This section presents prior studies in this article from two aspects: analysis and visualization of air pollution propagation and network visualization.

A. ANALYSIS AND VISUALIZATION OF AIR POLLUTION PROPAGATION

There is a large number of researches on air pollution propagation [16], which mainly focus on utilizing ideal simulating models or mining the co-occurrence relationships to find primary propagation paths. HYSPLIT is one of the most widely used simulating models for analyzing pollution sources [9], [17]. It is used to track the trajectories of individual particles in a given area at each time step. However, this model only focuses on the retention period, neglects the transport difference between different pollutants during air pollution propagation processes. To address these limitations, some articles utilize LDM (Lagrange dispersion model) [10], [18], [19], which take into account the effect of wind speed on the diffusion of pollution particles. It can well simulate the process of pollutants diffusing in the atmosphere.

On the other hand, some works focus on mining the hidden co-occurrence relationships of pollution from historical data by association rule mining methods, thereby inferring the relationship of pollution propagation. For example, Akbari *et al.* [6] proposed a co-occurrence pattern mining method, which extracts the implicitly contained spatial relationships algorithmically and allows mining a spatiotemporal co-occurrence pattern simultaneously in space and time. In this way, the method was applied to a real case study for air pollution, where the objective is to find correspondences of air pollution with other parameters that affect this phenomenon. Recently, He *et al.* [5] proposed an adaptive spatiotemporal episode pattern mining algorithm, which can discover the candidate driving factors for the occurrence of complex geographic events. The proposed approach was applied to analyze the air pollution in the region of Beijing-Tianjin-Hebei. Although the existing methods have been proven useful for air pollution analysis, they still have other restrictions. These models or methods are often used independently and lack auxiliary background information and necessary interactions in the analysis process. For example, when using HYSPLIT to analyze the propagation difference of different seasons, users can only use the color to code time attributes of the path lines. However, this simple visual method is not intuitive when the number of path lines is massive, or the path lines are interlaced seriously. For the post-analysis of model results, more data mining methods and visual analytics technology can be introduced so that users can more conveniently explore the time-varying laws of pollution propagation.

For exploring the air pollution propagation deeply, visualization has been introduced into recent studies as a powerful tool to bridge the gap between human cognition and automated tools. These methods utilize plenty of linked views (polar view, thermodynamic chart, etc.) to reveal he time-varying law of air pollution, the correlation between pollutants and the difference among geographical regions [20]–[23]. Zhang *et al.* [20] proposed a visual system and novel glyph designs to explore regular multivariate patterns as well as abnormal temporal and spatial cases in air pollution. Guo *et al.* [22] proposed Time-Correlation Partitioning (TCP) tree that represents the correlation of multiple pollutants and their evolutions. Zhou *et al.* [21] proposed a storyline design to depict the evolution of different stations. Different from these works, we focus on the spatiotemporal multivariate patterns of air pollution propagation based on air quality and meteorological data. Qu *et al.* [24] used parallel coordinates to analyze the relationship between different pollutants and meteorological factors. However, these views have severe occlusion when handling large-scale data. There is also a lack of interactions in these visualization designs, leading obstacles in understanding the pollution propagation. Deng *et al.* [25] proposed a novel visual system, which supports extracting and interpreting the uncertain air pollution propagation patterns. They extracted patterns based on frequent propagation routes and presented temporal uncertainty of patterns by glyph designs.

Inspired by the works mentioned above, we propose a visual analytics system based on a comprehensive pollutant particle tracking model and novel visual representations. Specifically, our work focuses on the air quality monitor stations, with extracting the spatiotemporal patterns of propagation from other surroundings, which makes it easier for experts to propose a joint plan of urban air quality management.

B. NETWORK VISUALIZATION

The dynamic geographic network is an important research object in the field of visual analytics. It represents a network structure with locations or regions as nodes and the relationships between them as edges. It has a wide range of applications in international trade, group movement, and transportation.

The OD (Origin to Destination) matrix is the most direct method for visualizing geographic networks. However, due to the lack of geographic location information, analysts can't discover geographic-related patterns. Yang *et al.* designed the MapTrix view [26], combining the starting point map, the ending point map and the OD matrix. It can solve the problem of visualization of many-to-many flowing data on the map. The flow graph is another method widely used to display the structure of geographic networks. Still, this expression's visual confusion will increase significantly as the complexity of the data increases. Filtering [27], edge binding [28], and geographic aggregation [29] are commonly used optimization methods to improve the readability of flow graphs. However, there is still a lack of clear and effective visualization methods for the air pollution propagation network and multivariate characteristics.

For dynamic network evolution analysis, animation and timeline are two commonly used methods [30]–[32]. They display the network structure of a single time step frame by frame or side by side. However, this type of method relies heavily on the analyst's recognition ability, and how to intuitively and accurately display the changing laws of network features is still a challenge. In recent years, some efforts try to embed graphs in a vector space and visualization. They usually reduce snapshots or nodes of the network to points in a high-dimensional space by deriving representations from some feature extraction methods. Then project snapshots or nodes to scatter points in 2D space or 1D space and connect the scatter points in chronological order. In this way, users are enabled to detect stable, recurring, or abnormal status during the network evolution. For example, Elzen *et al.* [33] considered the adjacent matrices of snapshots as points in a vector space and projected them to 2D space. To assist visual analytics, they also proposed two juxtaposed views, one for showing a snapshot and the other for showing the network's evolution. Xu *et al.* [34] extracted latent representations of nodes' structural and temporal features by a novel diachronic node embedding method, and then projected nodes to the 1D space and formed the trend view for the evolution visualization. Hajij *et al.* [35] proposed a novel method using persistent homology to quantify structural changes in time-varying graphs and performing visual analytics in the form of a timeline. Similarly, Linsen *et al.* [36] converted the network structure at each time step into the state of the node, Law *et al.* [13] defined the sequence of events as the feature vector of the self-network, so that the overall evolution of the dynamic network is revealed by measuring the changes in features. As for the dynamic air pollution propagation network, there is currently very little work to analyze its timing changes and extract significant patterns.

Aiming at the dynamic network of air pollution, we propose a pattern extraction method based on propagation features, and provide a system to support interactive exploration and interpretation of the multivariate, geographical, and time-varying laws of the extracted patterns.

III. TASKS AND OVERVIEW

After conducting literature review and discussions with experts, we finally summarize three analytical tasks as follows:

1) **T1: Explore an overview of a dynamic network.** For a propagation network centered with a specific target station, the geographical distribution of the source stations, the temporal transport laws, and the transport values of various pollutants should be demonstrated intuitively, which is helpful for experts to quickly understand how an area is affected.

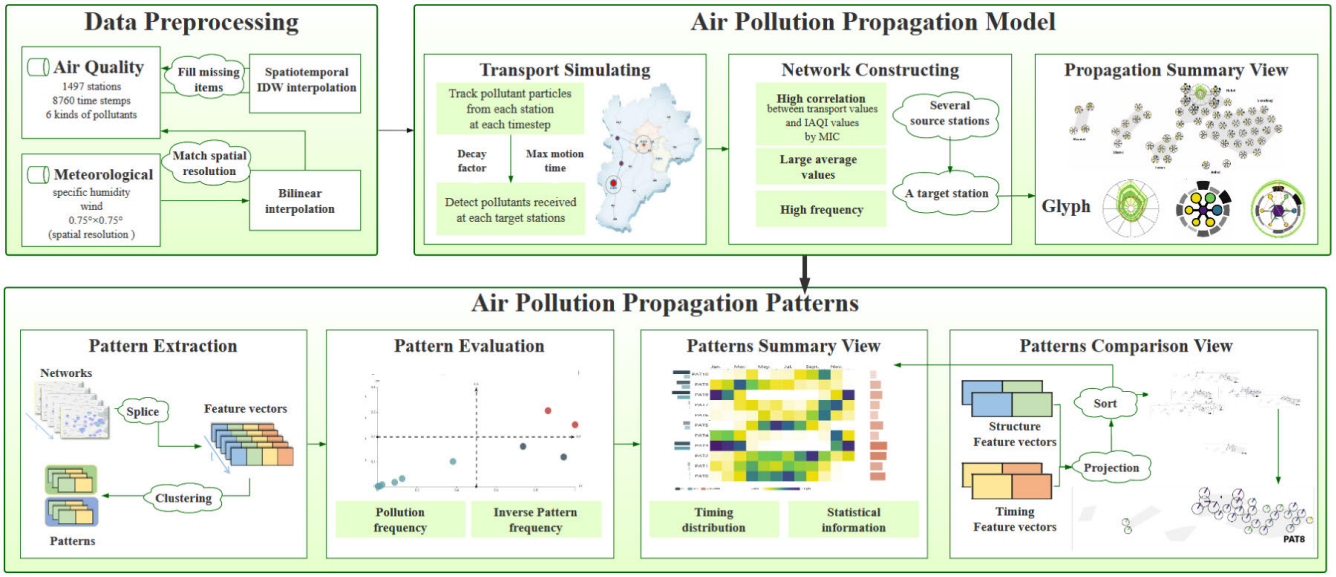


FIGURE 1. The visual analytics framework.

2) **T2: Extract patterns from the evolution of the network.** The experts aim to find what similar propagation network structures repeat diachronically. Extracting such typical patterns is essential for analyzing which areas are easily coordinated to transport pollution and defining pollution sources for different periods.
3) **T3: Reveal the differences between patterns.** Each pattern involves several networks of different moments, associating with specific pollution source nodes and edges information about the transport of pollutants. Experts want to compare these patterns based on the temporal distribution, the topology properties, and the impact on the target area.

In this paper, air pollution propagation analysis aims to identify the potential sources, explore the temporal laws and the major pollutants of delivery, which are the main factors experts concentrate on. To help users achieve this goal, we propose a visual analytics framework (FIGURE 1) that comprises three major phases: (1) data preprocessing, (2) air pollution propagation model, and (3) air pollution propagation patterns.

The data used in this paper consists of an air quality dataset and a meteorological dataset. We employ spatiotemporal IDW interpolation to fill the missing values and bilinear interpolation to match these two types of data's spatial resolutions.

Given the multi-source heterogeneous spatial data, we propose an air pollution propagation model that tracks pollutant particles' movement and detects pollution received at the target stations. This model consists of two parts. One is the transport simulation based on particle tracking, which considers the role of humidity and motion time in pollutant transport compared to traditional methods. Another is the model optimization based on data characteristics, which detects the active source stations group of each target station according to the average value, frequency, and correlation. On this basis, we design a propagation summary view with several intuitive glyphs, which provides the context of spatiotemporal multivariate information, and reveals how a target station is polluted by the surrounding areas.

Furthermore, to exploit the air pollution propagation to facilitate the exploration of the structural and temporal properties from the synergistic pollution source perspective, we extract patterns from a focused target station's propagation networks. First, each network's structural characteristics are represented by a high-dimensional feature vector reflecting six kinds of pollutant values transported from active source stations. A clustering algorithm then groups networks into patterns where they have similar features. We propose pollution frequency (PF) and inverse pattern frequency (IPF) to evaluate the importance of patterns, to provide guidance for further exploration. In addition, we design the patterns summary view and the patterns comparison view to demonstrate the statistical information, timing distribution and structural features of patterns.

Finally, we develop a system integrating the above methods and the linked views, such as map view and word clouds view. It supports the complete analysis process when users select a station of interest.

IV. DATA PREPROCESSING

Air quality is monitored and recorded by countless monitoring stations ceaselessly through the whole world. The air quality data used in this paper are crawled from the website PM25.in (http://www.pm25.in/). This dataset involves 1497 stations covering 375 cities in China and in the time range of 2015. Each station's geographical information consists of longitude, latitude, station number, station name,

and the city where it locates in. Each station records data every hour, comprising the concentrations of six kinds of air pollutants: $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 and CO. Therefore, the air quality data are featured with spatiotemporal, high-dimensional and large-scale properties.

Several missing and wrong items inevitably exist in the collected air quality data due to the incidental cases of the power outage and sensor damage. Since the missing values exist from both spatial and temporal views, traditional geographical interpolations, such as IDW and Kriging, only focus on spatial distance and ignore the continuity in time. For some sparsely distributed sites, using their own adjacent time value interpolation is often more accurate than the value on geographic adjacent. To solve this problem, we fill them in by utilizing a spatiotemporal interpolation strategy [37]. It improved traditional inverse distance weighting (IDW), by considering not only the data of the surrounding areas but also the adjacent timestamps, to achieve accurate results.

Besides, different kinds of pollutants cannot be directly compared by concentration values because of its various units. Thus, we normalize each kind of pollutant by calculating individual air quality index (IAQI) based on their concentrations. The higher the value is, the more serious the pollution and the more harmful to humans it means. As shown in FIGURE 2, air monitoring stations are marked as circles on the map, and circles' colors encode the annual average values of $IAQI$. It is obvious that the pollution of stations in the north is more serious than the stations in the south.

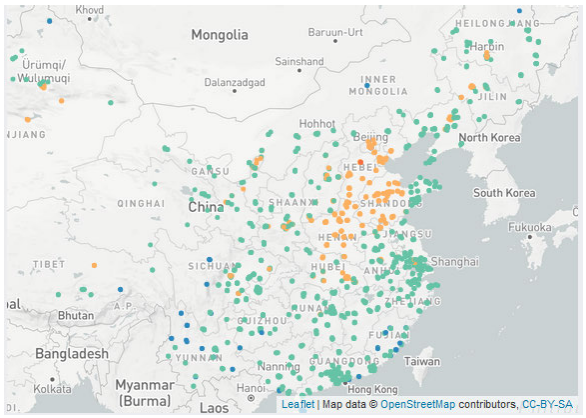


FIGURE 2. Spatial distribution of air quality monitoring stations in this research.

As for the meteorological data, we obtain it from European Centre for Medium-Range Weather Forecasts (ECMWF), which records specific humidity, U velocity and V velocity every three hours at the spatial resolution $0.75° \times 0.75°$. Table 1 details the three attributes.

Due to the difference between the spatial resolution of meteorological data and air quality data,, meteorological data are spatially interpolated to attain the spatial resolution matching air monitoring stations by the bilinear interpolation method.

TABLE 1. Variable description of ECMWF meteorological data.

Variable	Symbol	Description	Unit
Wind U velocity	U	U>0 means that the wind field component is northward, U<0 means that the wind field component is southward. |U| means the wind speed in the corresponding direction.	m/s
Wind V velocity	V	V>0 means that the wind field component is eastward, V<0 means that the wind field component is westward. |V| means the wind speed in the corresponding direction.	m/s
Specific humidity	S	The ratio of the mass of water vapor in air to the total mass of the mixture of air and water vapor.	kg/kg

V. AIR POLLUTION PROPAGATION NETWORK MODEL

A. TRANSPORT SIMULATING

Meteorological properties are essential factors affecting the propagation of air pollution. The speeds and directions of wind play a decisive role in the transport velocities and sedimentation of pollutants. The higher the wind speed is, the farther the air pollutants are transported, and the lower the concentration retains. At the same time, precipitation can purify the pollutants in the atmosphere. Taking the influence of meteorological factors into account, we propose a particle tracking method, which makes full use of the collected multi-source heterogeneous data and quantifies pollutant transport between stations.

In the domain of vector field research, Lagrangian Particle Dispersion model [38] is a common method to get traces by setting some particles and tracking their moving positions. Inspired by this, we regard pollutants with specific concentration as particles starting from each station, the wind field as a vector field. The transport of pollutants between different districts can be equivalent to particles' movement in the vector field. In this way, we apply air pollution transport simulation.

Obviously, according to each position's velocity vector in the wind field, we can track the trajectory of a particle m according to the following formulas:

$$x_m^{i+1} = x_m^i + V_{p_m^i}^i \Delta t \tag{1}$$

$$y_m^{i+1} = y_m^i + U_{p_m^i}^i \Delta t \tag{2}$$

where x_m^{i+1} and y_m^{i+1} indicate the position of particle m at time step $i+1$, x_m^i and y_m^i indicate the position of particle m at time step i, Δt is the time interval between two time steps. $V_{p_m^i}^i$ and $U_{p_m^i}^i$ are the velocities on position $p_m^i(x_m^i, y_m^i)$ at time step i.

Since the wind field is a volatile velocity field, to improve particle tracking accuracy, we introduce the Runge-Kutta strategy, taking the average speed of the current time step and the next time step as the actual movement velocity of the particle, as shown in the following formulas:

$$V_{p_m^i}^i = (V_{p_m^i}^{i\,\prime} + V_{p_m^i}^{i+1\prime})/2 \tag{3}$$

$$U_{p_m^i}^i = (U_{p_m^i}^{i\,\prime} + U_{p_m^i}^{i+1\prime})/2 \tag{4}$$

where $V'_{p_m^i}$ is the original V vector in the wind field, and $V'_{p_m^i}{}^{i+1}$ is the original V vector at position p_m^i and time step $i+1$. Besides, since the position of the particles moving in the wind field is uncertain, we use bilinear interpolation each time step to calculate the meteorological data at each particle's position.

More complex than particles in traditional vector fields, pollutants' concentrations are inconstant and affected by humidity and wind speed, causing the concentrations to decrease during the propagation process. Therefore, we propose a notion called decay factor (DF) that controls the particle's life during its motion. It is inversely proportional to wind speed and humidity and is defined as:

$$DF_p^i = Vel_p^i * S_p^i. \tag{5}$$

where DF_p^i represents the degree of contaminant particles' attenuation at position p and time step i. S_p^i is humidity of the position p at time step i. Vel_p^i is the speed of the position p at time step i, and it is calculated as follow:

$$Vel_p^i = \sqrt{U_p^{i\,2} + V_p^{i\,2}}. \tag{6}$$

Based on the definitions above, we employ an iterative algorithm to trace pollutants movements. In each iteration, the existing life value of particle m moving from time step i to $i+1$ is calculated as:

$$Life_m^{i+1} = Life_m^i * (1 - DF_{p_m^i}^i * \Delta t). \tag{7}$$

$Life_m^0$ is initialized to the concentration value at the start. The life of particle m is decaying over time, and the particle will stop moving when it decays to zero. During the iteration, we also set a maximum motion time threshold θ, which constrains the motion termination of the particle together.

The whole iterative process is given in Algorithm 1. For each source station a, there are six kinds of pollutants starting at each time step i, moving in the wind field until they vanish. In this way, the trajectories of pollutants can be tracked, and the life value at various locations along the way can be detected. In fact, each station sends pollution while also receiving pollution from others. From another perspective, we treat a station a as the recipient of pollution, when a particle starting from the station b passes through it, the particle's life value can be used as the pollutant transport value from the station b to the station a at timestep i. The particles are checked every timestep if it is within the neighborhood of any station. If the answer is yes, this time information and the life value at that time will be recorded to provide a data basis for the further analysis of the pollution recipient.

Algorithm 1 Transport Simulating

Input: Air quality data, meteorological data, particle motion time threshold θ, station neighborhood radius r

Output: Transport values of six kinds of pollutants at each timestep between stations

```
Initialize an empty particle set M.
for each time step i do
    for each particle m in M do
        if Life_m^i is greater than 0 and the motion time of m is less than θ then
            Calculate the speed of p_m^i where the particle m is located according to Eq.3 and Eq.4
            Update the position of the particle m according to Eq.1 and Eq.2
            Calculate the decay factor of p_m^i where the particle m is located according to Eq.5 and Eq.6
            Update the life Life_m^i of the particle m according to Eq.7
            Update the motion time of the particle m
            for each station a do
                if the particle m passes through the neighborhood (a circle with radius r) of the station a then
                    Record Life_m^i as the transport value from the station b where m starts to the target station a at the current time step i
                end if
            end for
        else
            Remove the particle m from the set M
        end if
    end for
    for each station a do
        Add six new particles to the particle set M. The motion duration of particles are 0, and the life values correspond to the six kinds of pollutant values
    end for
end for
```

B. NETWORK CONSTRUCTING AND POLLUTION SOURCES IDENTIFYING

Based on the proposed pollution transport method, we can track the propagation trajectories of pollutants and detect the impact of source stations on each target station. However, it is built on ideal assumptions, and the real pollution toward target stations are affected by many factors and cannot be simulated. To improve the method, a large amount of data provides the possibility to analyze pollution problems from the data science perspective that comprehensively considers the real impact of pollution propagation. Thus, we further integrate the real pollution co-occurrence relationship between regions to build a network model of air pollution propagation.

In data analysis research, correlation is often used to explore implicit co-occurrence relationships. Inspired by this, we introduce a multivariate timing correlation strategy to quantify the pollution-related strength between the source station and the target station and then filter the ideal propagation network's edges and nodes. In this paper, if one source station's transport value changes synchronously with

the real IAQI value of the target station, the correlation is high. That is, there is a significant air pollution impact that the source station makes on the target station. Here, we employ the maximum information coefficient (MIC) [39], a measure of correlation for a bivariate relationship, and considers six kinds of pollutants in combination comprehensively.

MIC was proposed based on mutual information, which applies the concepts of information theory and probability to continuous data, tries different bin numbers and observes which one will get the largest mutual information value, thus detecting the non-linear correlation between variables. MIC between a source station and a target station can be calculated as:

$$MIC(A, B) = \max_{|A||B|<\sigma} \frac{I[A; B]}{log_2(\min(|A||B|))}. \quad (8)$$

where A is the time-series of the transport value from the source station a to the target station b, and B is a time-series of IAQI values of b. Each time step in A and B contains information on six kinds of pollutants. $I[A; B]$ is the mutual information between A and B, and $|A|$ and $|B|$ are the bin number of two series divisions. The value of σ follows the work of Reshef *et al.* [39], which is equal to the $0.6th$ power of the total amount of data. The range of MIC is between 0 and 1, and higher value indicates a stronger correlation.

In the source stations initially filtered by the stronger correlation, we introduce additional metrics to filter the network further. It is hard to ignore the fact that our model studies six kinds of pollutants at the same time. The stations are initially filtered as active stations if they send at least one kind of pollutants with a large average value frequently and bring much negative influence towards the target station.

Based on the above definition, each station's active source station group can be detected. It affects the target station multiple times, transport at least one kind of massive pollutant, and the transport value maintains a high correlation with the target station pollution value. Then, with each target station as the center, it is possible to construct a pollution propagation network between it and the active source station group. This network can reveal many multivariate spatiotemporal laws and assist with air pollution control, which deserves further analysis. We model the diachronic propagation network $\urcorner$ of a target station as a sequence of T snapshots:

$$\urcorner = (G_1, G_2, \ldots, G_T) \quad (9)$$

where a snapshot is a directed graph $G_i = (V, E_i, t_i)$, with node set V and edge set E_i, and t_i represents the $i-th$ timestep. The tuples in V are fixed, which contains all the active source stations and the target station. The number of tuples in E_i equals the number of all the active source stations $|V|$. As the weight of an edge e linking the node m and the node n in snapshot i is defined as a 6-dimension vector, denoting the six kinds of pollutant transported values from the source station a to the target station b at the $i - th$ timestep.

C. PROPAGATION SUMMARY VIEW

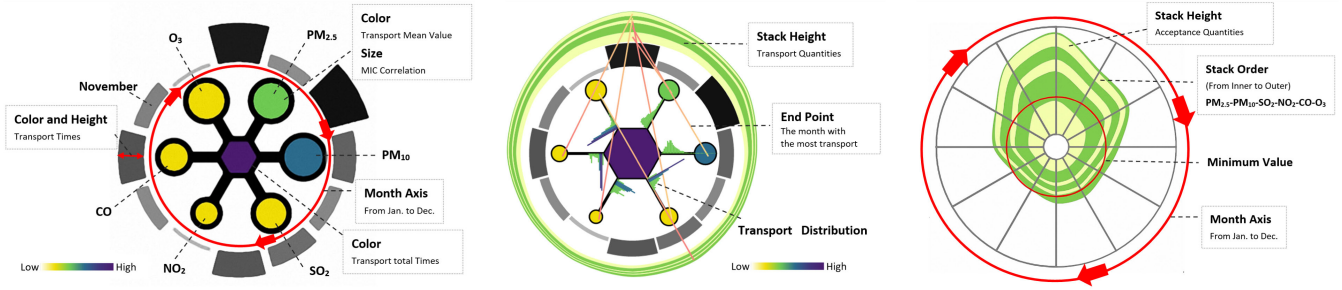
To enable visual exploration of air pollution propagation networks, we design the **propagation summary view** for providing a high-level overview of the suffering of the target station and the impact of each active source station.

When we focus on a target station centered propagation network, the source station group detected by the model contains a wealth of statistical information, including transport value and transport times of propagation in terms of pollutants and time, which provides the foundation for exploring pollution propagation mechanism. In this paper, we define every four hours as a time step and perform a transport simulating at each time step. In this way,the transport value of a pollutant at a time step determines the life value of the corresponding particle, and the transport times is equal to the number of time steps reaching the target station. While the transport times are different for the different periods, and the transport values of different pollutants are also different during the different periods. By referring to the literature on analyzing the time-varying law of air pollution propagation, we find that most studies were based on seasons or months, so we choose a month as the period granularity to analyze the time-varying law in the propagation summary view. After discussing with environmental experts, we summarize the key questions that need to be addressed by the view designs.

1) Q1. Which source stations have more frequent total transport times?
2) Q2. For a source station, how about the difference of the transport times for different months?
3) Q3. Which source stations have a large total transport value?
4) Q4. For a source station, how about the difference of the transport values for different pollutants?
5) Q5. For a source station, how about the difference of the transport values for different months?
6) Q6. For a source station, how about the difference of the transport values for different pollutants in different months?
7) Q7. For a source station, are the MIC correlation of different pollutants with the target station the same?

The information that needs to be displayed is complex. After multiple iterations of balancing the aesthetics and readability, we finally adopt a pair of tire-resembled glyphs based on multi-level design criteria, called **HighTireStation** and **LowTireStation**. The high-level **HighTireStation** is used to compare the transport of different source stations, while the low-level **LowTireStation** is used to show the pollution transport laws of a source station in detail.

As shown in FIGURE 3(a), HighTireStation consists of two parts, the outer tube, and the inner hub. The distribution of the temporal transport frequency is depicted with a histogram outside. Each bar represents a month, and the height and color represent transport times of the corresponding month(Q2). The center of the hub is a hexagon whose color encodes the total transport times(Q1). Purple represents a high value, and



(a) HighTireStation (b) LowTireStation (c) TargetStation

FIGURE 3. Designs of glyphs.

yellow represents a low value. The six axes that are sequentially arranged on the hub represent six kinds of pollutants. A circle is drawn at the end of each axis, whose color depicts the corresponding transport mean value(Q4), and its size represents the MIC correlation between the source station and target station(Q7). As for Q3, we find that the findings from Q3 and Q4 are consistent. The source station that has a high total transport value also has high pollutant transport values (the six circles have darker colors). To reduce redundant information, we abandoned coding Q3 by additional elements in the design.

Besides, users can also explore the source station of interest, such as the station that transports the most pollutants and the most frequently, or the noteworthy station in a particular season. After clicking one source station of interest, the detailed information will be added into the corresponding HighTireStation and brought to the foreground as the second level glyph LowTireStation. As shown in FIGURE 3(b), we incorporate the pollutant information by adding a radial stacking diagram outside to show the timing variations of six kinds of pollutants transport values in months(Q5 and Q6). Furthermore, we detect the month having the maximum value of each pollutant and use an orange line to connect it with the corresponding pollutant circle on the hub. Besides, the distribution of the transport values of various pollutants is laid out on the hub's six axes, facilitating more detailed analysis of each pollutant's transport laws.

As for the glyph used to identify the target station, we focus on the timing changes of its acceptance quantities of pollutants, which are shown in a radial stacking diagram. In order to enhance the comparison of the values in the radial layout, we add a baseline as the reference axis marked as a red circle in the stacking diagram, indicating the minimum value of the summed transport value of the six pollutants in twelve months. Through FIGURE 3(c), we can judge in which months the target station is more susceptible and which pollutant is the most influential. We put all the glyphs into a layout close to the real geographical distribution and form the propagation summary view. A force-directed collision detection method is implemented to separate the overlapping glyphs. This function can solve the visual clutter problem and keep the original layout as much as possible.To achieve this, we place glyphs as the scaled real geographic positions first. Then we calculate the maximum width or height of each glyph and use half of its length as the collision radius. We introduce the force-directed algorithm, with glyphs both drawn towards their original positions and by the spring force exerted by others. In this way, if a glyph does not collide with others, move it to the original position by a small step, and then iteratively move all glyphs until the layout reaches a balanced state. As a result, glyphs are laid out without clutter and maintaining geographic relative directions. Besides, we draw the same province's stations into a convex hull to facilitate the users to compare the real geographical locations. For example, FIGURE 4 demonstrates the transport network targeting Beijing. From the position of glyphs and the information they display, the users can explore various relationships between the pollution propagation and the geographic locations.

VI. AIR POLLUTION PROPAGATION PATTERNS

As mentioned above, we construct the air pollution propagation model to analyze how a selected target station is affected and identify its active source station group. The propagation summary view directly reveals the differences between the source stations from an overview perspective. However, how the source stations work together to influence the target at different periods is still unknown, which is of great significance for the joint management of pollution sources.

Therefore, based on the constructed propagation network, we extract the cooperative pollution propagation pattern between source stations. Furthermore, the extracted patterns are displayed and interpreted in the context of temporal, geographical, and multivariate information. This assists users in identifying which stations are synergistically affecting the target, at what time, and what are the major pollutants delivered.

A. PATTERN EXTRACTION

When exploring the pollution situation for the target station, consideration should be given to the air quality of the target station as well as the delivery values from the affecting station group. In the study of dynamic networks, analyzing the network structure by transforming a network into a high-dimensional vector based on the adjacency relationship

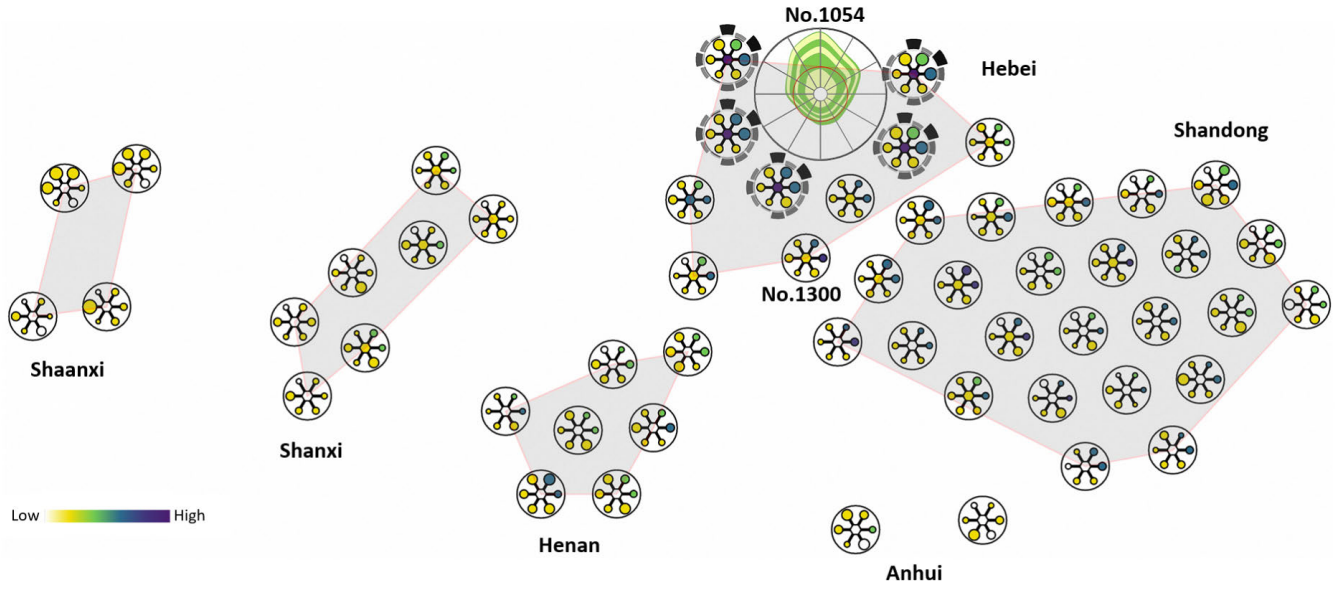


FIGURE 4. Propagation summary view. The air pollution propagation network of the station No.1054.

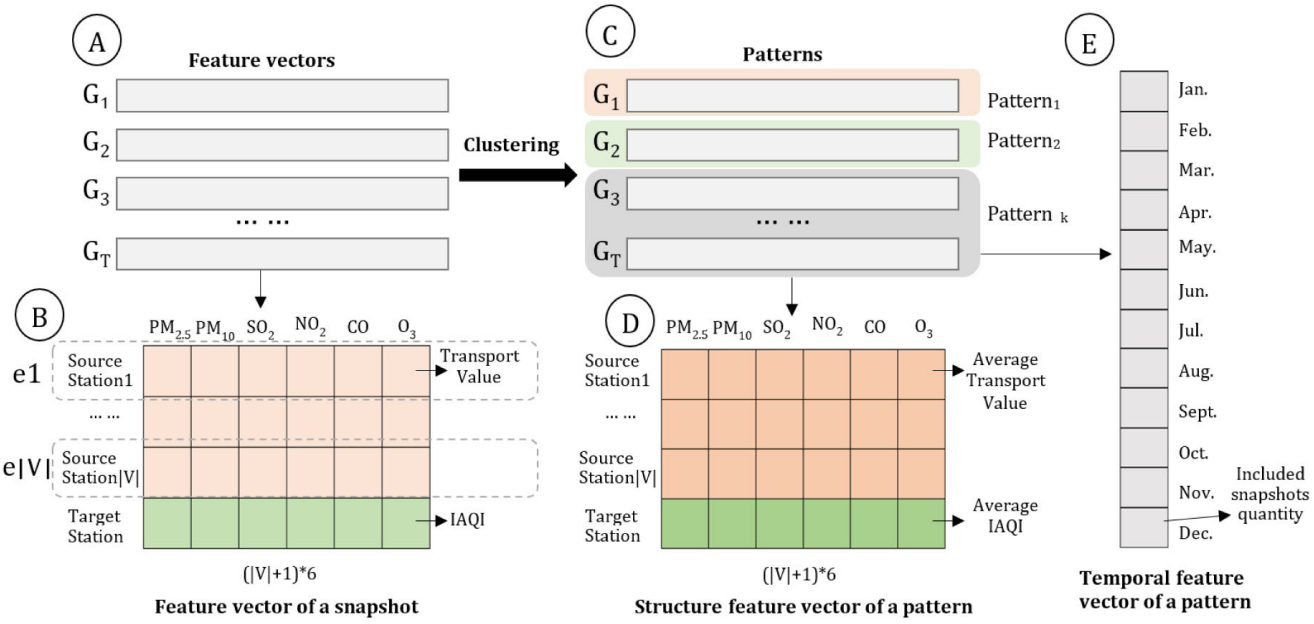


FIGURE 5. The framework of extracting patterns.

has been proved effective by some researches [33]. Inspired by this, we vectorize the diachronic propagation network of a target station and extract the patterns.

As shown in FIGURE 5(a), each snapshot G_i has a single feature vector, and each vector is expressed as a matrix with $|V+1|$ rows and 6 columns (FIGURE 5(b)). Each of the first $|V|$ rows corresponds to an edge, and each orange element in the first $|V|$ rows represents the transport value of corresponding pollutant transported by a source station to the target station at time step i. The last row of the matrix (green elements) shows the 6 pollutants IAQI values of the target station at time step i.

As a result, we have T feature vectors for T time step that define the diachronic propagation network. To further examine the similar snapshots, we perform clustering on the feature vectors and identify the extracted clusters as patterns. SF-kmedios algorithm [40] is employed to achieve this goal. It improves the traditional PAM algorithm [41] and requires the distance between every pair of items only once. This leads to high accuracy and less computation.

Another critical issue that needs to be solved in clustering is to determine the number of clusters. As the number of clusters increases, the items will be partitioned more refined, the degree of aggregation of each cluster will gradually increase, and the sum of the squared errors (SSE) will become smaller gradually. However, when the number of clusters is higher than the number of real clusters in data, SSE will tend to be stable. In this paper, we adopt the elbow method to find the inflection point of SSE and set the corresponding result as the number of clusters, thereby achieving a reasonable clustering performance.

In this way, we can automatically obtain the clustering results, and each cluster represents a kind of pattern of air pollution propagation. As shown in FIGURE 5(c), all snapshots are assigned into several patterns. Therefore, we can calculate the structure feature vector for a pattern by calculating the mean value of the original vectors included (FIGURE 5(d)). In addition, for the snapshots contained in a pattern, we mark the month that each snapshot is located in. Then we count the number of snapshots contained in each month, and create a 12-dimensional vector (FIGURE 5(e)) as the pattern's temporal feature. In the subsequent analysis, the structural feature vector of each pattern is displayed in the pattern comparison view(Section VI.C), and the temporal feature vector is displayed in the pattern summary view(Section VI.D), as well as they are used in the two sorting methods of the pattern(Section VI.C and Section VI.D).

B. PATTERN EVALUATION

In the propagation patterns obtained above, it is apparent that the patterns synchronized with the pollution of the target station are more worthy of attention. Towards this goal, we propose two evaluation indices, called pollution frequency (PF) and inverse pattern frequency (IPF), to assist users in locating meaningful patterns. The two indices are inspired by term frequency and inverse document frequency [42], which are the common concepts in natural language processing for detecting the keywords of an article. In this paper, PF and IPF are used to jointly evaluate the importance of each pattern for the polluted target stations.

Pollution frequency (PF): Probability of the target station polluted when this pattern occurs, which is given as:

$$PF = NtpP/Ntp \tag{10}$$

where Ntp denotes the number of time steps included in the focused pattern, and $NtpP$ represents the number of time steps included in the focused pattern when the target station is seriously polluted. In this paper, we define severe pollution as the case that the AQI value is over 150.

Inverse pattern frequency (IPF): Probability of this pattern occurred when the target station polluted, which is given as:

$$IPF = NtpP/Nt \tag{11}$$

where Nt represents how many time steps the target station gets polluted seriously within a specific period.

For a focused pattern, the higher PF, the more prone to the target station being polluted within the period it occurs. The higher IPF, the greater contributions it makes to the air pollution events. These two indices provide users with tools for effective air pollution propagation analysis. Moreover, we plot both of them on the same graph, as shown in FIGURE 6. The horizontal axis indicates PF, and the vertical axis indicates IPF. Patterns are marked as circles and classified into four categories according to their coordinates.

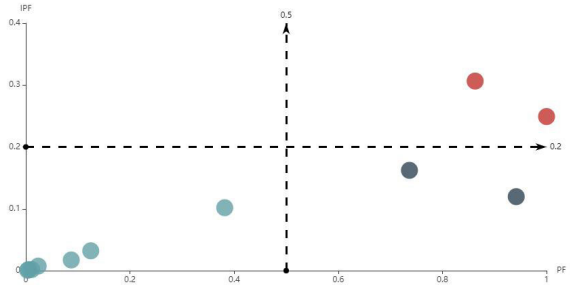


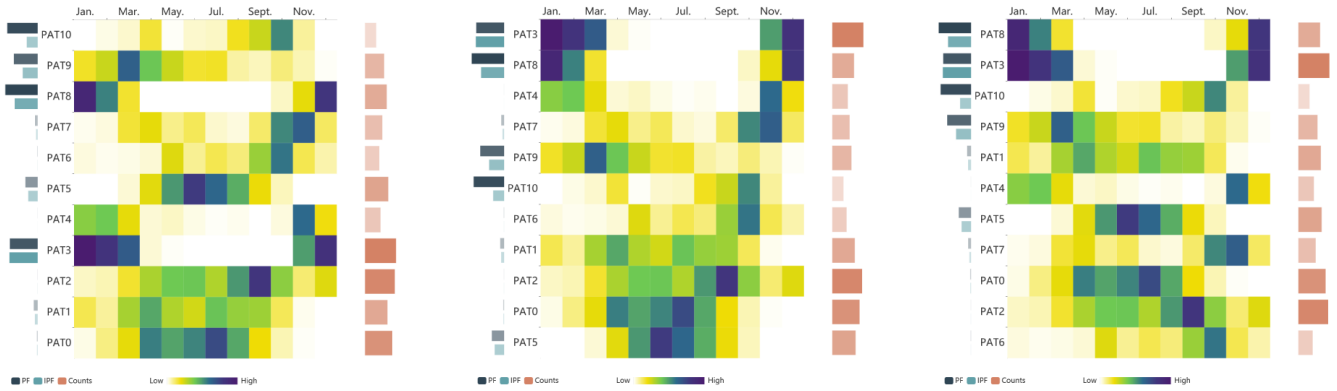
FIGURE 6. Pattern evaluation indices.

1) High PF and high IPF: They represent some long-term patterns of pollution propagation, which best meet our expectations and can be defined as the most noteworthy patterns.
2) High PF and low IPF: It means that even if the occurrence of these patterns signifies more likely excessive pollution of the target station, the frequency is very low. So, these patterns can reveal abnormal pollution propagation events.
3) Low PF and high IPF: As far as we know, this situation is contradictory and this model cannot exist.
4) Low PF and low IPF: When these patterns occur, the air quality at the target station is relatively good, which may not be the focus of joint pollution control.

C. PATTERN SUMMARY VIEW

In addition to the associations between the patterns and the pollution of the target station, other statistics information can also be used to evaluate patterns. We design the **pattern summary view** (FIGURE 7(a)) that provides the user with an overview of the pattern evaluation in terms of temporal distribution and statistics.

There is a matrix heatmap in the center of the pattern summary view, where each row represents one pattern, and each column represents one month. The color of a cell denotes the amount that the pattern appears in the month. There is a horizontal axis on the left of the heatmap showing the range of the evaluation indexes. We draw two bars for each pattern to indicate PF and IDF. Meanwhile, we add bars on the right of the heatmap to denote the number of timesteps included in the corresponding patterns. Under the three bars' mutual guidance, users can also quickly locate the pattern of interest, such as the pattern of serious pollution, for the next detailed analysis. From this view, users can perceive



(a) Unsorted. (b) Sorted by similar temporal distributions. (c) Sorted by similar network structure.

FIGURE 7. Pattern summary view.

whether the patterns that occur each month are diversiform or specific. Meanwhile, the temporal uncertainty of each pattern is revealed.

To better demonstrate this kind of uncertainty and reveal the difference of temporal distribution between the patterns, we reorder the rows in the heat map with specific rules. As shown in FIGURE 7 (b), patterns with similar temporal distributions are placed together. To get this result, we construct a feature vector for each pattern that stores the frequency of occurrences on each month. Euclidean distance is introduced to measure the similarity between the vectors, and then the vectors are projected into 1-dimension space by Multidimensional Scaling(MDS) [43]. In this way, we sort the rows of the heat map in 1D order. Different from the previously chaotic heat map, the temporal distribution laws can be observed more clearly.

D. PATTERN COMPARISON VIEW

We design the **pattern comparison view** to compare the average propagation network structure of each pattern. As shown in FIGURE 8, it consists of small multiples representing the patterns. In each multiple, the target station is marked as a black rhombus. At the same time, source stations are demonstrated by a glyph like a clock, in which the max mean transport value of the six pollutants is drawn as a hand and encoded by both length and color. Here, the node positions of the stations are geographically determined and shifted slightly according to the propagation summary view.

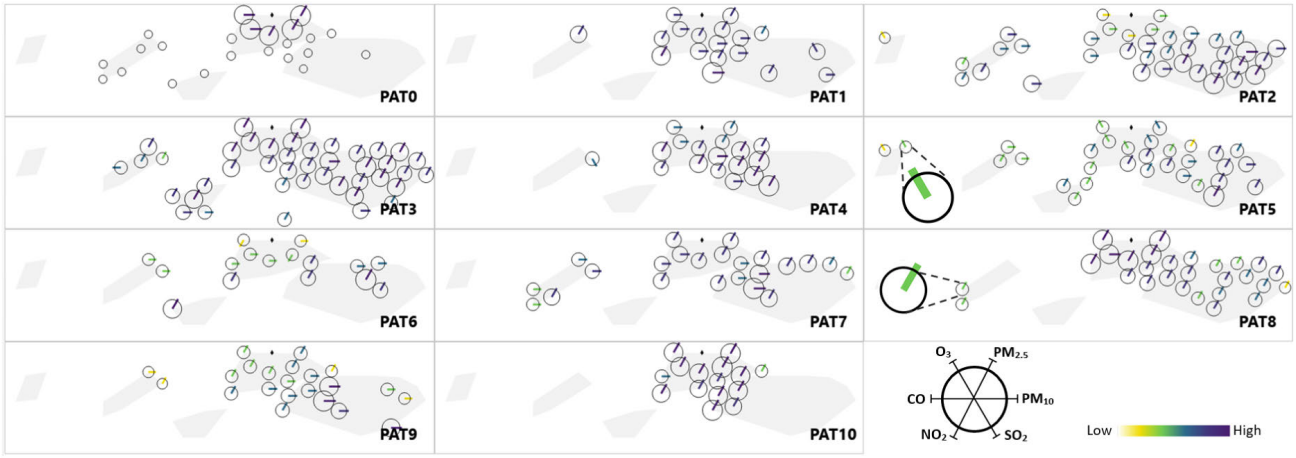
The position of pattern multiples on the screen can be set in three ways. First, the so-called sequential positioning lays the pattern multiples next to each other in one or several rows (FIGURE 8(a)). Second, we position the pattern multiples according to the similarity of their temporal distributions (FIGURE 8(b)), as mentioned in section VI.C. Another way to place patterns is based on the patterns' structural similarities (FIGURE 8(c)). The similarity function is the same as section VI.A. We place buttons on the control panel to switch the multiples positioning strategies flexibly. In addition, the third sorting method can also be transferred to the heatmap mode sorting (FIGURE 7(c)). Through this, users can explore the structural differences in the context of temporal information or explore the temporal differences in the context of structural information. For example, users can lock patterns with similar temporal distribution and compare their network structure with the varying composition (source stations, transport frequencies and transport values).

VII. CASE STUDIES

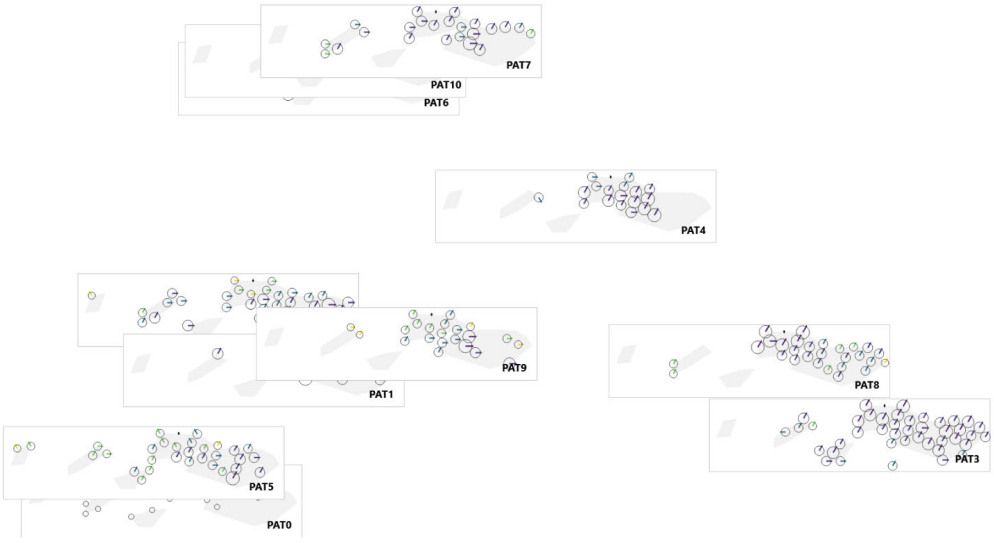
The preprocessing is written in C++, while the visualization is performed using D3.js and Echarts.js. We tested the proposed visual analytics system on a computer with a 4.00 GHz Intel Core i7-6700K CPU and 16 Gbytes of memory. FIGURE 9 shows the exploration pipeline of our system.

First, the user chooses what original information (air pollution or meteorological data on a specific time step) is demonstrated in the map view (FIGURE 9(a)) and then observes the map (FIGURE 9(d)(e)). Then the user selects a station by clicking the circle on the map (FIGURE 9(d)) or the word in the word cloud (FIGURE 9(c)). Next, the user explores the propagation information of all active source stations displayed in the propagation summary view (FIGURE 9(f)), to explore the multivariate time-varying laws in different regions. The user can click a HighTireStation glyph representing a source station and then obtains a LowTireStation showing more detailed information (FIGURE 9(g)). After that, the user performs pattern analysis, compares the patterns' temporal features through the patterns summary view (FIGURE 9(h)), and compares the patterns'structural features through the patterns comparison view (FIGURE 9(i)). The user clicks the switches on the option (FIGURE 9(b)), selects the sorting method of interest, and resets the order or position of the patterns in the two views (FIGURE 9(j)(k)(l)(m)).

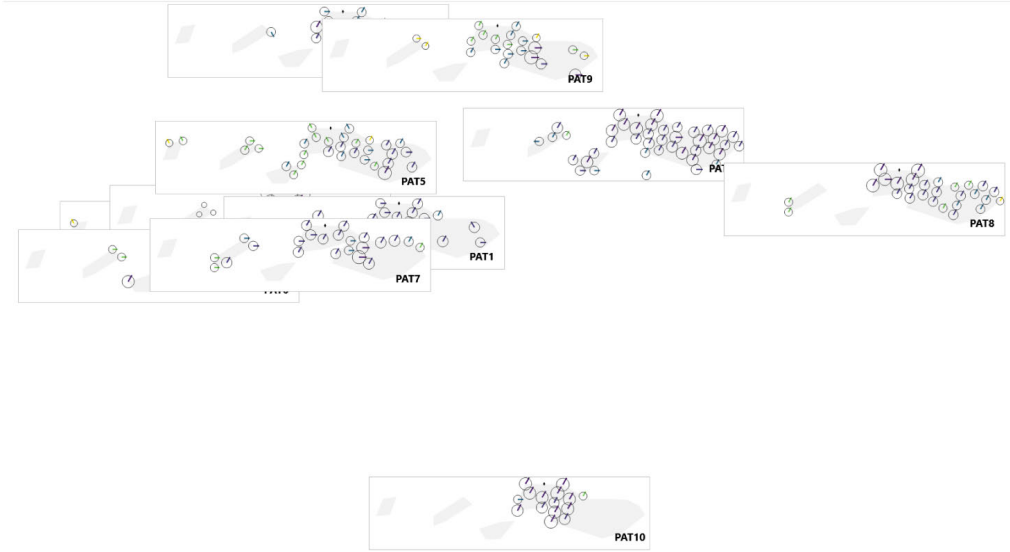
To demonstrate our system's effectiveness and usability, we conducte case studies with a domain expert who had studied the air-quality data for many years. The expert aims to identify the districts that act as the pollution sources in



(a) Unsorted.



(b) Projected by similar temporal distributions.



(c) Projected by similar network structural.

FIGURE 8. **Pattern comparison view.**

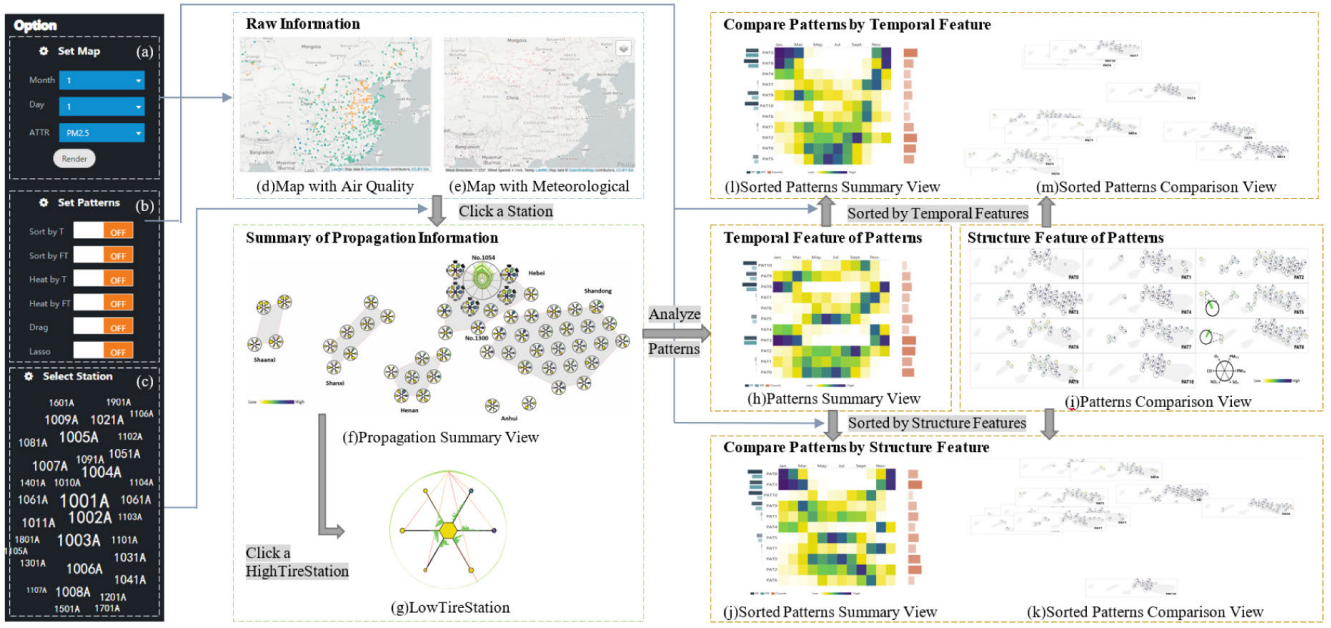


FIGURE 9. The exploration pipeline of our system.

air pollution propagation and understand how the districts interact with one another.

A. VISUAL ANALYTICS OF THE PROPAGATION NETWORK

Using our system, the domain expert starts with a brief overview (FIGURE 2) of China's air pollution, which shows the air quality monitoring stations' spatial distribution and acts as the interface of selecting stations for further research exploration. Each circle on the map represents a station, and the color represents the annual means of IAQI. The expert notices that North China stations are heavily polluted since most of the marks are orange or red. After zooming in on the map, he finds that the most polluted station No.1054 is in the Lianchi District of Baoding. He is greatly interested in it and hopes to recognize the stations that act as air pollution sources and understands how the source stations influence the target station No.1054 during different periods.

The expert clickes the mark of station No.1054, and then the map view and propagation summary view (FIGURE 4)are updated to focus the chosen station. The map view indicates that the major pollution comes from the surrounding areas and its southern areas. By referring to the propagation summary view, he immediately notices that the source stations are located in six provinces. The station No.1054 is closely related to the stations in Hebei Province where it is located. Multiple hexagons encoding the color-coded propagation frequencies have multiple purples, indicating long-term transports of pollutants to the target station. Shandong Province involves the largest number of source stations, although most stations only occasionally affect the target station, corresponding to the white or yellow hexagons in the figure. In addition, other areas, including Henan, Shanxi, Shaanxi, and Anhui, have also transport pollution.

The expert further examines the temporal and multivariate context information exhibited in the propagation summary view(FIGURE 7). He first focuses on the target station glyph and studies clear seasonal laws in the transport of air pollution. Station No. 1054 is affected by other areas more in winter, especially in December, but less affected in summer, especially in July. In addition, by examining each layer of the stacked graph's thickness, he realizes that $PM_{2.5}$ and PM_{10} are the main transport pollutants, and their transport values are consistent with the total value, keeping the transport pollution in winter more than summer. Surprisingly, O_3 is transported more often in summer than in winter, distinct from other pollutants, especially in May, June, and July. In order to assist in these findings, the expert continues to check the tires in the figure. He finds that the circles' color marking the mean values of $PM_{2.5}$ and PM_{10} transport are very dark and the PM_{10} value is slightly higher than $PM_{2.5}$. The expert then clicks on several high-frequency HighTireStation glyphs, and the LowTireStation glyphs are taken to the front desk. The shape of the gray lines in the glyphs are almost identical, and the pollutants are transported most in December except for the extreme value of O_3 in July. This complements the previous findings.

In addition, the expert also finds some phenomena that are inconsistent with common sense. The station No.1300 has the largest transport mean values of $PM_{2.5}$ and PM_{10}, but with a short geographical distance, as marked in FIGURE 4. By examining its low-level glyph(FIGURE 10), he finds that although it influences the target station fewer times but carries a lot of pollutants each time, especially in January. Another

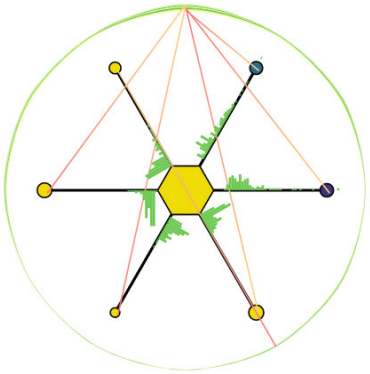


FIGURE 10. LowTireStation of the station No.1300.

finding is that there are some stations with small impacts, but they are highly relevant to the target station. As shown in the left half of the propagation summary view, their hexagons and pollutant circles are lighter in color, but the circle sizes of $PM_{2.5}$,CO, and O_3 are large. The expert realizes that these stations deserve further analysis in the future for finding similar pollution patterns.

B. VISUAL ANALYTICS OF THE PROPAGATION PATTERNS

After gaining the overall pollution situation of station No.1054, the expert hopes to explore further which regions jointly transport the pollution. So, using the system, he continues to explore the transport patterns to answer this question.

First, the expert sortes the patterns in the pattern summary view using the order of temporal distribution similarity(FIGURE 7(b)), and a series of apparent laws are revealed. At first glance, he is intrigued by the pattern PAT3 that occurs most frequently and with both high PF and IPF. All of this indicates that PAT3 is a common serious pollution mode that deserves special attention. Given that the heatmap demonstrates the temporal distribution, we can know PAT3 mainly occurs in January, February, and December, which are all in the winter. Then, he notices the pattern PAT8, which is next to PAT3 and also concentrated in winter with high PF and IPF. These findings serve as evidence that PAT3 and PAT8 can be defined as the typical pollution transport patterns in winter. Curious about which districts are involved in these patterns, the expert further inspects them in the pattern comparison view and maps the pattern tiles as the structure of transport networks(FIGURE 8(c)). In particular, all the patterns distribute unevenly and form several clusters. Even so, there is a close relation between PAT3 and PAT8. The expert unfolds these two patterns to study their transport networks. In the pattern tiles illustrated in FIGURE 8, most of the regions located in Hebei and Shandong transport pollutants to the target station at the same time, especially $PM_{2.5}$ and PM_{10}. From the source stations' colors, he finds that Hebei's stations are the main sources in PAT8. As for PAT3, although the degree of pollution transport is slightly lower, it involves more stations in Henan, and Zhengzhou is the main source. It is also worth mentioning that their pollution sources involve Shanxi province, but the stations are completely different. PAT3 involves some stations in the northeast, while PAT8 only involves stations in the southwest. Such an observation indicates that Shanxi stations do not transport pollution in the same direction in winter. The expert speculates that the different meteorological factors and terrain structures have led to this phenomenon.

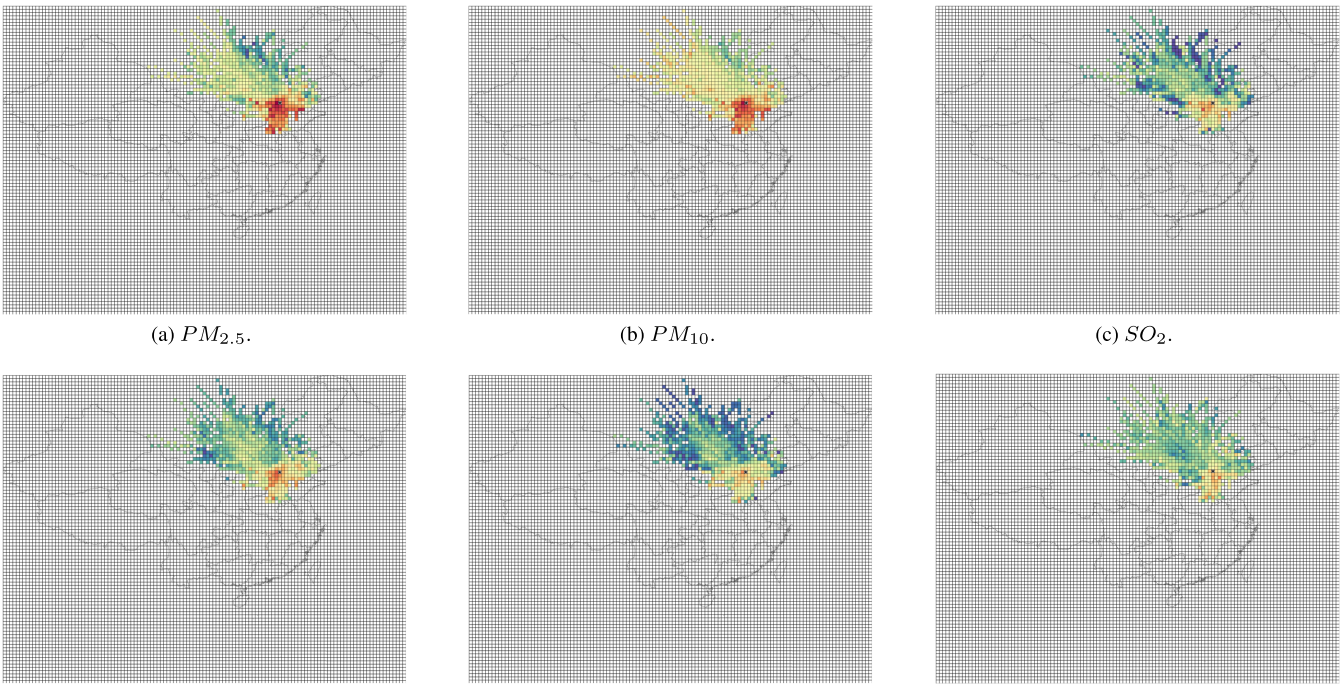
At the bottom of the pattern summary view, some patterns are completely different from the temporal distribution of PAT3 and PAT8, which mainly appear in the summer. Among them, PAT5 has relatively high PF and IPF values, which can be defined as the pattern with the pollution transport in summer. Meanwhile, PAT0 with low PF and IPF values appears more frequently in the summer, representing a better summer situation. The expert skims through these patterns and quickly finds that Shandong's southeastern stations are the main cause of the difference between PAT5 and PAT0, and the target station is polluted more heavily when they work in summer.

In addition, the expert also finds a strange phenomenon. There is an isolated pattern below the pattern view, PAT10, with the highest PF but a lower IPF and lower frequency, which can be regarded as an unusual serious pollution pattern. Such an observation triggers the expert's interest, so he unfolds the pattern PAT10. FIGURE 8 reveals that only Hebei Province and the northwestern Shandong region transport pollution to the station No.1054 at the same time, and this phenomenon mainly occurs in the autumn. The expert continues to observe the heatmap, and he is surprised to find that PAT2 with a lower indicator value is the most common pattern in autumn, indicating that the target station is not susceptible to contamination in the autumn.

VIII. USER STUDY

Our system received a lot of positive feedback. The expert has highly recognized the usefulness of glyph designs. He appreciated that glyphs have the ability to summarize the laws of air pollution propagation during a period. The first level of the glyph HighTireStation enabled him to distinguish differences among different areas at a glance. While, the second level LowTireStation can help the expert to explore the transport of pollutants in different months. These designs beyond the past traditional analysis work using existing software. The expert also appreciated the design of the pattern summary view, for revealing temporal distribution patterns of air pollution using a heatmap, which brings the convenience of subsequence analysis. Combining with pattern evaluation function, he could locate the special month pollution occurred in, and moved on further analysis on different patterns comparisons.

For problem-driven visual analytics system design, it is not feasible to use qualitative indicators to prove the system's effectiveness. Besides, different from our work, there is few visualization research focusing on summarizing the law of propagation and exploring patterns from a multivariate



(a) $PM_{2.5}$. (b) PM_{10}. (c) SO_2.

(d) NO_2. (e) CO. (f) O_3.

FIGURE 11. The Meteoinfo results of No.1054 station in 2015.

perspective. Subjecting to similar restrictions, most of the current works prove the effectiveness by case study and user study. Through the evaluation of experts who have rich domain knowledge and experience, they can verify if our system can help them in finding significant information about air pollution propagation and complete complex analysis tasks.

To prove our work's effectiveness, we design a task-driven qualitative user study on our system and a comparable baseline. We consulted the environmental experts and some students and learned that MeteoInfo [44] is a widely used software with some functions similar to our system. It is a widely used air quantity analysis software which well combines model HYSPLIT and concentration weighted trajectory(CWT) [45] method together. It can track the impact of surrounding areas on the air quality to the destination and the pollutant transport path over a period of time. However, this software could only analyze one pollutant at one time. Also, it could only provide trajectory map and contribution thermograph to users, lacking comprehensive information.

Using MeteoInfo, we choose the No. 1054 station as the target, same as the previous user study, and run the HYSPLIT model to simulate the propagation trajectory. On the basis of environmental experts' advice, we set the parameter initial height to 10m near the ground, and perform simulations every 8 hours, each time tracking the backward trajectory for 24 hours. After that, we select the entire outline of China as the map range, divide the map into a 0.5*0.5 uniform grid, and perform CWT analysis. Different from our work, the traditional CWT method only concerns the target's pollution value, but ignores the pass areas' air pollution during the propagation process. To make results comparable with our work, we define the contribution of a pass area to the target that equals the product of the residence time in the grid and the pollutant value in the grid. Same with the visualization available from MeteoInfo, we plot the results as grid heat maps for each pollutant (FIGURE 11).

So, based on what has been discussed above, we recruited ten graduate students as volunteers, including four environmental majors who used MeteoInfo software and six students majoring in computer science but have no expertise in air pollution. They were asked to complete four tasks in this study:

1) Task 1: Find the source station with the most pollutants transport.
2) Task 2: Identify the month with the most pollution.
3) Task 3: Summarize the different laws of geographical transport of different pollutants.
4) Task 4: Describe the transport laws in summer and winter respectively.

After a brief introduction to our framework, the participants were asked to complete the tasks using the heat maps and our system. After summarizing all the questionnaires, we obtained the following evaluations and suggestions from the participants:

The participants started to solve the tasks by heat maps. For Task 1, all participants circled dark grids of the six heat maps and quickly completed the task. For Task 3, most of them only used vague geographic areas to describe key contribution areas. One participant said ''Heat map is an effective method

when exploring one kind pollutant alone. However, when you find an active $PM_{2.5}$ source grid, it is hard to analyze other pollutants' transport contribution in this grid. In that case, it is scarcely possible to locate and compare the same area in six heatmaps simultaneously, especially when the color of the heat map is complex.'' Then participants were asked to complete Task 2 and Task 4 that need exploring time-varying information. For Task 2, we explored the Meteionfo software and found it cannot automatically calculate the total value of transport pollution and visualization by month, so this task failed. For Task 4, we performed additional CWT analysis for summer and winter, and then provided the heat map to the participants. However, this task requires 12 snapshots to be observed at the same time, which makes the dilemma of exploration Task 3 more serious.

By contrast, our system can help users solve these tasks effectively. For Task 1 and Task 2, all of the participants got definite answers using the summary view. They agreed that the source station's glyph and the target station's glyph design in this view were very easy to accept and learn. For Task 3, the six computer science students explored only using the high level of the glyph. They thought the high level of the glyph was enough for acquiring the analysis result of this task. For Task 4, all participants said that the coordinates after pattern projection were quite convenient for analysis according to the time distribution. However, two participants thought that the projection view could not help discover the difference between patterns. So, we plan to add more functions to help compare different patterns in future work. For the entire system, two environmental majors presented that the summary view and the pattern view lacked map background, which was not as convenient as MeteoInfo they were familiar with. While, on the other hand, the other two participants thought the convex hull and the relative position could already prompt geographic information. For further exploration of details, users can view related information in the associated map view.

Based on the above comparison, our system has apparent effectiveness in analyzing the multivariate time-varying geographical law of pollution propagation. Still, MeteoInfo is a mature analysis platform that can solve other analysis tasks. It has many powerful functions that our system cannot complete. Since our system aims to analyze a sub-problem of pollution propagation, we hope that the system can be used as an auxiliary tool that supports the exploration of the time-varying and multivariate patterns more conveniently.

IX. CONCLUSION AND FUTURE WORK

In this paper, we focus on the spatiotemporal multivariate features of air pollution propagation networks. From the global summary and local comparison perspectives, we have presented our visual analytics framework and system design. Our approach not only helps the quantification of pollution propagation but also supports the interactive visual exploration of the propagation patterns in a rich context. Through systematic views and intuitive glyph designs, users can recognize at first glance the geographical distribution and time-varying laws of the pollution transport, as well as identify the spatiotemporal differences of joint propagation patterns. We also described two case studies based on real-world datasets and user evaluations to demonstrate our approach's effectiveness.

In future work, we plan to explore and analyze air pollution propagation with geographic grids under different spatial scales. Besides, since uncertainty is inevitable in the process of model simulating, we also plan to take uncertainty into account and explore uncertainty propagation networks.

REFERENCES

[1] B. I. Escher, H. M. Stapleton, and E. L. Schymanski, ''Tracking complex mixtures of chemicals in our changing environment,'' *Science*, vol. 367, no. 6476, pp. 388–392, Jan. 2020.
[2] X. Li, L. Jin, and H. Kan, ''Air pollution: A global problem needs local fixes,'' *Nature*, vol. 570, no. 7762, pp. 437–439, Jun. 2019.
[3] R. Kumar, V.-H. Peuch, J. H. Crawford, and G. Brasseur, ''Five steps to improve air-quality forecasts,'' *Nature*, vol. 561, no. 7721, pp. 27–29, Sep. 2018.
[4] P. Li, R. Yan, S. Yu, S. Wang, W. Liu, and H. Bao, ''Reinstate regional transport of PM2.5 as a major cause of severe haze in beijing,'' *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 21, pp. E2739–E2740, May 2015.
[5] Z. He, M. Deng, J. Cai, Z. Xie, Q. Guan, and C. Yang, ''Mining spatiotemporal association patterns from complex geographic phenomena,'' *Int. J. Geograph. Inf. Sci.*, vol. 34, no. 6, pp. 1162–1187, Jun. 2020.
[6] M. Akbari, F. Samadzadegan, and R. Weibel, ''A generic regional spatio-temporal co-occurrence pattern mining model: A case study for air pollution,'' *J. Geograph. Syst.*, vol. 17, no. 3, pp. 249–274, Jul. 2015.
[7] J. Li, S. Chen, K. Zhang, G. Andrienko, and N. Andrienko, ''COPE: Interactive exploration of co-occurrence patterns in spatial time series,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2554–2567, Aug. 2019.
[8] D. Byun and K. L. Schere, ''Review of the governing equations, computational algorithms, and other components of the Models-3 community multiscale air quality (CMAQ) modeling system,'' *Appl. Mech. Rev.*, vol. 59, no. 2, p. 51, 2006.
[9] A. F. Stein, R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. Cohen, and F. Ngan, ''Noaa's hysplit atmospheric transport and dispersion modeling system,'' *Bull. Amer. Meteorolog. Soc.*, vol. 96, no. 2, pp. 2059–2077, Dec. 2016.
[10] M. Bahiraei and S. M. Hosseinalipour, ''Thermal dispersion model compared with euler-Lagrange approach in simulation of convective heat transfer for nanoparticle suspensions,'' *J. Dispersion Sci. Technol.*, vol. 34, no. 12, pp. 1778–1789, Dec. 2013.
[11] G. Zhao, G. Huang, H. He, and Q. Wang, ''Innovative spatial-temporal network modeling and analysis method of air quality,'' *IEEE Access*, vol. 7, pp. 26241–26254, 2019.
[12] B. Bach, E. Pietriga, and J.-D. Fekete, ''Visualizing dynamic networks with matrix cubes,'' in *Proc. 32nd Annu. ACM Conf. Hum. Factors Comput. Syst. CHI*, 2014, pp. 877–886.
[13] P.-M. Law, Y. Wu, and R. C. Basole, ''Segue: Overviewing evolution patterns of egocentric networks by interactive construction of spatial layouts,'' in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2018, pp. 72–83.
[14] D. Wang, P. Cui, and W. Zhu, ''Structural deep network embedding,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1225–1234, doi: 10.1145/2939672.2939753.
[15] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec, ''Learning structural node embeddings via diffusion wavelets,'' in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1320–1329, doi: 10.1145/3219819.3220025.
[16] A. Daly and P. Zannetti, ''Air pollution modeling—An overview,'' Ambient Air Pollut., Tech. Rep., 2007, pp. 15–28.
[17] H. McGowan and A. Clark, ''Identification of dust transport pathways from lake eyre, Australia using hysplit,'' *Atmos. Environ.*, vol. 42, no. 29, pp. 6915–6925, Sep. 2008.
[18] J. C. Carvalho and M. T. M. B. de Vilhena, ''Pollutant dispersion simulation for low wind speed condition by the ILS method,'' *Atmos. Environ.*, vol. 39, no. 34, pp. 6282–6288, Nov. 2005.

[19] K. Manomaiphiboon and A. G. Russell, "Effects of uncertainties in parameters of a lagrangian particle model on mean ground-level concentrations under stable conditions," *Atmos. Environ.*, vol. 38, no. 33, pp. 5529–5543, Oct. 2004.

[20] H. Zhang, K. Ren, Y. Lin, D. Qu, and Z. Li, "AirInsight: Visual exploration and interpretation of latent patterns and anomalies in air quality data," *Sustainability*, vol. 11, no. 10, p. 2944, May 2019.

[21] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao, and W. Su, "Visual analytics for spatial clusters of air-quality data," *IEEE Comput. Graph. Appl.*, vol. 37, no. 5, pp. 98–105, Sep. 2017.

[22] F. Guo, T. Gu, W. Chen, F. Wu, Q. Wang, L. Shi, and H. Qu, "Visual exploration of air quality data with a time-correlation-partitioning tree based on information theory," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 1, pp. 1–23, Mar. 2019.

[23] J. Li, Z. Xiao, H.-Q. Zhao, Z.-P. Meng, and K. Zhang, "Visual analytics of smogs in China," *J. Visualizat.*, vol. 19, no. 3, pp. 461–474, Aug. 2016.

[24] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo, "Visual analysis of the air pollution problem in hong kong," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1408–1415, Dec. 2007.

[25] Z. Deng, D. Weng, J. Chen, R. Liu, Z. Wang, J. Bao, Y. Zheng, and Y. Wu, "AirVis: Visual analytics of air pollution propagation," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 800–810, Jan. 2019.

[26] Y. Yang, T. Dwyer, S. Goodwin, and K. Marriott, "Many-to-many geographically-embedded flow visualisation: An evaluation," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 411–420, Jan. 2017.

[27] A. Nocaj, M. Ortmann, and U. Brandes, "Untangling the hairballs of multi-centered, small-world online social media networks," *J. Graph Algorithms Appl.*, vol. 40, no. 2, pp. 977–985, 2015.

[28] B. Bach, N. H. Riche, C. Hurter, K. Marriott, and T. Dwyer, "Towards unambiguous edge bundling: Investigating confluent drawings for network visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 541–550, Jan. 2017.

[29] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren, "MobilityGraphs: Visual analysis of mass mobility dynamics via Spatio-temporal graphs and clustering," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 11–20, Jan. 2016.

[30] S. V. D. Elzen, D. Holten, J. Blaas, and J. J. van Wijk, "Dynamic network visualization withExtended massive sequence views," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 8, pp. 1087–1099, Aug. 2014.

[31] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "A taxonomy and survey of dynamic graph visualization," *Comput. Graph. Forum*, vol. 36, no. 1, pp. 133–159, Jan. 2017.

[32] M. Greilich, M. Burch, and S. Diehl, "Visualizing the evolution of compound digraphs with TimeArcTrees," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 975–982, Jun. 2009.

[33] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk, "Reducing snapshots to points: A visual analytics approach to dynamic network exploration," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 1–10, Jan. 2016.

[34] J. Xu, Y. Tao, Y. Yan, and H. Lin, "Exploring evolution of dynamic networks via diachronic node embeddings," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 7, pp. 2387–2402, Jul. 2020.

[35] M. Hajij, B. Wang, C. Scheidegger, and P. Rosen, "Visual detection of structural changes in time-varying graphs using persistent homology," in *Proc. IEEE Pacific Visualizat. Symp. (PacificVis)*, Apr. 2018, pp. 125–134, doi: 10.1109/pacificvis.2018.00024.

[36] Q. Q. Ngo, M.-T. Hütt, and L. Linsen, "Visual analysis of governing topological structures in excitable network dynamics," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 301–310, Jun. 2016.

[37] D. Qu, X. Lin, K. Ren, Q. Liu, and H. Zhang, "AirExplorer: Visual exploration of air quality data based on time-series querying," *J. Visualizat.*, vol. 23, no. 6, pp. 1129–1145, Dec. 2020.

[38] J. Brioude, D. Arnold, A. Stohl, M. Cassiani, D. Morton, P. Seibert, W. Angevine, S. Evan, A. Dingwell, J. D. Fast, R. C. Easter, I. Pisso, J. Burkhart, and G. Wotawa, "The lagrangian particle dispersion model FLEXPART-WRF version 3.1," *Geosci. Model Develop.*, vol. 6, no. 6, pp. 1889–1904, Nov. 2013. [Online]. Available: https://gmd.copernicus.org/articles/6/1889/2013/

[39] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[40] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009.

[41] M. Van der Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Stat. Comput. Simul.*, vol. 73, no. 8, pp. 575–584, Aug. 2003.

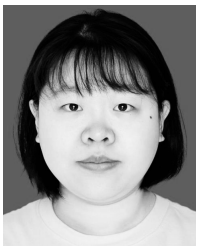
[42] M. Yamamoto and K. W. Church, "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus," *Comput. Linguistics*, vol. 27, no. 1, pp. 1–30, Mar. 2001.

[43] J. B. Kruskal, M. Wish, and E. M. Uslaner, *Multidimensional Scaling*. Norderstedt, Germany: Book On Demand Pod, 1978.

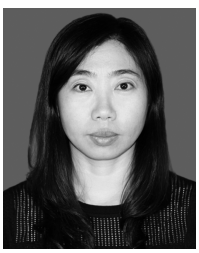
[44] Y. Q. Wang, "MeteoInfo: GIS software for meteorological data visualization and analysis," *Meteorolog. Appl.*, vol. 21, no. 2, pp. 360–368, Apr. 2014.

[45] N. Gao, M.-D. Cheng, and P. K. Hopke, "Potential source contribution function analysis and source apportionment of sulfur species measured at rubidoux, CA during the southern california air quality study, 1987," *Analytica Chim. Acta*, vol. 277, no. 2, pp. 369–380, May 1993.

KE REN received the B.Sc. degree from Northeast Normal University, in 2016, where she is currently pursuing the Ph.D. degree with the School of Information Science and Technology. Her current research interests include visual analytics, information visualization, scientific visualization, spatial-temporal data visualization, uncertainty visualization, and anomaly visualization.



YIYAO WU received the B.Sc. degree from East China Normal University, in 2017. She is currently pursuing the M.Sc. degree with the School of Information Science and Technology, Northeast Normal University. Her current research interests include visual analytics, scientific visualization, and interactive visualization with machine learning.



HUIJIE ZHANG (Member, IEEE) received the Ph.D. degree from Jilin University, in 2009. She is currently a Full Professor with the School of Information Science and Technology, Northeast Normal University. Her main research interests include scientific visualization, information visualization, visual analytics, computer graphics, 3D model simplification, multiresolution modeling for terrain, 3DGIS, and optimization algorithm. She is a Senior Member of the China Computer Federation (CCF) and the China Society of Image and Graphics (CSIG).

JIA FU received the B.Sc. degree from Northeast Normal University, in 2018, where she is currently pursuing the M.Sc. degree with the School of Information Science and Technology. Her current research interests include visual analytics, information visualization, spatial-temporal data visualization, multivariate data visualization, and anomaly visualization.

DEZHAN QU received the M.Sc. degree from Northeast Normal University, in 2017. He is currently an Assistant Librarian with the Library, Northeast Normal University. His current research interests include visual analytics and scientific visualization.

XIAOLI LIN received the M.Sc. degree from Northeast Normal University, in 2019. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Northeastern University. Her current research interests include visual analytics, information visualization, and scientific visualization.