A topology-based algorithm using corrective distance matrix for air pollution tracing

Ziao Liu[1], Lin Gao[1], Yuyang Hong[1] and Haibo Hu*

School of Software Engineering, Chongqing University, Chongqing, 410331, China.

ARTICLE INFO

Keywords:
Air pollution propagation
topology
graph visualization

ABSTRACT

The regional transmission of air pollutants is an important factor affecting climate and environment. Many existing studies are committed to mining potential propagation models. However, most models are highly dependent on data. Meanwhile, there is a lack of research on accurate pollutant tracing within and between regions. This paper presents a topology-based algorithm using correction distance matrix for air pollution tracing. The geographical distance is corrected according to meteorological factors. The distance relationship between monitoring stations is expressed as a directed graph. The distance threshold is used to filter the directed graph to obtain relevant subgraphs of the target station. By constructing pollutant concentration transfer equations, the reachability matrix of pollution propagation is calculated to get the pollution contribution of each node. The single point tracing result of the target station is obtained from the maximum contribution. Meanwhile, we design the air pollution tracing visualization system. The multi-dimensional views in the system are convenient for users to perceive propagation laws and explore results. Based on above work, we conduct an intraregional case study of Beijing in 2013 and an interregional case study of central and eastern China in 2014. The results of case studies confirm the effectiveness of the algorithm and visualization system in the field of air pollution propagation. Moreover, we evaluate the performance of the algorithm and get good feedback according to user's experience with MeteoInfo software and our system. Therefore, our work can accurately trace the source of air pollutants and provide technical support for meteorological experts.

1. Introduction

As one of the environmental problems, air pollution has attracted continuous attention all over the world. Relative research found that people's burden of diseases such as respiratory infection, heart disease, lung cancer may be led by air contamination(Jin et al., 2021). The Health Effects Institute report that the death rate of air pollution in 2019 has exceeded that of traffic accidents. Morever, Parajuli et al. (2021) found that a single pollutant has more frequent and significant effects on human body. Vo et al. (2022) found that 80 % of PM2.5 will be deposited in the head airway, while 75 % of PM0.1 will be deposited in the alveolar area. In addition, environment pollution has a negative impact on the economic situation, resulting in the total loss rate of India reaching 1.36% of GDP in 2019 (Pandey et al., 2021). There has always been an interdependent relationship between urban sustainable development and air pollution(Ahmad et al., 2021). According to demographic monitoring, excessive SO_2 in the air leads to the reduction of the overall growth and development of plants(Cotrozzi, 2020). In order to effectively reduce the harm of air pollution to these aspects, accurately determining the source of pollutants has become an essential part.

There has been a lot of research on the propagation of air pollution. The main research direction is to build an ideal propagation model and mine the symbiotic relationship under association rules. HYSPLIT4 (Stein et al., 2015) is a system that can not only calculate simple single trajectory problems but also solve complex distribution and precipitation problems through the smoke and particle method. Although HYSPLIT4 is most widely used in air pollution source identification, it ignores the problem of small-scale air transmission. Considering the influence of wind speed and direction on pollution propagation, some articles use Lagrange dispersion model (LDM) (Carvalho and De Vilhena, 2005) to avoid the numerpical pseudo diffusion. On the other hand, some researches focus on mining hidden symbiotic relationships through association rules to infer the relationship between pollution transmission. For example, the visual exploration framework Co-Occurrence Pattern Exploration (COPE) (Li et al., 2018) looks for co-occurrence patterns based on time-space series to evaluate the intensity of occurrence patterns and make predictions. Fan et al. (2020) made a comprehensive analysis of the spatio-temporal variation of urban air pollution in China during 2014–2018. Guo et al. (2019) proposed Time Correlation Partition (TCP) tree, which provides situational awareness analysis between correlation data and variables. Huang et al. (2018) adopted CAMx adjoint propagation model to realize dynamic inversion and grid quantitative propagation of emission sources. In this way, the spatio-temporal characteristics of pollution sources and their contribution rates can

*This work was supported in part by the National Natural Science Foundation of China under Grant U1836114.

*Corresponding author. Haibo Hu PhD, Associate Professor, School of Software Engineering, Chongqing University, Room 503, Software Building, No.55 South Road of University Town, Shapingba District, Chongqing, 410331 CN.

haibo.hu@cqu.edu.cn (H. Hu)

ORCID(s): 0000-0003-3237-0690 (Z. Liu); 0000-0001-9115-730X (L. Gao); 0000-0003-3901-7769 (Y. Hong); 0000-0001-8442-5222 (H. Hu)

[1]Authors contributed equally.

be evaluated quantitatively. The point-to-point propagation of pollutants can be mapped to the directed graph structure. The knowledge of graph theory and topological geometry can be feasible to apply to the pollution propagation model. Reachability matrix is often used to describe the degree that can be reached between nodes of a directed graph after a certain length of path. Topology-based reachability analysis plays an important role in exploring the relationship between two points of a directed graph. Xu and Hong (2012) proposed a finite solution method based on reachability matrix. The sufficient and necessary conditions for the reachability of deterministic and nondeterministic automata are obtained by using the semi-tensor product of matrices. Scarpato et al. (2019) proposed the Reachability Matrix Ontology (RMO) to describe the network structure and network security. RMO proposes an innovative method to investigate whether one node can reach another node through the ISO/OSI layer protocol. Kim et al. (2021) proposed a new incremental algorithm framework AORM, which is a reachability matrix framework of any order. AORM can handle both directed and disconnected networks, such as reference networks as well as various types of real data sets.

According to the feedback of meteorological experts, we find that there is a lack of effective visualization system to display data and analyze tracking results. Visual presentation relies on a lot of data analysis technology. Regression analysis (Demuzere et al., 2008), statistical analysis (Sharovsky et al., 2004) and correlation analysis (Qu et al., 2007) are often used as models for analyzing air quality data. The establishment of visualization system is helpful to discover the potential information and explore the value behind the data. There are a lot of researches focused on geographic information due to the close relationship between air pollution data and geographic information. By introducing the concept of air wrapping to draw the pollutant propagation map,Deng et al. (2019) developed the AirVis system, which is based on Word2Vec model to effectively extract transmission models. Ren et al. (2020) abstracted the complex inter regional propagation relationship into a dynamic network. In their research, visual analysis is introduced to explore the spatio-temporal multiple propagation model. At the same time, a model based on particle tracking is proposed to construct the pollution propagation network under multi-source factors. Deng et al. (2021) proposed VisCas system, which combines the extraction of pollution patterns with interactive visualization technology. Based on the triple optimization strategy, a novel influence view is designed to infer and explain the potential cascading patterns in the spatio-temporal context. Tian et al. (2021) developed an inline visualization system from the perspective of fine-grained air quality. Liu et al. (2021) proposed an integrated AQEyes system. By integrating multiple pipelines such as data preprocessor and characterizer, the unsupervised anomaly detection method is adopted to effectively check the anomalies. Qu et al. (2020) constructed the AirExplorer system. Combined with hierarchical piecewise linear representation and dynamic time warping, the time series patterns of interest are accurately queried. Zhou et al. (2017) used Multidimensional Scaling (MDS) to convert data into 2D maps. Combining hierarchical clustering, Voronoi diagram and interactive technology help experts explore and extract the propagation modes.

Inspired by above work, we propose a topology-based algorithm using correction distance for air pollution tracing. In order to maximize the data value, we preprocess the high-dimensional time series air pollution data. The reachability matrix between pollution monitoring stations is calculated by using classical topological geometry. According to the set thresholds, the correlation between monitoring stations is analyzed. Based on tracing results, we build the air pollution tracing visualization system. Our focus is on air quality monitoring stations. The goal of the system is to facilitate experts to explore the propagation law clearly and put forward urban air quality management plans.

In the remainder of this article, we describe the preprocessing work of the data set in Section 2. We present the details of the tracing algorithm in Section 3. In addition, we describe the visualization work in Section 4. We conduct two case studies with and between regions in Section 5. We evaluate the performance of algorithm and conduct user study in Section 6. We finally present the summary and outlook of this work in Section 7.

2. Data preprocessing

2.1. Data set

The data set used in this paper is high-resolution air pollution reanalysis data set of China released by the Institute of Atmospheric Physics, Chinese Academy of Sciences and other units. The data set includes the grid data of six conventional pollutants in China's ambient air quality standards. It is also the reanalysis data of the national ambient air quality monitoring network and nested grid air quality prediction model (naqpms) of China National Environmental Monitoring Station (CNEMC)(Kong et al., 2021). https://doi.org/10.11922/sciencedb.00053

The data set involves 2751 monitoring stations, covering 375 cities in China, with a time range from 2013 to 2018. Stations have collected daily mean data and hourly data in January from 2013 to 2018. Each data information includes the concentration of six air pollutants: $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 and CO_2. In addition, each station also recorded the hourly or daily wind direction, wind speed, temperature, humidity and air pressure information, as well as the longitude and latitude data of the station. The basic information is in Table 1:

2.2. Inverse geocoding

The position variables in the data set are expressed in longitude and latitude. However, We hope to obtain provinces, cities, districts and counties. This needs to obtain the geographic information of each city and carry out inverse geographic coding according to the longitude and latitude. In this paper, our local geographic information database is used to realize the inverse geocoding. We code the longitude and

Table 1
Data set variables

Data type	Name	Unit	Explain
Pollutant	$PM_{2.5}$	$\mu g/m^3$	Diameter $\leq 2.5\mu$m
	PM_{10}	$\mu g/m^3$	Diameter $\leq 10\mu$m
	NO_2	$\mu g/m^3$	Nitrogen dioxide
	SO_2	$\mu g/m^3$	Sulphur dioxide
	O_3	$\mu g/m^3$	Ozone
	CO	$\mu g/m^3$	Carbon monoxide
Meteorology	U	m/s	Latitudinal wind speed
	V	m/s	Radial wind speed
	RH		Relative humidity
	TEMP	K	
	PSFC	Pa	Ground pressure

latitude of cities above the county level in China by geohash. The steps are as follows:

- Convert the longitude and latitude information into geohash and query the candidate cities;
- Find the nearest city according to the distance.

The method converts the initial latitude and longitude data into city names that can be understood and used, such as 'Jiangsu-Suzhou-Wuzhong'.

2.3. Aggregation and reduction

We use the air quality index (AQI) to simplify the monitored pollutant concentration into a single conceptual index value. It classifies the degree of air pollution and air quality status, which is suitable to represent the short-term air quality status and change trend of cities. Individual air quality index (IAQI) refers to the air quality index calculated based on single pollutants.

3. Methodology

The topology-based algorithm using correction distance matrix for air pollution tracing is introduced in this part. Flow chart of algorithm steps is shown in Fig. 1.

The pseudo code of algorithm implementation is shown in Algorithm 1:

Algorithm 1 The topology-based algorithm for air pollution tracing

Input: $data$: air pollution data set; Dth: distance threshold; Pth: pollution threshold; i: target station number; t: initial date; r: sliding time window unit; $times$: sliding times
Output: k: source station number

```
for station i do:
    move the sliding window unit r;
    while times > 0 do:
        load data;
        compute humidity diffusion coefficient H;
        compute wind direction and wind speed diffusion coefficient W;
        compute real spherical distance distance;
        use comprehensive evaluation mode to get the modified distance matrix D;
        if D_ij > Dth then:
            compute reachability matrix A
            use logarithm and diagonal differentiation to get A'';
            compute the maximum value A''_ik;
            if A''_ik > Pth and i ≠ k then:
                station k is the source of pollution at time t;
            end if
        end if
    end while
end for
```

3.1. Correction distance matrix calculation

According to the basic theory of multivariate diffusion, air humidity mainly affects the molecular diffusion coefficient. The binary Chapman-Enskog humidity diffusion coefficient H (Chapman and Cowling, 1990) is calculated from the relative humidity. In the Eq. 1, P represents the total pressure of the gas; T is the temperature of the gas; M_A and M_B are molar masses and $\sum vol_A$ and $\sum vol_B$ are molecular diffusion volumes.

$$H = \frac{0.0101T^{1.75}\sqrt{\frac{1}{M_A}+\frac{1}{M_B}}}{P[(\sum vol_A)^{1/3}+(\sum vol_B)^{1/3}]} \tag{1}$$

According to the research of Li et al. (2017) and Li et al. (2018), wind direction and wind speed diffusion coefficient W is calculated in Eq. 3.1. Where d_A and d_B represent the azimuth represented by the wind direction at grid cells A and B respectively. d_M represents the azimuth of AB direction. F is the influence of wind direction on diffusion. The diffusion distance of downwind direction is small and that of upwind direction is large. The weights w_A and w_B are the influence of wind speed on diffusion. For the downwind direction, the wind speed and diffusion distance are negatively correlated. For the upwind direction, the wind speed and diffusion distance are positively correlated. L_{AB} represents the type of adjacency relationship between A and B. If it is edge adjacency, $L_{AB} = 1$. If it is point adjacency, $L_{AB} = 2$. v_A and v_B represent the wind speed. $sgn(x)$ represents that the higher the wind speed is, the greater the influence on the diffusion distance becomes, and vice versa.

(2)

The real spherical distance $dist$ between monitoring stations can be calculated by Gaussian coordinate conversion of longitude and latitude in the data set. The humidity diffusion coefficient H is multiplied by the real spherical distance to obtain the corrected humidity distance. Multiply the coefficient W by the real spherical distance to obtain

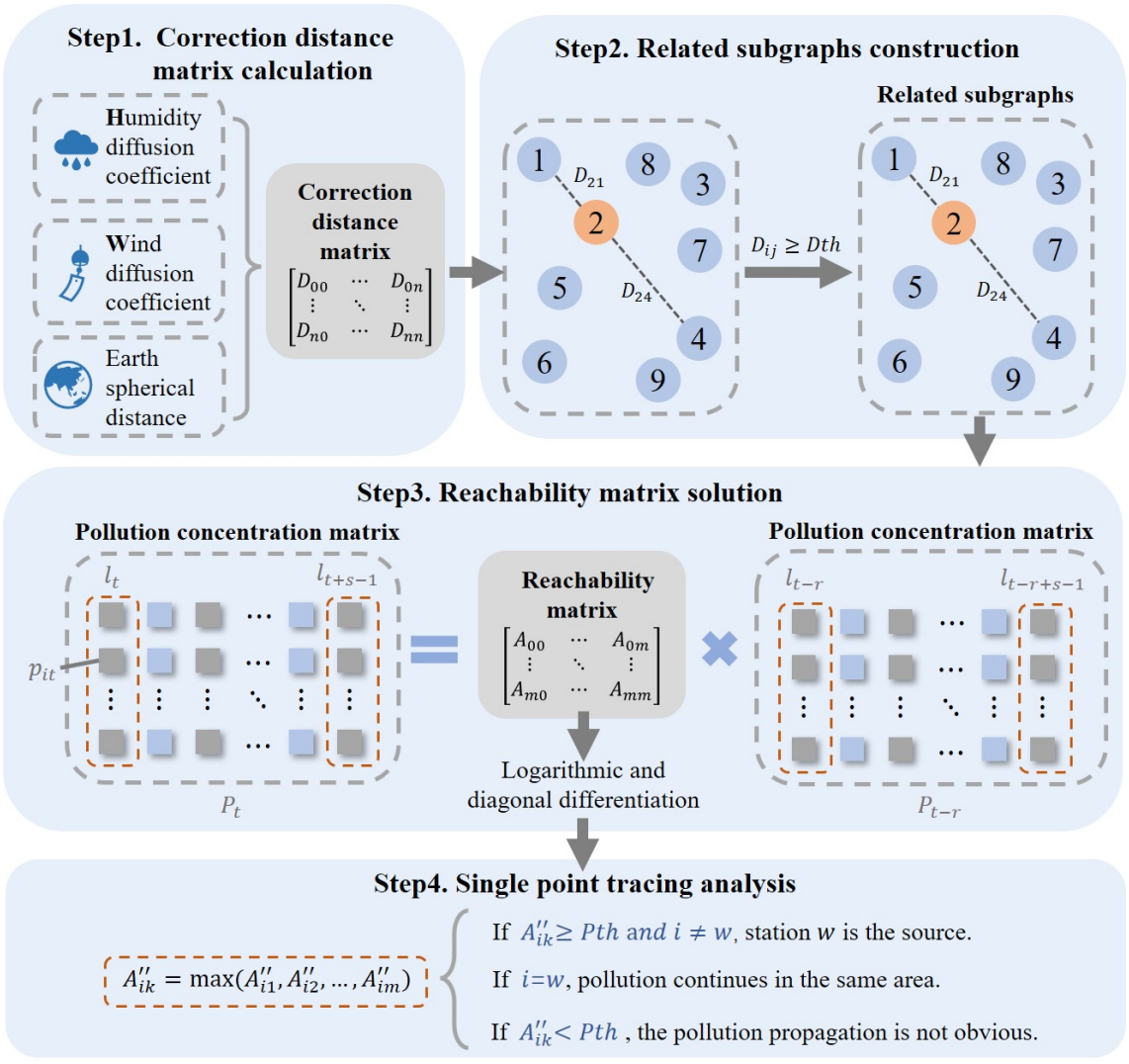

Fig. 1: Flow chart of algorithm steps

the corrected wind distance. The comprehensive evaluation model is used to give weight to the above three distances and the modified distance matrix D is obtained in Eq. 3.

$$D = w_1 \times dist + w_2 \times W \times dist + w_3 \times H \times dist \quad (3)$$

3.2. Related subgraphs construction

The distance matrix D between monitoring stations is regarded as a directed graph, where D_{ij} represents the influence distance of station j on station i. We suppose the ranges of i and j are positive integers from 0 to n. Determine the target station and filter the D_{ij} using the distance threshold Dth. Finally, the relevant subgraph of the target station are obtained. We assume that the relevant subgraph contains m stations at all.

3.3. Reachability matrix solution

The reachability matrix describes the degree that can be achieved after a certain length of path between the nodes of the directed connection graph. In this model, it is used to represent the pollution contribution between monitoring stations in the relevant subgraph. A_{ij} represents the pollution contribution of station j to station i in the relevant subgraph.

$$A = \begin{bmatrix} A_{00} & A_{01} & \cdots & A_{0m} \\ A_{10} & A_{11} & \cdots & A_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m0} & A_{m1} & \cdots & A_{mm} \end{bmatrix} \quad (4)$$

The solution of the reachability matrix depends on the pollutant concentration transfer equation and the solution process is as follows. In Eq. 5, P_t is the pollution concentration matrix of the monitoring station in the relevant subgraph. r is the sliding time window unit.

$$P_t = A \times P_{t-r} \quad (5)$$

According to the relevant subgraph, the pollutant concentration matrix of the monitoring station in the time interval is constructed in Eq. 6. s represents the time interval,

l_t is the concentration column vector of each station in the correlation subgraph at time t. p_{it} represents the pollution concentration of station i at time t.

$$\begin{aligned} P_t &= [l_t, l_{t+1}, \ldots, l_{t+s-1}] \\ P_{t-r} &= [l_{t-r}, l_{t-r+1}, \ldots, l_{t-r+s-1}] \\ l_t &= [p_{1t}, p_{2t}, \ldots, p_{nt}]^T \end{aligned} \tag{6}$$

We use logarithm and diagonal differentiation to preprocess the reachability matrix for subsequent tracing analysis:

1) Log the reachability matrix A to the matrix A'.

$$A'_{ij} = \begin{cases} \ln(A_{ij} + 1) & A_{ij} \geq 0 \\ -\ln(|A_{ij}| + 1) & A_{ij} < 0 \end{cases} \tag{7}$$

2) diagonally differentiate matrix A' to the matrix A''.

$$A''_{ij} = \begin{cases} A'_{ij} & i = j \\ A'_{ij} - A'_{ji} & i \neq j \end{cases} \tag{8}$$

3.4. Single point tracing analysis

Eq. 9 shows the process of extracting the station with the highest absolute value of contribution to the target station as the tracing result to achieve single point tracing. Take the corresponding row or column of the target station i in the reachability matrix as a set and calculate the maximum value A''_{ik}.

$$A''_{ik} = \max(A''_{i1}, A'_{i2}, ..., A''_{im}) \tag{9}$$

Set the pollution threshold Pth. If $A''_{ik} \geq Pth$ and $i \neq k$, then k is the tracing result of station i and A''_{ik} is the impact intensity of pollution source. If $A''_{ik} < Pth$, it is considered that the pollution transmission is not obvious in this time period. That is, the tracing result of station i is not obtained. If $i = k$, the pollution in the area where station i is located continues.

Use time units r as a sliding window. Move the sliding window and repeat the above steps to draw the pollutant propagation path. While the tracing results are obtained, the influence degree of each transmission process can also be obtained.

4. Visualization of our work

Although the algorithm has been able to give the tracing results in the third part, it can't intuitively tell users more detailed information about the source place and the target place. Therefore, we design a visual analysis system, so that users can more intuitively see the tracing results and detailed information of tracing locations. The exploration pipeline of air pollution tracing visualization system is shown in Fig. 2.

4.1. Option window

The leftmost part of the system is the option window, which is composed of views A, B and C. The user selects the date through view A (the time range is from January 1, 2013 to December 31, 2018). View B uses the method of transforming word cloud to enable users to select the research area. In this view, the user can determine whether the area is selected according to the bold of the font. At the same time, users can select multiple areas. View C inputs control variables for tracing results. Users can generate tracing results by setting distance threshold, pollution threshold, sliding window size, sliding step and sliding times.

4.2. Basic information window

The basic information window are determined by views A and B. View D shows the AQI thermodynamic map of the selected area over a specific time range. The six colors in View D (varying from pale white to dark blue) correspond to each of the six levels of pollution. View E shows the change of climatic conditions during this period, specifically wind direction and wind speed. The color from white to converge with blue shows wind speed decrease. The color from white to converge with red represents wind speed increase.

4.3. Tracing result window

View F is the pollution propagation path map. After the user determines the parameters such as time, tracing range and pollution threshold, the tracing results of the algorithm can be visually displayed by the figure. The line between the two places indicates the transmission route of pollutants and the direction of the arrow indicates the transmission direction of pollutants. We also draw pie charts at the locations of each monitoring station. And the three components of each pie chart are W_1, W_2, and W_3 respectively, which form a one-to-one correspondence with the distance weight distribution chart (view I).

4.4. Multi-view analysis window

4.4.1. Trend comparison chart

View G is called the trend comparison chart. This view is designed to show the AQI trend comparisons between the source city and the target city. Among them, the red column is the AQI value of the source city and the blue column is the AQI value of the target city. Through the AQI value display for four consecutive days, we can see the trend change of air pollutants in the source and target points. On the right axis is the range of the AQI value. Through a lot of tests, we find that setting the threshold to 0-1000 can cover all situations. The lowest coordinate axis is the number of days displayed.

4.4.2. Influence degree sankey chart

View H is called the influence degree sankey chart. The tracing results of polluted areas reflects the relationship between data flow. This view is convenient for meteorological experts to observe the relationship between areas, to clarify the trend of pollutants between areas. The left vertexes are the source areas. The right vertexes are the target areas. The edges of the view constitute the pollution relationship. The

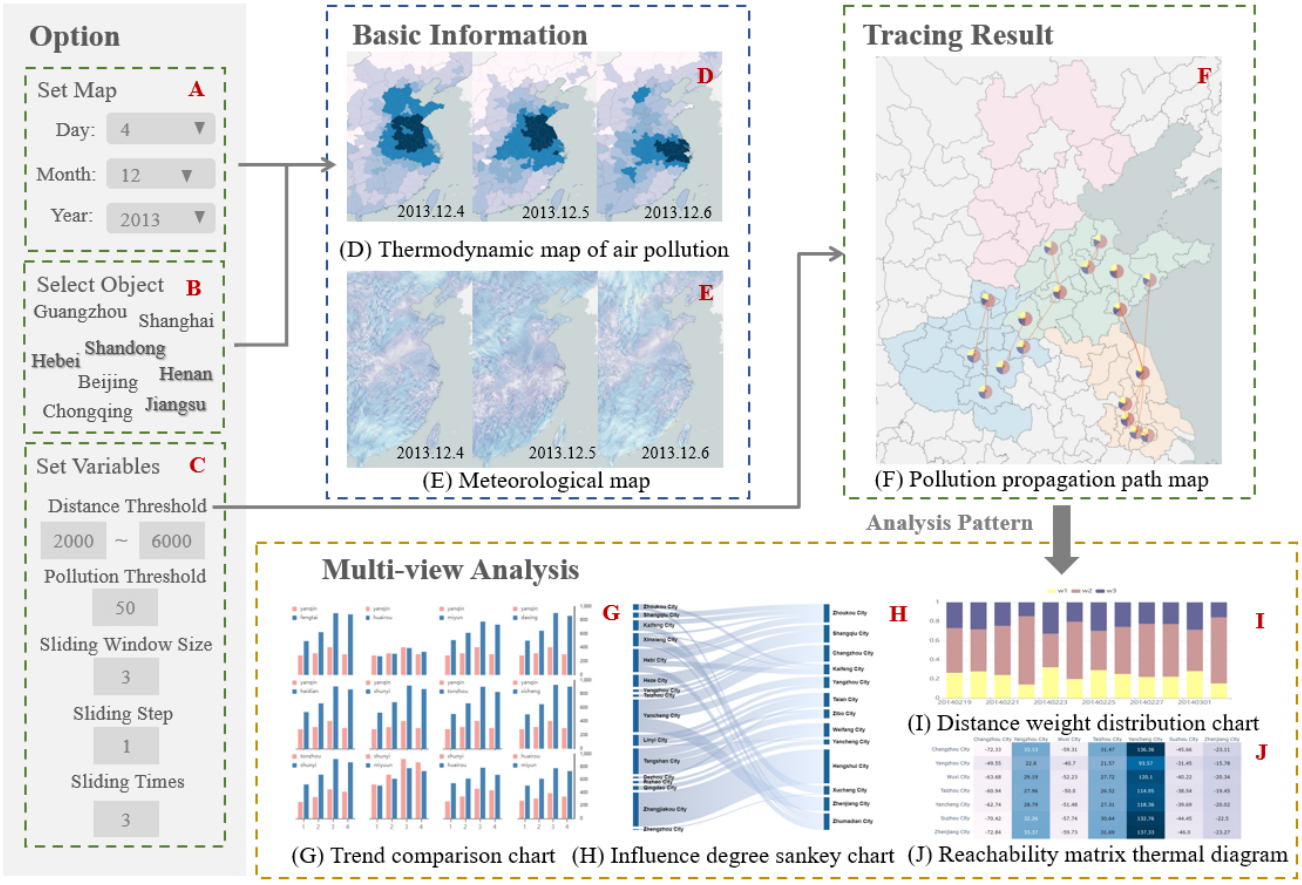

Fig. 2: The exploration pipeline guides users to employ the air pollution tracing visualization system. After setting variables in option window, users can explore the tracing results from algorithm with the help of basic information window. The multi-view analysis window helps users make some analysis work better.

flow is the degree of influence. The greater the degree of influence, the greater the flow. We can usually find that multiple areas will cause varying degrees of pollution to a area. At the same time, one area will also cause varying degrees of pollution to different areas. Therefore, the area with the largest flow value will become the focus of researchers.

4.4.3. Distance weight distribution chart

View I is called distance weight distribution chart. From the introduction of the algorithm, we can know that in order to consider meteorological factors, we modify the original distance. A comprehensive evaluation model is adopted to assign weights to each correction distance. We all visualize these three weights.w_1 represents the weight of the true distance of the sphere. w_2 is the weight of humidity correction distance. w_3 is the weight of wind direction and wind speed correction distance. Given the fact that the sum of the three weights always adds up to 1, we use stacked histograms. The abscissa is the time point and the ordinate is the weight value. By comparing the distribution of weights at different time points, it is found that the degree of meteorological impact changes. At the same time, the maximum weight represents the dominant factor.

4.4.4. Reachability matrix thermal diagram

View J is called the reachability matrix thermal diagram. From the introduction of the algorithm, we can know that the reachability matrix represents the pollution contribution between monitoring stations in the relevant subgraph. We use visual methods to facilitate meteorologists to find more relevant information about urban pollution. The matrix items are represented by squares. The color of the box indicates the size of the contribution. The greater the contribution is, the darker the color is. The contribution is positive or negative. Users can query the specific value of the contribution according to the tag. According to the algorithm, the station with the greatest contribution is the tracing result.

5. Case studies

The effectiveness of the algorithm is verified in this part. In order to explore the path of pollution propagation in the region, we take Beijing, China as the research object. The algorithm is also suitable for the research between regions. Therefore, we conducted pollution propagation restoration for the severe haze event in central and eastern China.

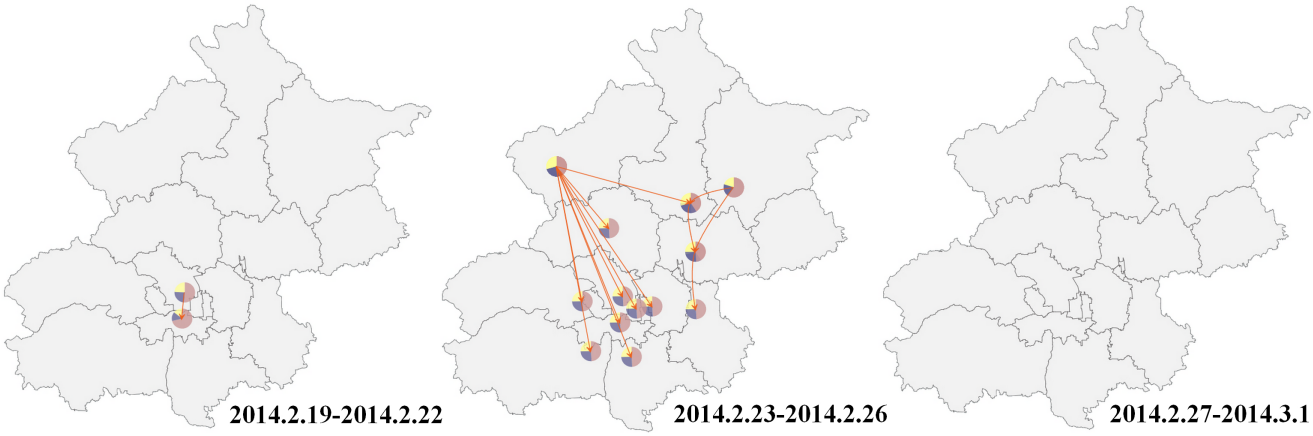

Fig. 3: Pollution propagation path map of Beijing

Combined with the follow case studies, the algorithm and system are further improved.

5.1. Beijing pollution event

As the capital of China, Beijing's industrial economy is developing rapidly. However, the excessively rapid industrial development has led to a sharp deterioration of air conditions in Beijing. Many smog incidents have seriously endangered the lives and health of local people. From 2013 to 2017, in order to prevent the spread of air pollutants in the region or even outside the region, Beijing stepped up the rectification of Beijing's industry and formulated a number of measures against air pollution. As a typical city of air pollution transmission, we choose Beijing as the first case to analyze. Based on the data set and related work, we find that Beijing had a very obvious phenomenon of pollution transmission from February 23, 2014 to February 26, 2014. In order to study the process of air pollution event in Beijing, we set the length of the algorithm's sliding window as 3 days/cycle and the sliding step as 1 day/time. Meanwhile, in order to achieve the optimal effect, we adjust the distance threshold to 300 and the pollution threshold to 200. The result of tracing path is shown in Fig. 3. It shows the source of pollutants and propagation paths in districts of Beijing.

5.1.1. Early stage of pollution event (2014.2.21-2014.2.23)

The data set shows that the air pollution index was good before February 23, with the AQI remaining between 51 and 100. The tracing results only show that the pollution spread from Fengtai District of Beijing to Haidian District, with a low degree of influence.

5.1.2. Middle stage of pollution event (2014.2.24-2014.2.26)

At this stage, the overall AQI of Beijing increased to 300-350, reaching the level of severe pollution. Compared with the early stage, the tracing results become more complex and diverse. We can know the pollution sources of each area through Fig. 3. Meanwhile, The degrees of pollution in the polluted area are shown in Fig. 4. We find that Tongzhou District was seriously affected since the pillar pointing to it is the thickest.

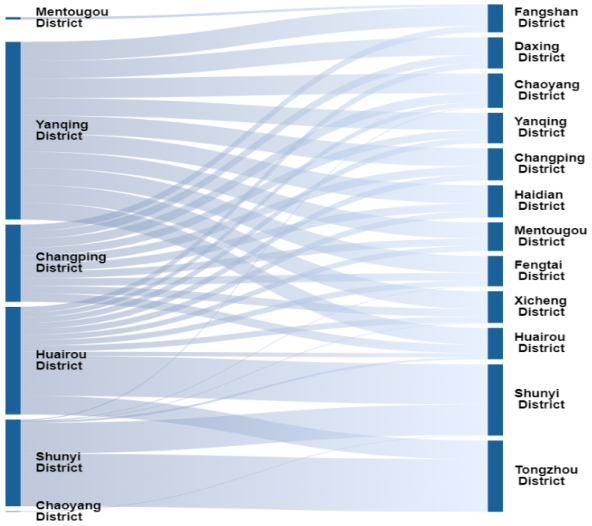

Fig. 4: Influence degree sankey chart of Beijing

To analyze the pollution factors in Tongzhou District, we draw Fig.5. Fig. 5 shows that the main factor of transmission is wind. The northwest wind gradually strengthened in Tongzhou area from 23rd to 26th, which makes Shunyi in the northwest spread the pollutants to Tongzhou. The concentration of pollutants in Shunyi is higher, leading to more pollutants spreading to Tongzhou. Combined with the Fig. 6, we can see that the pollutants in Tongzhou rose rapidly in these four days, reaching a maximum of 460.38. All these information above shows that Tongzhou was infected by Shunyi and Huairou because of the wind.

Simultaneously, Fig. 7 shows that Yanqing District had a wide spread of influence and had a certain degree of pollution impact on 9 surrounding areas, including Changping,

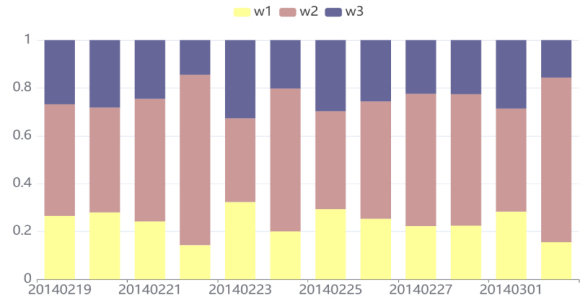

Fig. 5: Distance weight distribution chart of Beijing

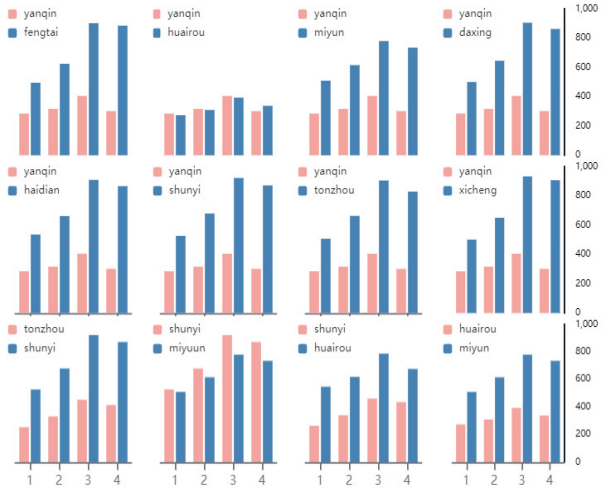

Fig. 6: Trend comparison chart of Beijing

	Fengtai District	Daxing District	Yanqing District	Huairou District	Changping District	Chaoyang District	Haidian District	Xicheng District	Tongzhou District	Mentougou District	Shunyi District
Fengtai District	-108.66	-171.07	304.3	75.04	152.53	-41.68	-10.43	-64.45	-109.61	-59.34	34.7
Daxing District	-94.65	-149.01	265.07	65.36	132.87	-36.29	-9.08	-56.14	-95.47	-51.69	30.24
Yanqing District	-106.05	-166.95	296.98	73.23	148.86	-40.67	-10.18	-62.9	-106.97	-57.92	33.87
Huairou District	-103.9	-163.58	290.98	71.75	145.86	-39.85	-9.97	-61.63	-104.8	-56.75	33.19
Changping District	-105.83	-166.61	296.37	73.08	148.56	-40.59	-10.16	-62.77	-106.75	-57.79	33.8
Chaoyang District	-108.73	-171.18	304.5	75.09	152.63	-41.7	-10.44	-64.49	-109.68	-59.38	34.72
Haidian District	-105.88	-166.7	296.53	73.13	148.63	-40.62	-10.17	-62.8	-106.81	-57.81	33.81
Xicheng District	-98.53	-155.13	275.95	68.05	138.32	-37.79	-9.46	-58.44	-99.39	-53.81	31.47
Tongzhou District	-100.92	-158.9	282.67	69.72	141.68	-38.72	-9.69	-59.87	-101.81	-55.1	32.22
Mentougou District	-103.1	-162.31	288.73	71.19	144.73	-39.53	-9.89	-61.15	-103.99	-56.32	32.93
Shunyi District	-99.59	-156.78	278.89	68.77	139.8	-38.19	-9.56	-59.07	-100.45	-54.39	31.81

Fig. 7: Reachability matrix thermal diagram of Beijing

Huairou and Haidian. Although Yanqing District spread pollution to surrounding areas, the overall trend of transmission was from the northwest to the south of Beijing. Combined with the dominant factor shown in the trend comparison chart, we infer that the wind in Beijing was mainly from the west, and the wind tended to be excessive to the northwest. To verify the visual results, we investigate the meteorological conditions in Beijing at this stage and get the same results. This verification confirm the accuracy of the model inference and confirm the validity of the model.

5.1.3. Late stage of pollution event (2014.2.27-2014.3.1)

The air pollution level in Beijing improved significantly on 27th. The AQI index was below 50, reaching an excellent level. The tracing results only show the pollution influence of Daxing District on Xicheng District, and the pollution influence degree was extremely low which is almost negligible. The humidity changed greatly from 68.096 on the 26th to 25.298 in this stage. The temperature droped more and the possibility of rainfall was great. Weather conditions greatly promoted the deposition of pollutants and severely reduced the serious pollution from 23rd to 26th. The actual situation shows that under the action of strong cold air and north wind, pollutants were effectively removed. The monitoring data shows that the concentration of fine particles was between 3-21 $\mu g/m^3$, which means the air quality reached the first-class excellent level.

The study of Beijing pollution event is a tracing study in a polluted region. According to above analysis, the dominant factor of the event is wind direction and wind speed. Our algorithm and visualization system make an accurate and comprehensive analysis of the event process. The transmission path of pollutants is shown for relevant experts to make follow-up treatment research. Relevant experts show that the tracing results are effective and beneficial to pollution control and policy-making.

5.2. Severe haze event in central and eastern China

The central and eastern regions of China have less wind in winter, and the meteorological conditions are not conducive to the diffusion of air pollutants from central and southern North China to the Yangtze River Delta region. Internal settlement and diffusion of air pollutants in the central and eastern regions often occur.

In December 2013, the most serious long-term and large-scale spread process of high-concentration particulate matter occurred in the middle and eastern regions of China. This event is called the severe haze event in central and eastern China in December 2013(Li et al., 2015). The pollution spread of this incident involved Tianjin, Hebei, Shandong, Jiangsu, Shanghai and other provinces. The daily average concentration of $PM_{2.5}$ exceeded 150 $\mu g/m^3$, and even reached 300-600 $\mu g/m^3$ in some areas, with the air quality index reaching level 6 (Fig. 8).

Therefore, we carries out a series of studies on the tracing process of air pollution in some central and eastern provinces (Hebei, Shandong, Jiangsu and Henan). We hope that the algorithm could identify effective pollution sources and provide suggestions for further treatment and policy intervention. This paper studies the track of $PM_{2.5}$ in four central and eastern provinces from December 1st, 2013 to December 10th, 2013. To verify whether the tracing results are in line with the actual situation, this paper uniformly sets the sliding window value to 3 days/cycle, the sliding step size to 1 day/time, the distance threshold to 2000 km, and the pollution threshold to 50. The second distance threshold is set as 6000 km to effectively study the cases of medium

and long-distance tracing process. The final tracing results are shown in Fig. 9.

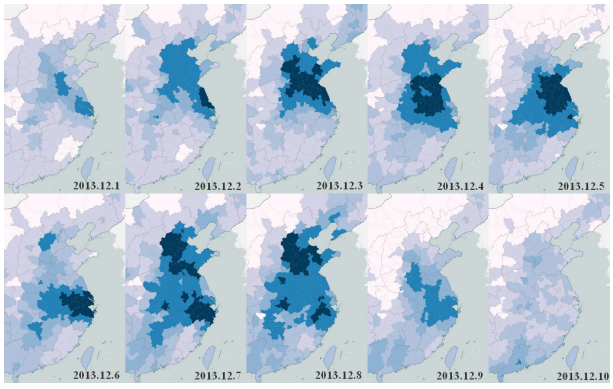

Fig. 8: Thermodynamic map of haze event

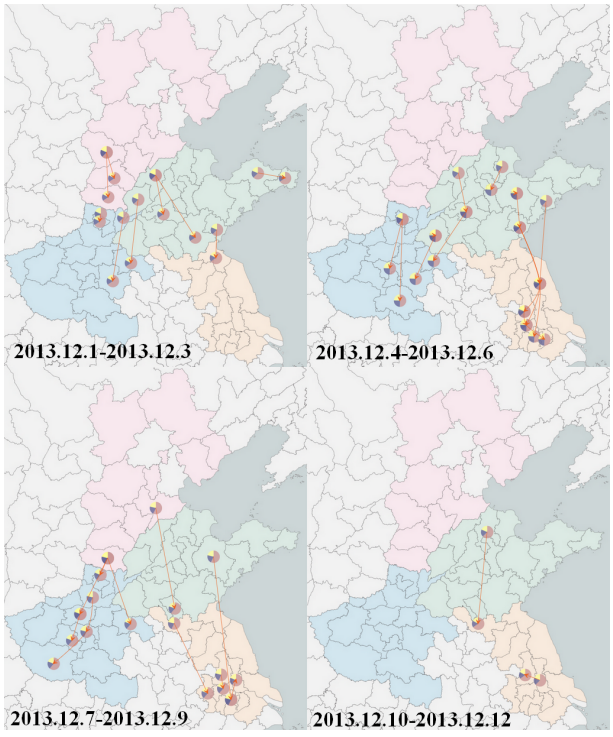

Fig. 9: Pollution propagation path map of haze event

5.2.1. *Early stage of haze event (2013.12.1-2013.12.3)*

Since December 1, $PM_{2.5}$ had appeared in Shandong and Henan. The tracing results show that the phenomenon of fine particulate pollutants transmitted to the south in Hebei. Liaocheng, Shandong and Puyang, Henan also gradually transmited fine particulate pollutants to the southwest, resulting the increase of $PM_{2.5}$ in Shangqiu and Zhoukou. The internal pollutants in Shandong not only affected the southwest cities but also transported fine particles of pollutants to the southeast of Shandong and the northeast of Jiangsu. For example, Rizhao City affected the aggravation of pollution in Lianyungang City. According to the diffusion and aggravation of pollution, the scope of pollution was expanded to most areas of north and east China on December 2. On December 3, the pollution situation continued in most parts of the East. AQI indexes of several cities were higher than 200 and up to more than 400.

5.2.2. *Middle stage of haze event (2013.12.4-2013.12.6, 2013.12.7-2013.12.9)*

The middle stage of the pollution transmission event is the period when the pollution was the most serious and the pollution transmission was more obvious and complicated. Therefore, we focus on using views to analyze the middle stage of pollution transmission event. Fig. 8 shows that Jiangsu was seriously polluted and Fig. 9 shows that it had a wide range of pollution transmission. Firstly, we analyze that the wind is the dominant factor in the stage. According to the meteorological map (Fig. 10), we find that the wind direction gradually changed to the northeast in December, which caused pollutants spreading from Zhangjiakou City to Hengshui City. Meanwhile, pollution spreaded more seriously in Jiangsu, starting from Yancheng to Changzhou City, Yangzhou City and other cities in the southeast of Jiangsu.

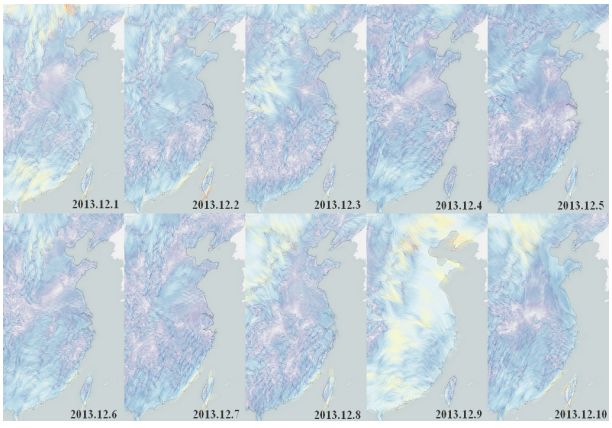

Fig. 10: Meteorological map of haze event

Although the atmospheric diffusion conditions improved during December 5 and 6, the pollution influence in Jiangsu was still increasing. In this period, the AQI in southeast Jiangsu increased sharply, exceeding 600 and reaching the serious pollution level.

From December 7 to December 9, affected by the high-altitude cyclone, the relative humidity in Shandong increased and the wind speed decreased. The air condition basically stopped the spread of internal pollutants. However, there was a wide spread in Henan. From Fig. 9, since December 7, the northeast of Henan (such as Hebi City and Xinxiang City) had a pollution impact on the cities in central Henan. From December 8 to 9, the pollutants moved further to the southwest, affecting as far as Nanyang City.According to Fig. 11, the AQI values of all cities in Henan showed an upward trend during this stage. The AQI value of Xuchang City reached 600 and the main influent was $PM_{2.5}$. It reached the level of moderate pollution.

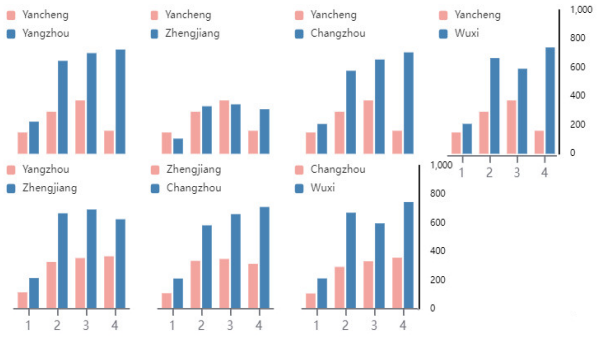

Fig. 11: Trend comparison chart of haze event

According to Fig.10, we can see that see an increase in wind speed on December 9, resulting in a decrease of pollutants. The same information can also be reflected in Fig 9. We can see that due to the increase of the wind speed on the 9th, the number of pollution tracks from December 10th to December 12th also decreased sharply.

5.2.3. Late stage of haze event (2013.12.10-2013.12.12)

During this period, the pollution transmission in various provinces and cities was greatly reduced. The pollution transmission events basically stop. According to the query of relevant meteorological factors, there was a large area of rainfall in the later stage of transmission and the humidity increased in many cities in the southeast. This leaded to a large settlement of pollutants. The inference obtained in this paper is basically consistent with the objective conditions.

5.3. Effectiveness analysis of algorithm

The severe haze event in central and eastern China is a case of pollution transmission involving many provinces. According to the above research, our algorithm can effectively restore the pollution propagation paths. Assisted by the visualization system, relevant experts can better explore the law of pollution propagation. Experts believe that inter regional collaborative governance is essential in air pollution control. The division of multi regions should consider the links between regions in economy, transportation, policy and so on. Therefore, the case further verifies the effectiveness of our work.

6. Discussion

6.1. Universality analysis of algorithm

This part mainly discusses the universality test of algorithm input. The algorithm is less dependent on the data set than the traditional forward propagation and backward trajectory algorithms. The algorithm can also guarantee the accuracy of backward trace tracing results when the requirement of data set is not high.

The common PSCF factor method (Begum et al., 2005) for identifying sources based on air flow track analysis requires obtaining the residence time of air mass in each region. However, it is difficult to collect this data set. HYSPLIT4 and LDM models are widely used to trace the particle sedimentation at different times, so they require higher height data. However, in the process of studying backward trajectory propagation, it is difficult to accurately obtain the data of every influence factor at every moment. Inaccurate impact factor data have a significant impact on both process variables and final results. Therefore, the above algorithms or models are highly dependent on the data set. The deviation of data sets often leads to error in tracing results.

Our algorithm requires less precise data input. In addition, user can effectively adjust some thresholds and parameters based on the system. It is of great help to remedy the defect of data accuracy and obtain satisfactory tracing results. Therefore, the algorithm is generally applicable to most data sets. It surpasses most existing algorithms and models in universality.

6.2. Sensitivity analysis of algorithm

This part is the sensitivity analysis of the algorithm. By analyzing the parameters of the algorithm, we find that the distance threshold and the pollution threshold are parameters with great uncertainty. So we focus on the sensitivity analysis of distance threshold and pollution threshold.We take Beijing from February 23, 2014 to February 26, 2014 as a case for analysis. The threshold value is taken as the variable, and the number of polluted routes output by the algorithm is taken as the dependent variable. We control the distance threshold value and pollution threshold value within the interval of [0,300] respectively, and collect the number of pollution routes output by the algorithm. Finally, we draw a line chart (Fig. 12). The abscissa is the value of the threshold, and the ordinate is the number of polluted routes. The blue line shows the relationship between the pollution threshold and the number of pollution routes. The orange line shows the relationship between the distance threshold and the number of polluted routes. According to the figure, when the distance threshold is constant, the number of pollution routes decreases with the increase of the pollution threshold. When the pollution threshold is constant, the number of pollution routes increases with the increase of distance threshold.

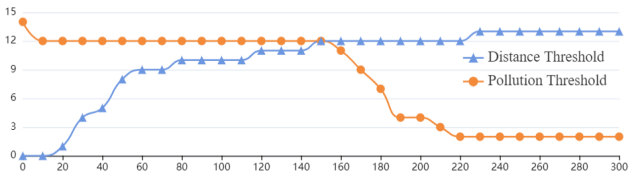

Fig. 12: Sensitivity analysis of algorithm

In the sensitivity analysis of this case, the number of pollution routes output by the algorithm changed within a reasonable range by adjusting the distance threshold and pollution threshold values respectively. This conforms to the principle that the output of the algorithm varies with the input parameters. At the same time, the range of variation

is within a reasonable range, which further proves that our algorithm has strong stability.

6.3. User study

The visualization system we designed is a tool to feed back the algorithm output to users in time. Users can further analyze the backward trajectory results based on the output of the algorithm, so as to explore the causes of pollution and explore the propagation law.

After investigation, we find that MeteorInfo (Wang, 2014) is a software for effective interaction of backward trajectory algorithm, which is widely used in the case study of atmospheric propagation. The software well simulates the results of HYSPLIT4 with Concentration Weighted Trajectory (CWT) algorithm. However, the software has high requirements for data and the universality of the algorithm is not high. In addition, based on the traditional backward trajectory algorithm, the output result of the software is single. The software can only provide users with trajectory map and contribution heat map and lacks the evaluation of influence factors.

For the problem driven visualization system, it is not feasible to use qualitative indicators to prove the effectiveness of the system. Therefore, based on above discussion, we recruited twenty volunteers, including ten students majoring in environment who have used MeteoInfo software and ten students majoring in computer science without professional knowledge in the field of air pollution transmission. In the survey, they were asked to complete the following three tasks and explain their experience:

- Task 1: Find the source station with the most wide-spreading transport of pollutants;
- Task 2: Make a knowledge of and compare the basic information of cities involved in the propagation process;
- Task 3: Determine the main influencing factor of pollution process in different stages.

After learning about our algorithm and system, volunteers began to explore MeteoInfo software and our system for the above tasks. We summarize the feedback and draw the following conclusions:

For Task 1, all volunteers can clearly find the source station with the most wide-spreading transport of pollutants through the path map and the corresponding pollution contribution graphic. However, the results of traditional algorithms are more limited than our algorithm. The traditional algorithm obtains the only station strictly based on data. According to parameter feedback of users, the results obtained by our algorithm are more consistent with the reality and also provide a more comprehensive technical reference for pollution control. At the same time, four volunteers mentioned that from the research objectives they selected, when the pollution involves multiple stations and the color of the heat map becomes complex, the heavily polluted areas will interfere with the research of other cities. However, the reachability matrix thermal diagram and the influence degree sankey chart provided by the visualization system do not have this problem.

For Task 2, we find that the MeteoInfo software can not understand the basic information of the city specified by the user. Therefore, volunteers can not complete the comparison and further exploration of urban information through MeteoInfo software. However, through our system, when users are curious about the information of a city in the pollution track or try to explore the change trend of stations, the trend comparison chart provides a comprehensive reference.

For Task 3, the HYSPLIT4 with Concentration Weighted Trajectory (CWT) algorithm adopted by MeteoInfo software lacks consideration of wind speed and direction. At the same time, the black box technology used in the software can not reasonably analyze the influencing factors of pollution for users. Therefore, all volunteers said that, the software could not complete Task 3. However, our algorithm comprehensively considers geographical and meteorological factors to realize distance correction. The pie distribution next to the station in the pollution propagation path map can make users clearly understand the distribution of pollution influencing factors in each path. The distance weight distribution chart also shows the influencing factors in the time dimension. The volunteers said they could clearly observe the change of the influence degree of humidity or wind and guess the occurrence of rainfall or extreme weather conditions.

Based on the above user study, our algorithm and system have obvious effectiveness in analyzing the causes of pollution transmission and comparing the trends of different regions. Nevertheless, MeteoInfo software is still a mature and classic analysis platform, which can solve many backward trajectory problems. Our system aims to become an auxiliary tool for algorithm output and user satisfaction to facilitate subsequent analysis and research more conveniently.

7. Conclusion and future work

The prevention and control of air pollutants is an important issue in environmental governance. In order to assist meteorologists in process of governance, a topology-based air pollution tracing algorithm is proposed in this paper. Combined with Chapman-Enskog binary diffusion theory and Gaussian diffusion model, the algorithm fully considers the spatial distance and meteorological factors. Compared with the traditional pollution propagation models, the advantages of our algorithm are as follows:

- This paper focuses on the backward trace of air pollutants. Compared with the forward air pollution propagation models, this algorithm is more suitable for the accurate location of pollutant sources. Therefore, the algorithm is more suitable for the environmental governance. Our work also provided a new idea for the prevention and control of air pollutants.
- By introducing the mathematical method of reachability matrix, the algorithm steps are reasonable and effective in the field of mathematical geometry. Topological geometry takes the positional relationship between objects

into sufficient consideration. Therefore, it is more suitable for tracing the source of pollutants between regions and under global. Our algorithm and system provide technical support for collaborative governance between regions.

- The algorithm goes beyond the ordinary Gaussian diffusion model and considers the transmission factors of pollutants more comprehensively. With full consideration of geographical and meteorological factors, the accessibility of pollution transmission path becomes more reasonable and accurate.

Combined with the designed air pollution tracing visualization system, this paper conducts the case analysis of the air pollution situation around 2014. Finally, the tracing results of this paper are basically consistent with the local investigation, which verifies that the algorithm provided in this paper is true and effective. However, there are still some improvements in the algorithm. For example, we can further consider the different transmission modes of six pollutants because of the difference of chemical composition. Due to the lack of description of extreme weather in data set, we do not consider different extreme weather when using this algorithm and system.

After perfecting the algorithm and system, we will promote the online release of the system, so that more users can use our system to trace air pollutants.

References

Ahmad, M., Chandio, A.A., Solangi, Y.A., Shah, S.A.A., Shahzad, F., Rehman, A., Jabeen, G., 2021. Dynamic interactive links among sustainable energy investment, air pollution, and sustainable development in regional china. Environmental Science and Pollution Research 28, 1502–1518.

Begum, B.A., Kim, E., Jeong, C.H., Lee, D.W., Hopke, P.K., 2005. Evaluation of the potential source contribution function using the 2002 quebec forest fire episode. Atmospheric Environment 39, 3719–3724.

Carvalho, J.C., De Vilhena, M.T.M., 2005. Pollutant dispersion simulation for low wind speed condition by the ils method. Atmospheric Environment 39, 6282–6288.

Chapman, S., Cowling, T.G., 1990. The mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction and diffusion in gases. Cambridge university press.

Cotrozzi, L., 2020. Leaf demography and growth analysis to assess the impact of air pollution on plants: A case study on alfalfa exposed to a gradient of sulphur dioxide concentrations. Atmospheric Pollution Research 11, 186–192.

Demuzere, M., Trigo, R., Vila-Guerau de Arellano, J., Van Lipzig, N., 2008. The impact of weather and atmospheric circulation on o3 and pm10 levels at a mid-latitude site. Atmos. Chem. Phys. Discuss 8, 21037–21088.

Deng, Z., Weng, D., Chen, J., Liu, R., Wang, Z., Bao, J., Zheng, Y., Wu, Y., 2019. Airvis: Visual analytics of air pollution propagation. IEEE transactions on visualization and computer graphics 26, 800–810.

Deng, Z., Weng, D., Liang, Y., Bao, J., Zheng, Y., Schreck, T., Xu, M., Wu, Y., 2021. Visual cascade analytics of large-scale spatiotemporal data. IEEE Transactions on Visualization and Computer Graphics .

Fan, H., Zhao, C., Yang, Y., 2020. A comprehensive analysis of the spatio-temporal variation of urban air pollution in china during 2014–2018. Atmospheric Environment 220, 117066.

Guo, F., Gu, T., Chen, W., Wu, F., Wang, Q., Shi, L., Qu, H., 2019. Visual exploration of air quality data with a time-correlation-partitioning tree based on information theory. ACM Transactions on Interactive Intelligent Systems (TiiS) 9, 1–23.

Huang, Liu, Sheng, 2018. On adjoint method based atmospheric emission source tracing (in chinese). Chin Sci Bull 63, 1594–1605.

Jin, L., Godri Pollitt, K.J., Liew, Z., Rosen Vollmar, A.K., Vasiliou, V., Johnson, C.H., Zhang, Y., 2021. Use of untargeted metabolomics to explore the air pollution-related disease continuum. Current Environmental Health Reports 8, 7–22.

Kim, S.S., Kim, Y.K., Kang, Y.M., 2021. Aorm: Fast incremental arbitrary-order reachability matrix computation for massive graphs. IEEE Access 9, 69539–69558.

Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., Wu, Q., Chen, H., Zhu, L., Wang, W., et al., 2021. A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in china based on the assimilation of surface observations from cnemc. Earth System Science Data 13, 529–570.

Li, J., Chen, S., Zhang, K., Andrienko, G., Andrienko, N., 2018. Cope: Interactive exploration of co-occurrence patterns in spatial time series. IEEE transactions on visualization and computer graphics 25, 2554–2567.

Li, J., Fan, Z., Deng, M., 2017. A method of spatial interpolation of air pollution concentration considering wind direction and speed. J. Geo Inf. Sci 19, 382–389.

Li, L., Cai, J., Zhou, M., 2015. Potential source contribution analysis of the particulate matters in shanghai during the heavy haze episode in eastern and middle china in december, 2013. Huan Jing ke Xue= Huanjing Kexue 36, 2327–2336.

Liu, D., Veeramachaneni, K., Geiger, A., Li, V.O., Qu, H., 2021. Aqeyes: visual analytics for anomaly detection and examination of air quality data. arXiv preprint arXiv:2103.12910 .

Pandey, A., Brauer, M., Cropper, M.L., Balakrishnan, K., Mathur, P., Dey, S., Turkgulu, B., Kumar, G.A., Khare, M., Beig, G., et al., 2021. Health and economic impact of air pollution in the states of india: the global burden of disease study 2019. The Lancet Planetary Health 5, e25–e38.

Parajuli, R.P., Shin, H.H., Maquiling, A., Smith-Doiron, M., 2021. Multi-pollutant urban study on acute respiratory hospitalization and mortality attributable to ambient air pollution in canada for 2001–2012. Atmospheric Pollution Research 12, 101234.

Qu, D., Lin, X., Ren, K., Liu, Q., Zhang, H., 2020. Airexplorer: Visual exploration of air quality data based on time-series querying. Journal of Visualization 23, 1129–1145.

Qu, H., Chan, W.Y., Xu, A., Chung, K.L., Lau, K.H., Guo, P., 2007. Visual analysis of the air pollution problem in hong kong. IEEE Transactions on visualization and Computer Graphics 13, 1408–1415.

Ren, K., Wu, Y., Zhang, H., Fu, J., Qu, D., Lin, X., 2020. Visual analytics of air pollution propagation through dynamic network analysis. IEEE Access 8, 205289–205306.

Scarpato, N., Cilia, N.D., Romano, M., 2019. Reachability matrix ontology: a cybersecurity ontology. Applied Artificial Intelligence 33, 643–655.

Sharovsky, R., César, L., Ramires, J., 2004. Temperature, air pollution, and mortality from myocardial infarction in sao paulo, brazil. Brazilian journal of medical and biological research 37, 1651–1657.

Stein, A., Draxler, R., Rolph, G., Stunder, B., 2015. B., cohen, md, and ngan, f.: Noaa's hysplit atmospheric transport and dispersion modeling system, b. Am. Meteorol. Soc 96, 2059–2077.

Tian, Li, Cheng, 2021. Visual analysis system for fine-grained inline relationship of air quality data. Journal of Computer-Aided Design Computer Graphics 33, 11.

Vo, L.H.T., Yoneda, M., Nghiem, T.D., Shimada, Y., Van, D.A., Nguyen, T.H.T., Nguyen, T.T., 2022. Indoor pm0. 1 and pm2. 5 in hanoi: Chemical characterization, source identification, and health risk assessment. Atmospheric Pollution Research , 101324.

Wang, Y.Q., 2014. Meteoinfo: Gis software for meteorological data visualization and analysis. Meteorological Applications 21, 360–368.

Xu, X., Hong, Y., 2012. Matrix expression and reachability analysis of finite automata. Journal of Control Theory and Applications 10, 210–215.

Zhou, Z., Ye, Z., Liu, Y., Liu, F., Tao, Y., Su, W., 2017. Visual analytics for spatial clusters of air-quality data. IEEE computer graphics and applications 37, 98–105.