Air pollution visual analysis system

Zhimo Li
Shandong University
Qingdao, Shangdong, China
202100202121@mail.sdu.edu.cn

Weiqi Quan
Shandong University
Qingdao, Shangdong, China
202100130177@mail.sdu.edu.cn

Xinghai He
Shandong University
Qingdao, Shangdong, China
202118130250@mail.sdu.edu.cn

Ruohuan He
Shandong University
Qingdao, Shangdong, China
202100130222@mail.sdu.edu.cn

ABSTRACT

With the development of the economic and social aspects and the subjective impressions of the air quality, the public's awareness of atmospheric pollution is deepening. Meteorological monitoring devices across various regions are becoming increasingly sophisticated, accumulating a large amount of analyzable data. Extracting relevant patterns of atmospheric pollution from these data and understanding its relationship with economic and health effects have become crucial and urgent issues. Today, the spatial and temporal display systems for atmospheric pollution have significantly improved. However, there is still limited analysis on pollutant classification, pollutant migration, pollutant attribution, and pollutant prediction, especially concerning their relationship with economic and health effects. Analyzing these relationships with big data technology and visualizing them effectively in the system remain challenges. In this paper, regarding the issue of pollutant prediction, an improved transformer and dataset provided in a referenced paper are used to predict various pollutants. The results are compared with the predictions made by LSTM. For pollutant attribution, a random forest method is employed to attribute pollutant indicators to factors such as temperature and GDP. Regarding pollutant classification and migration, visualization is achieved through libraries such as ECharts and D3, creating heatmaps, vector graphs, Sankey diagrams, etc. Importantly, this paper keeps up with the times by training relevant large models, utilizing web crawlers to gather related language data for training, and generating professional summaries for the convenience of experts.

1 INTRODUCTION

Meteorological monitoring devices across various regions have been increasingly refined, resulting in the accumulation of substantial analyzable data. Uncovering the patterns related to air pollution from these datasets, and understanding their correlation with economic and health effects, has become an essential and urgent research topic.Presently, spatial-temporal display systems for air pollution have seen significant improvement. However, there remains a scarcity of in-depth analyses regarding pollutant classification, migration, attribution, and prediction, particularly in relation to economic and health impacts. The challenge lies in leveraging big data technology to explore these relationships comprehensively and presenting the findings effectively within the system.Therefore, the motivation behind this project is to address this research gap, aiming to employ advanced big data techniques to thoroughly analyze the intricate connections between air pollution, economic repercussions, and health effects. This initiative seeks to contribute robust scientific support for more effective environmental management and public health policies.

The "China Environmental Status Bulletin" from 2016 to 2018 showed that 5 out of the 10 cities with the worst air quality comprehensive index rankings were located in the Beijing Tianjin Hebei region. Many studies have pointed out that the Beijing Tianjin Hebei region is a concentrated area for PM2.5 pollution and excess deaths, with about 10 percentage of excess deaths occurring in this region nationwide. Therefore, it is necessary to pay attention to the health effects of PM2.5 pollution on the population in the Beijing Tianjin Hebei region and conduct economic loss assessments.And with China's rapid economic growth and industrial development, the problem of air pollution is becoming increasingly prominent PM2.5 is one of the main atmospheric pollutants in China, and also one of the most consistent and significant health hazards in long-term exposure studies

To the best of our knowledge, existing large-scale models related to air pollution systems primarily focus on pollution prediction analysis. There hasn't been a specific task for language reporting on air pollution in these models. Therefore, we are the first to utilize a Chinese large-scale language model for the task of providing language-based reports on air pollution.

2 RELATED WORK

2.1 Large Language Model

ChatGLM-6B[1] is an open bilingual language model based on General Language Model (GLM) framework, with 6.2 billion parameters. With the quantization technique, users can deploy locally on consumer-grade graphics cards Due to the outstanding performance of ChatGPT in Chinese language tasks, we have chosen to use it as the model for our fine-tuning process.

2.2 Prediction

Cosine Transformer[2] is a Linear time-complexity Transformer with a non-linear re-weighting attention mechanism enforcing strict locality between neighboring tokens. First use of Transformers for multivariate pollutant forecasting.

2.3 Economic

The "Evaluation of Death Burden and Economic Losses from Long term PM2.5 Exposure in the Beijing Tianjin Hebei Region" points out that based on remote sensing PM2.5 data with a spatial resolution of 1 km and matched grid population data, combined with the latest Global Exposure Mortality Model (GEMM) modeled with Chinese queue data, the number of deaths attributed to long-term PM2.5 exposure in the Beijing Tianjin Hebei region in 2015 was estimated, and the corresponding economic losses were quantified.

2.4 Health

The "Health Benefit Assessment of PM2.5 Pollution Prevention and Control in China from 2013 to 2017" used a comprehensive exposure response model that combined population activity factors to evaluate the health effects attributable to PM2.5 in the eastern and central regions of China from 2013 to 2017. The relative effects of population size, aging degree, baseline mortality rate, and PM2.5 concentration on health burden assessment were quantified.

3 METHODOLOGY

3.1 Pollution Types

Based on the comparison of characteristic values of each pollutant with their standard values and upper/lower limits on specific dates and locations, pollution types are categorized. The atmospheric pollution types are classified into the following categories based on different combinations of pollutant characteristic values:

- **Standard Type:** Characteristic values of all pollutants do not exceed the upper/lower limits, and pollution characteristics show no significant changes.
- **Slightly Secondary Type:** PM2.5 exceeds the upper limit, indicating a significant influence of secondary particulate matter generation on pollution characteristics.
- **Slightly Dust Type:** PM10 exceeds the upper limit.
- **Slightly Motor Vehicle Type:** NO2 and CO characteristic values exceed the upper limit, indicating a significant impact of motor vehicle emissions on pollution characteristics.
- **Slightly Coal Burning Type:** SO2 characteristic values significantly exceed the upper limit, indicating a significant influence of coal burning emissions on pollution characteristics.
- **Slightly Fireworks Type:** PM2.5 and SO2 characteristic values significantly exceed the upper limit, combined with specific circumstances indicating an influence of fireworks burning on pollution characteristics.
- **Slightly Iron and Steel Type:** SO2, NO2, and CO characteristic values exceed the upper limit, indicating a significant impact of industrial emissions on pollution characteristics.
- **Other Types:** Other situations where there are no apparent pollution characteristic values, requiring specific analysis based on the circumstances.

3.2 Attribution Analysis

We use random forest model[3] to analysis what cause the air polution.Random forest is a machine learning method that includes multiple decision tree classifiers. Generally, the bootstrap sampling method is used, multiple samples are taken from the original sample, each bootstrap sample is modeled as a decision tree, and then these decision trees are combined to obtain the final result through voting scoring rules. Random forests are very convenient for processing large data sets, insensitive to multicollinearity, and robust to missing data and unbalanced data, and can provide prediction results with high relative accuracy. In reality, there are often hundreds of previous features in a dataset, and how to select the features that have the greatest impact on the results to reduce the number of features when building the model is a problem that we are more concerned about. In this case, we need to perform a Variable importance measure, or Feature importance evaluation, which is used to calculate the importance of the sample features and quantitatively describe the contribution of the features to the classification or regression. When the Gini index is used as the evaluation index, we denote the variable importance score as VIM and the Gini index as GI, assuming that there are J features X1... X2... X3... XJ, with I decision trees and C categories, now we need to calculate the Gini index score for each feature Xj, i.e., the average change in node splitting impurity of the jth feature across all decision trees in the RF. The formula for calculating the Gini index of node q on the ith tree node is as follows:

$$\mathrm{GI}_{\mathrm{q}^{(\mathrm{i})}}^{(\mathrm{i})} = \sum_{\mathrm{c}=1}^{|\mathrm{C}|} \sum_{\mathrm{c}' \neq \mathrm{c}} \mathrm{p}_{\mathrm{qc}}^{(\mathrm{i})} \mathrm{p}_{\mathrm{qc}'}^{(\mathrm{i})} = 1 - \sum_{\mathrm{c}=1}^{|\mathrm{C}|} (\mathrm{p}_{\mathrm{qc}}^{(\mathrm{i})})^2$$

where C indicates that there are C categories, and p_{qc} indicates the proportion of category C in node q. The importance of the feature X_j at the ith tree node q, i.e., the change of the Gini index before and after the branching of node q is:

$$\mathrm{VIM}_{\mathrm{jq}}^{(\mathrm{Gini})\,(\mathrm{i})} = \mathrm{GI}_{\mathrm{q}}^{(\mathrm{i})} - \mathrm{GI}_{\mathrm{l}}^{(\mathrm{i})} - \mathrm{GI}_{\mathrm{r}}^{(\mathrm{i})}$$

If the node where the feature X_j appears in the decision tree i is the set Q, then the importance of X in the ith tree is:

$$\mathrm{VIM}_{\mathrm{j}}^{(\mathrm{Gini})\,(\mathrm{i})} = \sum_{\mathrm{q} \in \mathrm{Q}} \mathrm{VIM}_{\mathrm{jq}}^{(\mathrm{Gini})\,(\mathrm{i})}$$

Finally, all the obtained importance scores are normalized:

$$\mathrm{VIM}_{j}^{(\mathrm{Gini})} = \frac{\mathrm{VIM}_{j}^{(\mathrm{Gini})}}{\sum_{j'=1}^{J} \mathrm{VIM}_{j'}^{(\mathrm{Gini})}}$$

3.3 Prediction

3.3.1 *Cos Transformer.*

Cosine Transformer is a Linear time-complexity Transformer with a non-linear re-weighting attention mechanism enforcing strict locality between neighboring tokens.

The attention score between the query $\tilde{Q}_i$ at position i and the key $\tilde{K}_j$ at position j is calculated using the following formula:

$$s(\tilde{Q}_i, \tilde{K}_j) = \tilde{Q}_i \tilde{K}_j^T \cos^2\left(\pi \frac{i-j}{2M}\right) = \frac{1}{2}\left[\tilde{Q}_i \tilde{K}_j^T + \tilde{Q}_i \tilde{K}_j^T \cos\left(\pi \frac{i-j}{M}\right)\right]$$

Here, $\tilde{Q}_i \tilde{K}_j^T$ represents the dot product between the query $\tilde{Q}$ at position i and the key $\tilde{K}$ at position j. The term $\cos^2\left(\pi \frac{i-j}{2M}\right)$ is the square of the cosine similarity, where i and j are position indices, and M is the maximum length of the sequence.

The final attention score is computed as the average of the dot product and the scaled cosine similarity, providing a balanced measure of similarity between the query and key at different positions.

Figure 1: prompt

4 EXPERIMENT

4.1 Random Forest

For the Random Forest model, we designate the feature columns as temperature, per capita GDP, power generation, per capita road area, population density, per capita park green space area, relative humidity, and wind speed.

The target variable is set to atmospheric pollutants (PM2.5, PM10, SO2, etc.). And we set the number of estimators to be 100.The goal is to investigate the importance of various indicators on pollutant levels. For the Random Forest model, $N_{\text{Estimators}}$ represents the number of sub-datasets generated by sampling the original dataset with replacement, essentially denoting the number of decision trees in the ensemble. Choosing an appropriate value for $N_{\text{Estimators}}$ is crucial—too small a value may lead to underfitting, while an excessively large value might not significantly improve the model [3].

	PM2.5	PM10	SO2	NO2	CO	O3
RMSE	7.45	8.77	5.02	2.45	0.13	3.16
MAE	6.26	7.10	4.08	1.94	0.11	2.46
R − squared	0.59	0.83	0.39	0.94	0.63	0.81

4.2 Large Language Model

In order to provide a comprehensible analysis of the air pollution situation, we fine-tuned a Large Language Model (LLM). Using data gathered through web scraping from the website of the Ministry of Ecology and Environment of the People's Republic of China (`mee.gov.cn`), we trained a system capable of delivering daily national air pollution reports.

Corpus Analysis. The target corpus mainly consists of the following content: a summary overview of the air quality in major regions, an assessment of key pollutants, and a final evaluation of specific locations within the region.

We input nationwide atmospheric pollution data, including AQI levels, pollution severity, and primary pollutants, into the model with a learning rate of 1×10^{-4} for 1000 steps. The training batch size is set to 1, and we choose the final checkpoint (checkpoint-1000) for inference.

Result Analysis. We compared our results with the untrained baseline model (see Appendix Figure 3) and observed significant improvement. The model can provide a comprehensive summary of

Ground Truth

6 月中上旬，我国北方地区预计受到 1 轮（4 日）弱冷空气过程影响，南方大部地区预计有持续性降水过程出现，全国扩散条件总体较好；但随气温升高，太阳辐射增强，区域臭氧污染问题凸显；伴随大风降温过程有扬沙天气出现。预计全国大部空气质量以良至轻度污染为主，其中华北、东北、西北、西南及长三角区域可能出现臭氧中度污染。

京津冀及周边区域：6 月中上旬，区域以良至轻度污染为主，首要污染物为臭氧。其中，1-3 日，山东西部和河北南部局地可能出现中度污染；4-5 日，区域北部为优良；6-9 日，区域中南部以轻至中度污染为主；10-15 日，区域整体以轻至中度污染为主。

• • •

Finetune Result

6 月 1 日，6 月上半月（6 月 1 日～15 日）全国空气质量预报。6 月上半月，全国大部扩散条件总体较好，空气质量以良至轻度污染为主，首要污染物以臭氧为主。京津冀及周边区域中南部、长三角中北部、苏皖鲁豫区域、汾渭平原和东北区域局地可能出现臭氧中度污染。新疆局地受沙尘天气影响，可能出现 PM10 中度及以上污染。

6 月上半月，京津冀及周边区域中南部、北部和南部空气质量以良至轻度污染为主，首要污染物为臭氧。其中，7-10 日以及 15-16 日，京津冀及周边区域中南部可能出现臭氧轻度至中度污染；11-12 日以及 13-15 日，京津冀及周边区域北部、苏皖鲁豫区域和汾渭平原局地可能出现臭氧轻度污染。

• • •

Base Model Result

以下是根据提供的数据生成的全国各个地区简要的大气污染通报：

东北地区：吉林省:空气质量良好,AQI 指数为 22

辽宁省：空气质量较差，AQI 指数为 74.69。

黑龙江省:空气质量较差,AQI 指数为 12.34

• • •

Figure 2: The output in Figure a represents the results of the base model, which failed to produce the expected outcome and performed poorly. Figure b shows the output after fine-tuning, with errors highlighted in red and correct information in blue annotations within the summary language.

the national atmospheric conditions, outline the pollution situation in major regions, and offer specific comments on areas with typical scenarios. The following figure is an example.

Deficiencies and Analysis. Repetition Phenomenon:1)Insufficient data volume and a low number of training epochs.2)The batch size is too small. The model may overfit on the training data, resulting in the generation of repetitive content. Fantasy:1)Lack of information in the input data.2)Certain information in the target data is disproportionately high. Errors:The model exhibits weak inference capability. Lack of certain informationThe provided data does not support the inference of certain details.

Prospects. Data Cleaning is needed,since we still find meaningless words in the generated sentences If more data and additional training resources are provided, the results may improve. With a better prompt design, the model may more easily recognize the significance. We need to designing a validation set (Issue of not having a partitioned validation set due to data volume) A more robust model may lead to better performance.

4.3 Prediction

The training set uses data from 2016-2017 and pollutant data from 2018 for prediction.30 consecutive days were randomly selected for the prediction. Figure 3, on the test data, we found that the results of CosineTransformer were not as good as those of LSTM, which may be due to the fact that the information in the dataset is less characteristic than that of the original author (such as transportation, power plant, etc.), and the data set size is different, which leads to the difference.

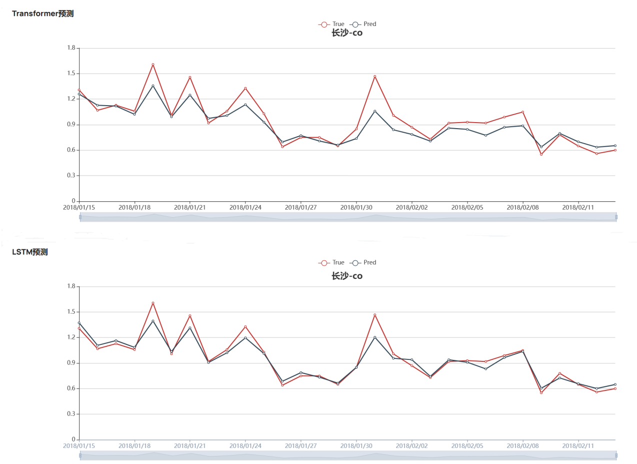

Figure 3: Transformer & LSTM

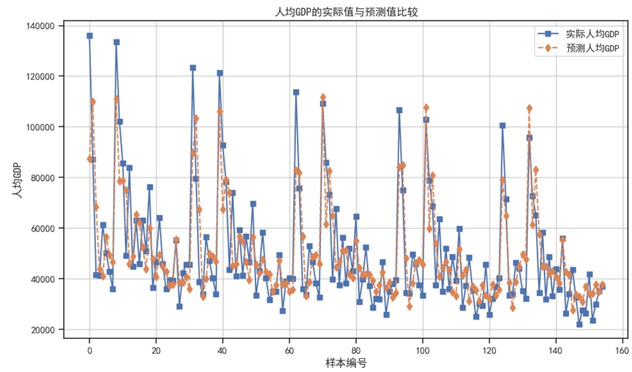

Figure 4: Fitting of Actual Values and Predicted Values

5 THE ANALYSIS OF ECONOMIC AND HEALTH EFFECTS

5.1 Economic Effects

Due to the pollution data with coarse spatial resolution, ignoring population data with spatial heterogeneity, and lacking exposure -response functions in highly polluted areas, existing research on long-term exposure disease burden of PM2.5 has significant uncertainty. Therefore, this study is based on remote sensing PM2.5 data with a spatial resolution of 1 km and matched grid population data, The Global Exposure Mortality Model (GEMM), which combines the latest data included in the Chinese cohort, estimates the number of deaths attributed to long-term exposure to PM2.5 in the Beijing Tianjin Hebei region in 2015 and quantifies the corresponding economic losses.

Based on the global exposure mortality model, evaluate the attributable deaths of long-term exposure to PM2.5 in the Beijing Tianjin Hebei region in 2015, and use the statistical life value method to estimate their corresponding health and economic losses. Method for assessing mortality effects: This study is based on the exposure response model, combined with the PM2.5 attributable mortality burden assessment method proposed by Cohen et al. in the 2015 Global Burden of Disease Study, to quantify the mortality effects of PM2.5 [4]. The calculation formula is:

$$M_{ij} = \gamma_{0i} \times AF_{ij} \times Pop_j; AF_{ij} = \frac{RR-1}{RR}$$

GEMM is an exposure response model proposed by the Burnett team in 2018 to evaluate the health effects of PM2.5. The model integrates information from 41 cohort studies in 16 countries, including large-scale cohort studies in China. It is currently the first exposure response model considering the Chinese population, covering the global exposure range of PM2.5 concentrations (2.4-84.0) G/m3). Calculation formula:

$$\text{GEMM}\left(\Delta z_j\right) = \exp\left\{\beta T\left(\Delta z_j \mid \alpha, \mu, v\right)\right\};$$
$$T\left(\Delta z_j \mid \alpha, \mu, v\right) = \log\left(1 + \Delta z_j/\alpha\right) / \left\{1 + \exp\left[-\left(\Delta z_j - \mu\right)/v\right]\right\};$$
$$\Delta z_j = \max\left(0, z_j - z_c\right)$$

Economic effect evaluation method: Using the widely used statistical life value method, the health and economic losses caused by PM2.5 pollution are further evaluated on a city by city basis. Due to the unavailability of statistical life value in some cities in the Beijing Tianjin Hebei region, this study will use Xie Xuxuan's 2010 research results on Beijing as the benchmark for statistical life value. Through the unit value transfer method, the statistical life value of 13 cities in the Beijing Tianjin Hebei region will be estimated, Specific calculation formula:

$$VSL_{l,2015} = VSL_{B,2010} \times \left(\frac{G_{l,2010}}{G_{B,2010}}\right)^{\beta} \times \left(1 + \%\Delta P_l + \%\Delta G_l\right)^{\beta}$$

In order to gain a deeper understanding of the relationship between air pollution and various economic indicators, this study employs the Ordinary Least Squares (OLS) model. Six pollutants are used as control variables, and various economic indicators serve as explanatory variables for model fitting. The results are depicted in the Figure 4:

5.2 Health Effects

In order to better evaluate the health benefits of air pollution prevention and control in China, this article uses a comprehensive exposure response model that combines population activity factors to evaluate the health effects attributable to PM2.5 in the eastern and central regions of China from 2013 to 2017. The relative effects of population size, aging degree, baseline mortality rate, and PM2.5 concentration on health burden assessment are quantified, Intended to provide a reference basis for differentiated control of air pollution and cost-benefit analysis of pollution control.

This study used the Integrated Exposure Response (IER) model developed by Burnett et al. to calculate the PM2.5 related health burden in adults aged 25 years old. The IER model was first proposed in the GBD2010 report and has been widely used in related studies on the health effects of PM2.5. We selected ischemic heart disease (IHD), chronic obstructive pulmonary disease (COPD) Lung

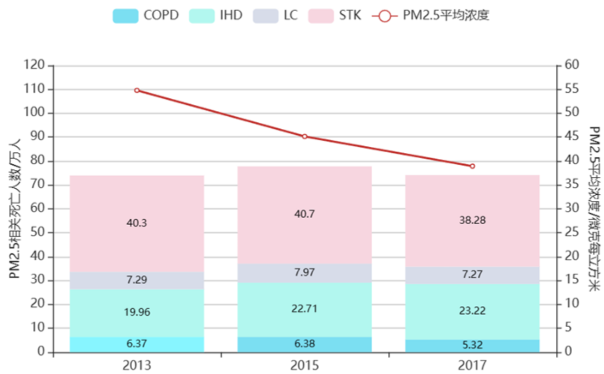

Figure 5: Pollution Change Chart

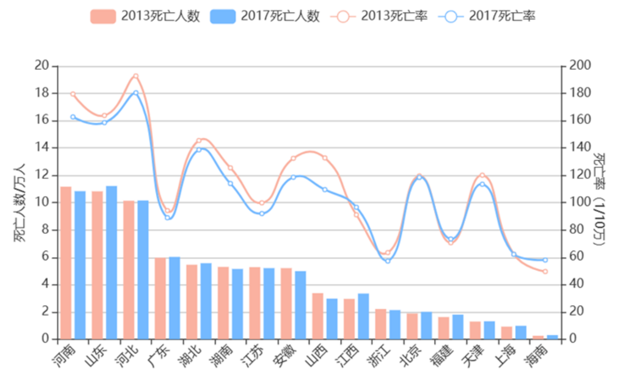

Figure 6: 2013 and 2017 Provincial Mortality Rates and Changes in the Number of Deaths

cancer (LC) and stroke (STK), the four major fatal diseases in adults, are used as health endpoints [5]. The relative risk (RR) is calculated using formula :

$$\mathrm{RR}_{\mathrm{IER}}(Z)=\begin{cases}1, & Z \leqslant Z_{\mathrm{cf}} \\ 1+\alpha\left\{1-\exp\left[-\gamma\left(Z-Z_{\mathrm{cf}}\right)^{\delta}\right]\right\}, & Z>Z_{\mathrm{cf}}\end{cases}$$

Calculate attributable deaths using formula Mor:

$$\begin{aligned}\Delta \text{ Mor } &= y_0 \times \text{ Pop } \times\left(\frac{\mathrm{RR}-1}{\mathrm{RR}}\right) \\ &= y_0 \times \mathrm{Pop} \times \mathrm{AF}\end{aligned}$$

5.2.1 Result. Figure 5, in 2017 compared to 2013, the total number of deaths attributable to PM2.5 did not change significantly. However, the mortality rate decreased from 127.68 per 100,000 people to 119.60 per 100,000 people. The number of deaths due to Ischemic Heart Disease (IHD) increased, while the numbers for Stroke (STK), Chronic Obstructive Pulmonary Disease (COPD), and Lung Cancer (LC) decreased. Notably, within these, the proportion of deaths attributed to Stroke (STK) remained the highest.

Figure 6, considering population factors, the results show that, although Guangdong has a higher total number of deaths, the mortality rate and the number of deaths per unit area are relatively low, indicating a comparatively lighter health burden. Hebei has the highest mortality rate, being three times that of Hainan. Shanghai, Tianjin, and Beijing have a lower total number of deaths, but the number of deaths per unit area is high. These three municipalities are economically powerful metropolises in China with extremely high population density, with Shanghai's population density even exceeding 10 times that of other provinces. Due to a strong positive correlation between the spatial distribution of PM2.5 pollution and population distribution, strengthening PM2.5 pollution control in densely populated mega-cities is crucial. This is more conducive to alleviating the health

The estimation results of PM2.5 health effects are not only related to changes in air quality but also associated with social factors such as total population, aging, and baseline mortality rates. While the decrease in mortality rates and the reduction in PM2.5 concentrations alleviate the health burden, these positive effects are offset by population growth and an aging population. Therefore, despite continuous improvements in PM2.5 pollution, there is no significant reduction in attributed death counts, and in some regions, an increase has been observed. Although changes in population factors are not favorable for alleviating the public health burden, controlling pollutant emissions and reducing PM2.5 concentrations remain the primary and most effective approaches to mitigate health burdens.

6 VISUALIZATION

Based on the purpose of data analysis, the system is divided into multiple analysis pages, including the main view, pollutant migration, pollutant attribution, pollution types and severity, pollutant prediction, economic benefit analysis, health benefit analysis, and large-scale models.

6.1 Main View Analysis Page

Figure 7, Main View Analysis Page includes a spatial-temporal display of China's map. Clicking on pollutants and meteorological indicators above the map displays corresponding indicators on both large and small maps, using different color schemes. Dragging the timeline below the map or clicking the calendar in the top right corner allows switching to a specific date, showing the pollution status for that day. Clicking on a province on the map enables map switching on the small map and displays a pollutant line chart for the corresponding province. Hovering over the small map displays indicator values and major pollutants. Clicking on a coordinate on the small map updates the radar chart with synchronized pollutant concentration information. A provincial AQI ranking is drawn based on the current day's AQI, with rankings changing synchronously over time. The pollutant line chart shows the concentration changes of pollutants over time in a province and can be zoomed in, and dragged to view relevant details.

6.2 Pollutant Migration Page

Figure 8, Pollutant Migration Page features a dynamic national wind field map on the left and various pollutant indicator charts on the right. Dragging the timeline or clicking the play button allows

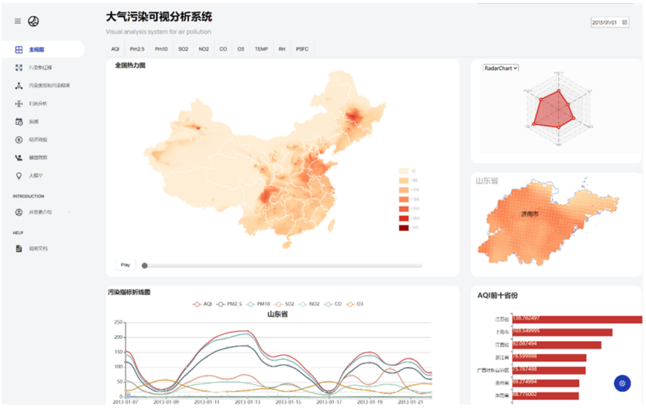

Figure 7: Main View Analysis Page

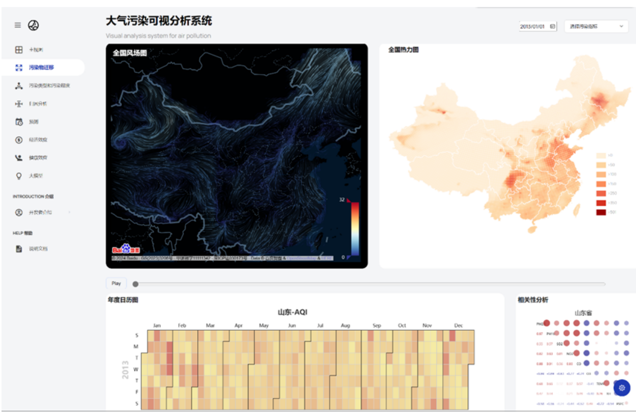

Figure 8: Pollutant Migration Page

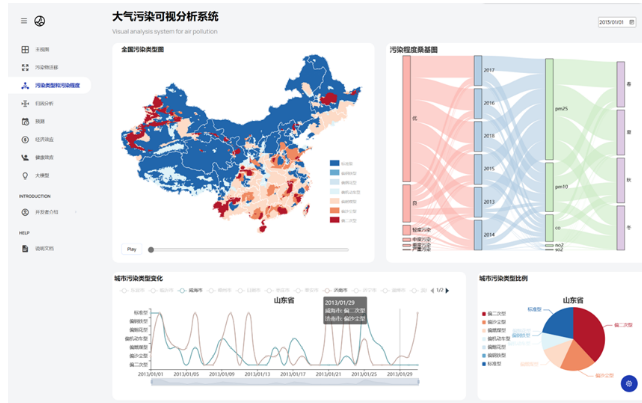

Figure 9: Pollution Types and Severity Page

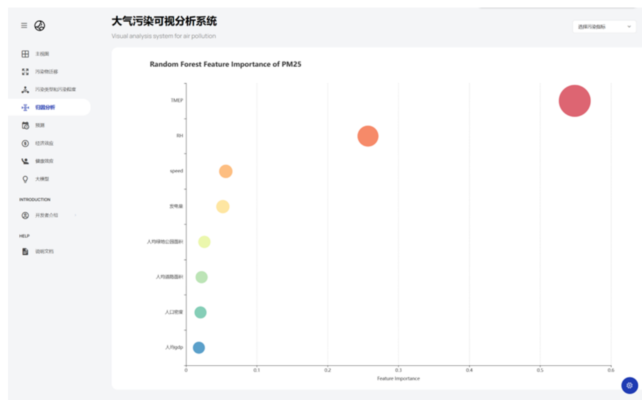

Figure 10: Pollution Attribution Analysis Page

for a dynamic display of pollutant migration. On the lower left, there is an annual calendar that enables interaction by selecting time and the pollutants to be displayed in the upper right corner. Simultaneously, it allows switching the map to display different provinces' annual indicator data. On the lower right, there is a correlation analysis showing the correlation between pollutants and meteorological indicators in different provinces, which can be switched by province.

6.3 Pollution Types and Severity Page

Figure 9, Pollution Types and Severity Page features a national pollution type distribution map on the left. Different colors on the map represent different pollution types, including standard, iron and steel, fireworks, and motorbike types, among others. The timeline can be dragged, or a calendar can be selected to display the distribution of pollution types at different times. On the right, there is a global six-year pollution severity Sankey diagram, with indicators ranging from air quality, year, major pollutants, to season. Hovering over corresponding nodes reveals the percentage distribution along the edges of the respective indicator nodes. Below is the chart depicting changes in pollution types in cities, showing the pollution types in different cities of a specified province for a month. Cities can be selected to view specific information about their pollution types. Choosing different provinces results in synchronized changes in the line chart and pie chart.

6.4 Pollution Attribution Analysis Page

Figure 10, Pollution Attribution Analysis Page, utilizing the Random Forest method to calculate the importance of various factors (such as temperature, per capital GDP, electricity generation, etc.). This process helps display the distribution of causes for different pollutants. Users can interactively choose different pollutants for analysis.

6.5 Pollution Prediction Page

Figure 11, Pollution Prediction Page employs both Transformer and LSTM models, trained on a dataset spanning from 2013 to 2017. The models are then tested using one month of data from 2018 to assess their reliability. Users can choose different cities and various pollution indicators, and the page displays corresponding predicted and actual data for the selected pollution indicator in that city. The black line represents the prediction, while the red line represents the actual data.

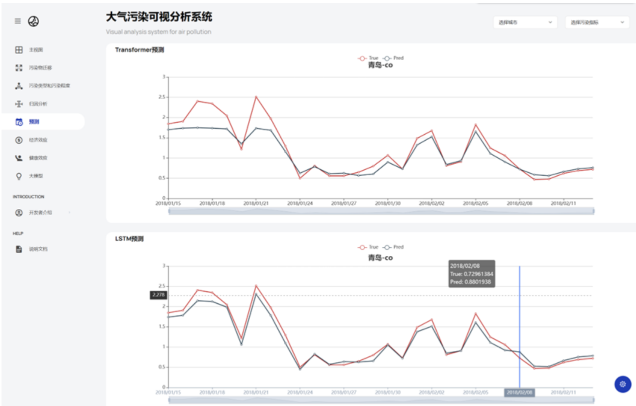

Figure 11: Pollution Prediction Page

Figure 12: Economic Benefit Analysis Page

6.6 Economic Benefit Analysis Page

Figure 12, Economic Benefit Analysis Page. The bubble chart at the top displays the three-dimensional distribution of life value, GDP, and economic losses for the Beijing-Tianjin-Hebei region. The chart on the right shows the percentage of the population in the study area. Below is a comparison between the actual and true values of economic indicators such as electricity generation and per capital GDP.

6.7 Health Impact Analysis Page

Figure 13, The Health Impact Analysis Page includes graphs depicting changes in mortality rates from 2013 to 2017, the number of deaths related to specific diseases, the number of deaths from related diseases in different provinces, and a chart illustrating avoidable deaths attributed to various factors.

7 CONCLUSION

This visualization system makes a significant contribution by offering an in-depth tool for understanding the spatiotemporal distribution and evolution patterns of air pollution, providing users with a comprehensive way to grasp environmental quality. It not only showcases the static distribution patterns of air pollution but

Figure 13: Health Impact Analysis Page

also delves into the spatiotemporal correlations between pollutants and meteorological factors, as well as the spatial distribution clustering of trends in atmospheric pollutants. This enables users to gain a more holistic understanding of the dynamic changes in the atmospheric environment.

Built on the foundation of accomplishing visual analysis tasks, the system enhances the user experience through meticulous design in visual encoding, innovative visualization techniques, and interactive features. The optimization of visual encoding ensures clearer information transmission, innovative visualization designs make complex data relationships more understandable, and improved interactivity allows users to engage more flexibly with the data, deepening their understanding of air pollution issues.

To facilitate ongoing improvements, future developments could consider incorporating more advanced data mining and machine learning techniques to enhance the understanding of complex relationships among pollution factors. Additionally, collecting user feedback and continuously optimizing the user interface and interaction design will make the system more user-friendly and intuitive. Regular updates and maintenance of data sources are crucial to maintaining the accuracy and practicality of the system. Through continuous improvement, this visualization system will provide increasingly robust support for scientific research, environmental monitoring, and decision-making.

REFERENCES

[1] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, Jie Tang.GLM: General Language Model Pretraining with Autoregressive Blank Infilling. arXiv:2103.10360
[2] Mayukh Bhattacharyya, Sayan Nag, Udita Ghosh.Deciphering Environmental Air Pollution with Large Scale City Data. arXiv:2109.04572
[3] ZHAO Yanyan, ZHANG Xiaoping, CHEN Mingxing, GAO Shanshan, LI Runkui. Regional variation of urban air quality in China and its dominant factors. Acta Geographica Sinice, 2021, 76(11): 2814-2829 doi:10.11821/dxb202111015.
[4] Wu Wenjing, Yang Xiaocui, Yao Minghong, Wu Gonghua, Xu Jiayue, Zhao Xing, Zhang Juying. Assessment of mortality burden and economic loss attributed to long-term PM2.5 exposure in the Beijing-Tianjin-Hebei area[J]. Chinese Journal of Epidemiology, 2020, 41(9): 1471-1476.
[5] ZHANG Meng-jiao, SU Fang-cheng, XU Qi-xiang, ZHANG Rui-qin, WEI Yu, XIE Liu-zhen. Health Impact Attributable to the Control of PM2.5 Pollution in China During 2013-2017[J]. Environmental Science, 2021, 42(2): 513-522.