A Time Series Analysis and Visualization of PM2.5 Distribution in China

https://github.com/Artemis20123/FangLiRenZhang_ENV872_EDA_FinalProject

Yixin Fang, Jiahuan Li, Yuxiang Ren, Jinglin Zhang

Contents

List of Tables

List of Figures

1 Rationale and Research Questions

Numerous studies have shown that air pollution has significant impacts on human health (Doe, Smith, & Williams, 2020), exposure to environments with air pollution can lead to a variety of diseases, including cardiovascular disease, mental illness, skin cancer, chronic respiratory diseases, and decreased lung function Dockery & Verrier (2005). Among the various pollutants in the environment, $PM_{2.5}$ is considered a major contributor to health problems (Pope III & Dockery, 2006). Therefore, monitoring and controlling $PM_{2.5}$ is crucial for the sustainable development of society.

With a booming economy and rapid urbanization, China has been suffering from severe air pollution problems. since China's economic reform in the 1980s, a large number of factories have been established rapidly in China Compared with the growth of industry, China did not invest a lot in environmental protection in the early stage. since 2006, China updated its air pollution control measures and tried to control the serious air pollution, In 2013, the central and eastern regions of China experienced severe haze weather, leading the central government implemented the Air Pollution Prevention and Control Action Plan to deal with serious air pollution problems, since that, China has continuously updated the "Ambient Air Quality Standards" and established a nationwide monitoring network covering nearly 400 cities, which for the first time included $PM_{2.5}$ as a monitoring target (Jin & Zhang, 2016). Due to the late start of China's $PM_{2.5}$ monitoring, the pollution status of $PM_{2.5}$ in China before 2013 was unclear. Wei et al. (2021) filled this information gap by interpreting MODIS Collection 6 MAIAC AOD to obtain the daily average concentration of $PM_{2.5}$ nationwide in China since 2000 for the first time.

China currently has seven megacities with a population of over 10 million. These cities have frequent economic activities and high population densities, understanding the changes in $PM_{2.5}$ concentrations in these areas and predicting their trends can provide important references for China's air pollution prevention and control. This project utilizes the data from Wei et al. to visualize the monthly average $PM_{2.5}$ concentrations in these areas from 2000 to 2021, allowing people to have a more intuitive understanding of the changes in local air pollution, and predicting the changes in these areas for the next three years. Our research question is:

1. How has the $PM_{2.5}$ concentration in the seven cities changed?
2. What might be the monthly average $PM_{2.5}$ concentrations in these seven cities for the next 5 years?

2 Dataset Information

2.1 Data sources

This project relies on two primary datasets. The first dataset, China $PM_{2.5}$ pollution data, is available via the Zenodo data repository(Wei et al., 2021). This dataset provides information on the daily concentration of particulate matter ($PM_{2.5}$) in the atmosphere over China, as measured by satellite sensors. The dataset contains daily average $PM_{2.5}$ concentration values for a period from 2000 to 2021, covering the entirety of China's land area. The data is presented in the form of raster images, with a spatial resolution of 0.01 degrees. This high-resolution dataset allows for accurate identification of high pollution cities and regions, and is an important resource for understanding the severity and extent of air pollution in China.

The second dataset utilized in this project is a shapefile of Chinese cities boundaries, provided by the statistical bureau of the Chinese government. The shapefile includes Chinese city names, unique identifiers, and geometry information for the city polygons. This dataset enables the mapping of $PM_{2.5}$ concentrations across different Chinese cities and supports the time series analysis. Together, these datasets form the foundation of our dashboard and other relevant analysis, providing important insights into the spatial and temporal distribution of $PM_{2.5}$ pollution in China.

2.2 Wrangling process

The project involves two main wrangling processes. The first is to convert raw NetCDF files to GeoTIFF files using Python code provided by the original dataset creators. The original NetCDF files are not included in our repository and can be downloaded from the open-access Zenodo repository. The conversion code uses the `GDAL` package, a powerful translator library for working with raster and vector geospatial data formats. It reads the SDS data from the NetCDF files and handles missing values, extracts longitude and latitude information, and finally writes the output as GeoTIFF files. The use of `GDAL` in this project ensures that the NetCDF files are accurately transformed into GeoTIFF format, and that the resulting raster data can be easily integrated with other geospatial datasets for further analysis and visualization.

In the second data wrangling process, zonal statistics are computed to aggregate the information of raster pixels within the city boundary defined in the shapefile. The `terra` package's zonal function is utilized to extract the statistics. The mean and sum $PM_{2.5}$ values of cities are generated as the outputs written in a CSV table. To extract zonal statistics for each raster file, a loop function is used to process each file in sequence. Two data frames are created to store the sum and mean values separately. These data frames are then merged together into a single data frame with two value columns. But only the mean $PM_{2.5}$ value is used in the following analysis.

Finally, all the CSV files containing separate daily $PM_{2.5}$ information are merged into one large CSV file. The daily records are then summarized at the monthly level to facilitate dashboard visualization and time series analysis.

2.3 Data structure

The $PM_{2.5}$ pollution data used in the project is derived from the ChinaHigh$PM_{2.5}$ dataset, which provides a seamless 1 km ground-level resolution of $PM_{2.5}$ distribution in the context of China. Since the original data (NetCDF files) are stored as images, tables of the derived zonal statistics are shown below:

Table 1: Monthly average $PM_{2.5}$ concentrations of Chinese cities

Variables	Class	Units	Ranges
year	numeric	T	[2000,2021]
month	numeric	T	[1,12]
city_id	numeric	N/A	1100 ~ 659001
cityname	character	N/A	N/A
meanPM	numeric	μg/m^3	[6.437489, 200.2201]

Additional metadata information can be found in the `Metadata.Rmd` file located in the "Data" folder and "Raw" subfolder of the project. This file contains detailed information on the data sources, data characteristics, variable definitions, as well as information on the geospatial reference system and units.

3 Exploratory Analysis

To explore the data, we wanted to create a visualization of the monthly $PM_{2.5}$ for the cities from 2000 to 2021. The data was first filtered to include only the seven cities of interest. Then we grouped the data by city, year, and month and summarized the mean $PM_{2.5}$ value for each month. A monthly $PM_{2.5}$ line plot was created for each city using ggplot2.

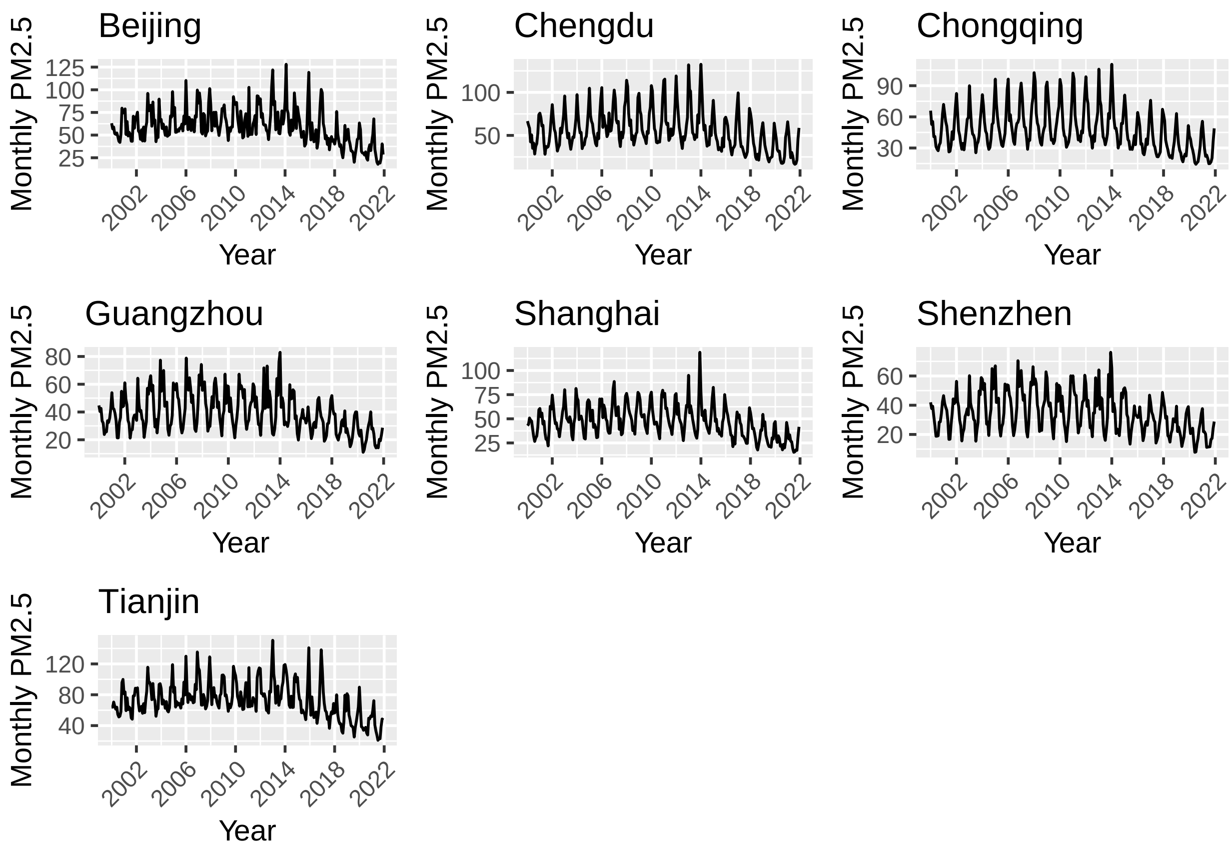

Figure 1: Time Series Visualization of the Seven Cities

Figure 1 showed that Beijing and Tianjin had the highest $PM_{2.5}$ values among the seven cities, and Shenzhen had the lowest values. Most of the cities seemed to have a slow increase in $PM_{2.5}$ from 2000 to 2006, then stayed relatively constant from 2006 until 2014. In 2014, most cities had a spike in $PM_{2.5}$. From 2014, $PM_{2.5}$ in all cities showed a decreasing trend. To better understand the changes in $PM_{2.5}$, we conducted time-series analysis and predicted the monthly $PM_{2.5}$ for the seven cities from 2022 to 2026.

4 Analysis

4.1 Part1: Time Series Analysis

4.1.1 How has the $PM_{2.5}$ concentration in the seven cities changed?

Using the STL function, we decomposed the $PM_{2.5}$ time series data of the target cities into seasonal and trend components and performed a comparative analysis. The $PM_{2.5}$ values in these cities are generally lower during the summer months (June, July, August, and September) and higher during the winter months (November, December, January, and February) (Figure 2). There are several possible factors contributing to this seasonality: First, In the summer, temperatures are higher, leading to more active air convection, which facilitates the dispersion and dissipation of pollutants in the air. Secondly, there tends to be more rainfall in the summer, which can wash $PM_{2.5}$ particles from the atmosphere to the ground, reducing their measured values (Pu, Zhao, Zhang, & Ma, 2011). Thirdly, in the winter, heating demand increases, and in some areas, coal combustion remains the primary method of providing heat (Liang et al., 2015). This leads to increased emissions of coal-related pollutants, causing a rise in $PM_{2.5}$ levels.

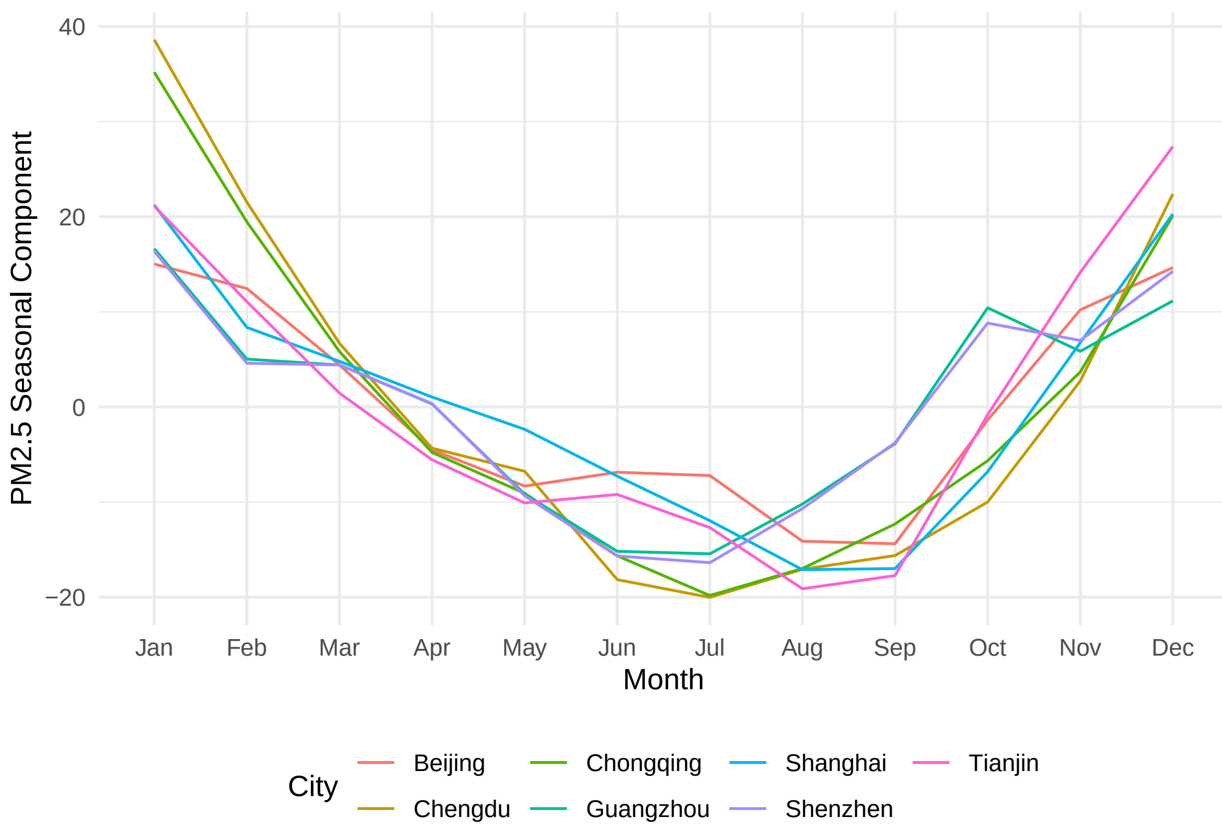

Figure 2: Seasonality in PM2.5 in 7 Chinese megacities

The trend results for the target cities are consistent with those observed in the exploratory analysis (Figure 3). Notably, around 2014 marked a turning point when $PM_{2.5}$ levels in major cities began to decrease gradually. This shift is likely attributed to the Chinese government's implementation of the Air Pollution Prevention and Control Action Plan (Cai et al., 2017). It required that, by 2017, the annual average $PM_{2.5}$ concentration in cities at or above the prefectural level should be reduced by more than 10%. In key regions such as Beijing-Tianjin-Hebei, the Yangtze River Delta, and the Pearl River Delta, the plan aimed for reductions of 25%, 20%, and 15%, respectively.

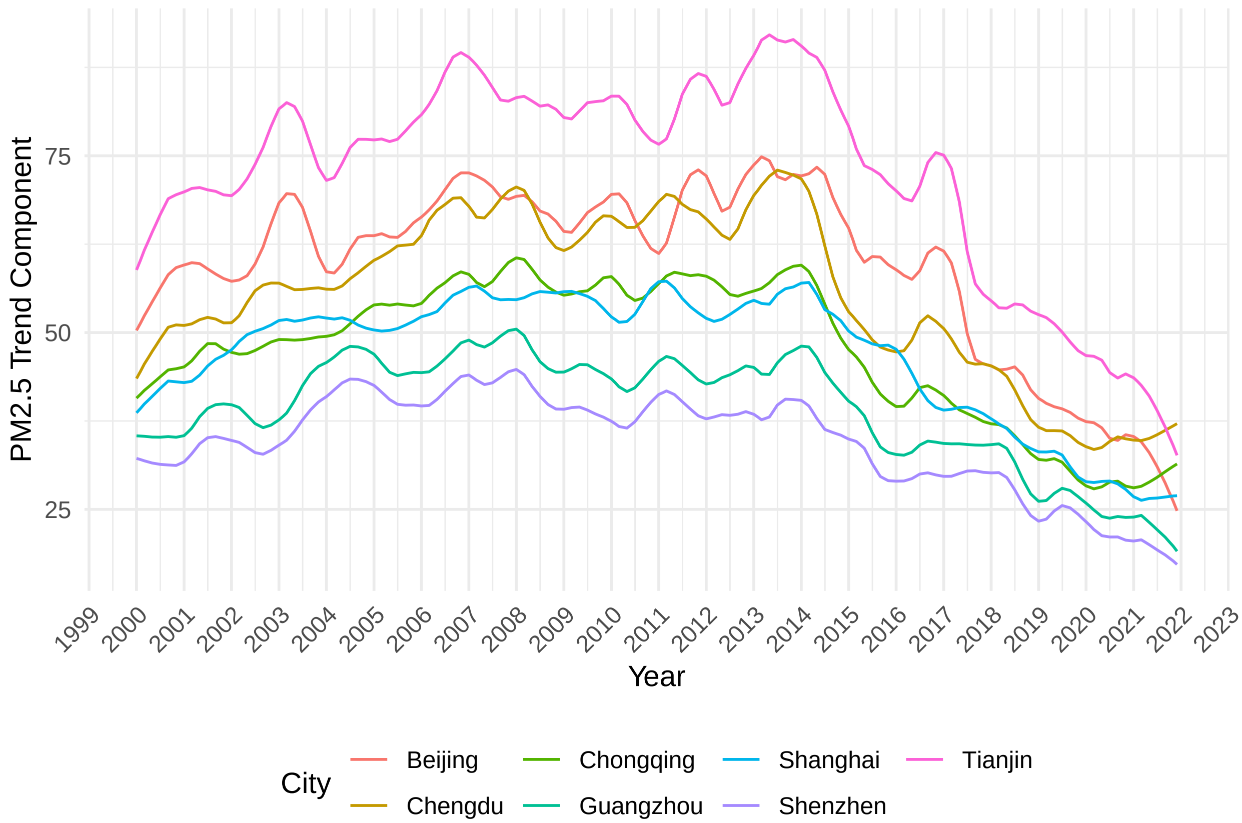

Figure 3: Trends in PM2.5 in Chinese megacities

4.1.2 What might be the monthly average $PM_{2.5}$ concentrations in these seven cities for the next three years?

In order to obtain better forecasting results, we conducted a predictive accuracy test on multiple time series models based on Beijing's data from 2000 to 2020 (using actual 2021 data as a benchmark for comparison). These models include:

1. Arithmetic Mean Model (Mean): This model assumes that future observations will equal current observations, i.e., the predicted future value equals the current value, making the model a constant average value model.
2. Seasonal Naive Model (SNAIVE): It assumes that future observations will equal the most recent observations from the same season. The model focuses solely on historical data within the same season.

Table 2: Forecast Accuracy for Seasonal Data

	ME	RMSE	MAE	MPE	MAPE
MEAN	-28.1059	31.6898	29.2560	-119.0707	120.7650
SNAIVE	-2.4987	13.4377	9.3796	-17.3618	28.6924
SARIMA	-4.3333	14.5863	12.2154	-32.3573	45.6127
SSES	4.8008	12.3008	7.7936	5.2446	19.6416

Table 3: Forecast Accuracy for Seasonal Data

	ME	RMSE	MAE	MPE	MAPE
MEAN	-28.105858	31.68980	29.256003	-119.070730	120.76499
SNAIVE	-2.498701	13.43770	9.379570	-17.361823	28.69245
SARIMA	-4.333324	14.58628	12.215420	-32.357291	45.61268
SSES	4.800801	12.30076	7.793640	5.244604	19.64161
SNAIVE_SARIMA	-3.416013	13.00090	9.967983	-24.859557	35.26887
SSES_SARIMA	0.233739	12.50184	9.574127	-13.556343	31.19364

3. Seasonal Autoregressive Integrated Moving Average Model (SARIMA): SARIMA is a widely-used seasonal time series model that builds upon the ARIMA model by incorporating seasonal variations in the time series data.
4. Seasonal Simple Exponential Smoothing (SSES) Model: This is a time series forecasting method based on weighted averages.

In these four models tested, the SSES model had the smallest Root Mean Square Error (RMSE) and the best forecasting capability (Table 1). However, considering that the forecasting target is for the next five years, both SSES and SNAIVE models predict the same values for the first, second, third, fourth, and fifth years when forecasting multiple years. As a result, we combined the results of SSES and SNAIVE with the SARIMA model, creating two new models:

5. SNAIVE_SARIMA. The average of SNAIVE and SARIMA.
6. SSES_SARIMA. The average of SSES and SARIMA.

Although the accuracy results show that the RMSE of SSES_SARIMA is the second smallest, its predictive accuracy is still slightly lower than that of SSES (Table 2). However, considering the forecasting objective and the comparison with other models, we ultimately chose to use the SSES_SARIMA model for predicting $PM_{2.5}$ levels in the target cities.

The final forecast results of Chinese megacities are shown in Figure 4.

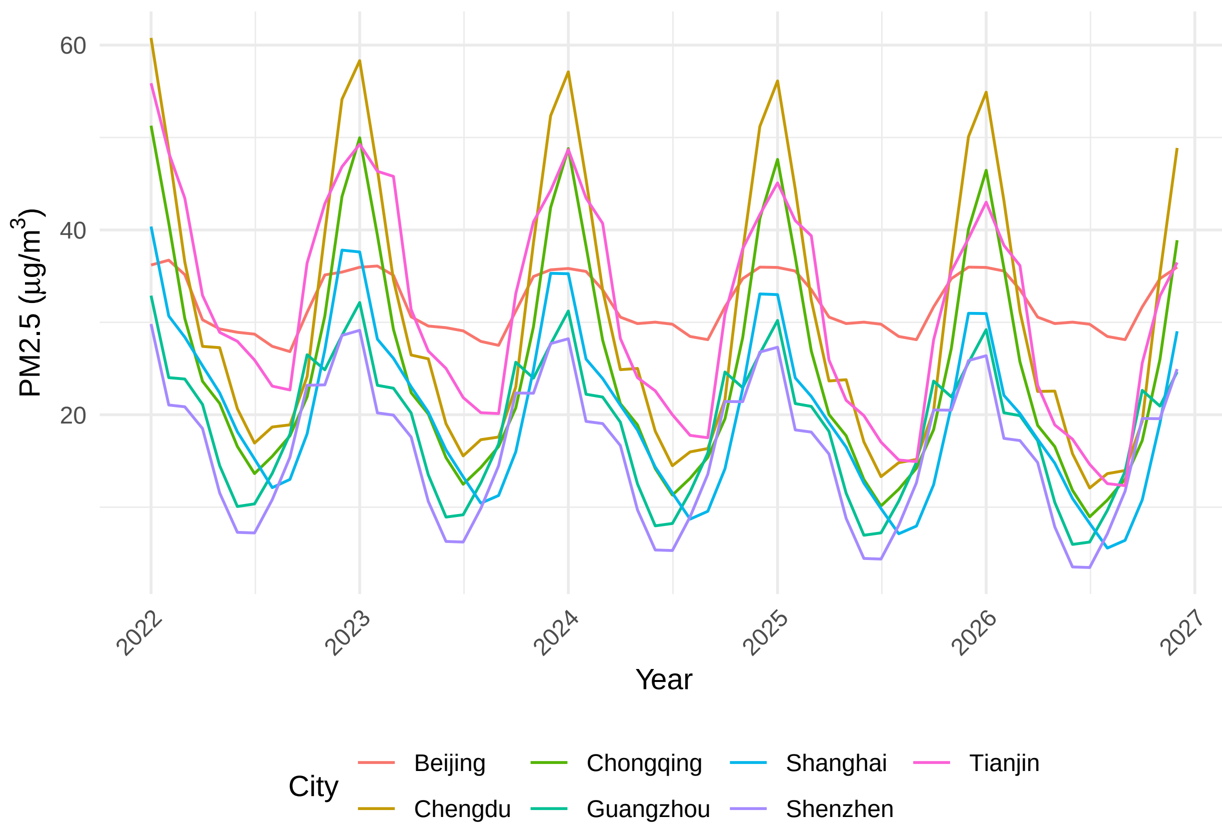

Figure 4: PM2.5 forecast in Chinese megacities

4.2 Part2: Dashboard:

To better visualize and allow users to interact with the data, we created a dashboard using Shiny. The dashboard consisted of three side panels:“PM2.5 National Distribution”, “Time Series Visualization by City”, “PM2.5 Prediction by City” and three corresponding tab panels: “Map”, “TSA”, and “Prediction”.

The “PM2.5 National Distribution” allows users to drag the map in the “Map” tab and explore the $PM_{2.5}$ distribution pattern in the country. The “Time Series Visualization by City” panel has a dropdown box that allows users to select the monthly $PM_{2.5}$ visualization of each city. The results are displayed in the “TSA” tab. Finally, the “PM2.5 Prediction by City” panel has a dropdown box and slider. Users can select the city and the time range to view the monthly $PM_{2.5}$ prediction in the “Prediction” tab.

5 Summary and Conclusions

«««< HEAD We visualized the monthly average concentration of PM2.5 in Chinese cities from 2000 to 2021 and analyzed the historical trends and predicted the concentration changes for the next 5 years in megacities. Our results show that the PM2.5 concentration increased from 2000 to 2013, with higher concentrations in the eastern and northen regions compared to the western and southern regions. Since 2014, PM2.5 concentration in China has been decreasing. The concentration during autumn and winter was higher than that during spring and summer. Combining the SRIMA and SSES models resulted in the most accurate prediction. Our prediction suggest that at the end of 2026, the highest and lowest pm2.5 concentration (” * mu * “g/m”ˆ3 * “) in Beijing, Shanghai, Guangzhou, Shenzhen, Tianjin, Chongqin, Chengdu will be 35.97, 29.04, 24.63, 24.94, 36.47, 38.88, 35.97 and 28.47, 5.57, 5.97, 3.46, 12.54, 8.95, 28.12. ======= We visualized the monthly average concentration of $PM_{2.5}$ in Chinese cities from 2000 to 2021 and analyzed the historical trends and predicted the concentration changes for the next 5 years in megacities. Our results show that the $PM_{2.5}$ concentration increased from 2000 to 2013, with higher concentrations in the eastern and northen regions compared to the western and southern regions. Since 2014, $PM_{2.5}$ concentration in China has been decreasing. The concentration during autumn and winter was higher than that during spring and summer. Combining the SRIMA and SSES models resulted in the most accurate prediction. Our prediction suggest that at the end of 2027, the highest and lowest $PM_{2.5}$ in Beijing, . »»»> 9495f5cbc0e9280d5b10949ea3d1cb4d3daeea4b

References

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, *360*(9341), 1233–1242.

Cai, S., Wang, Y., Zhao, B., Wang, S., Chang, X., & Hao, J. (2017). The impact of the "air pollution prevention and control action plan" on PM2. 5 concentrations in jing-jin-ji region during 2012–2020. *Science of the Total Environment*, *580*.

Dockery, L.-G., D. W., & Verrier, R. L. (2005). Association of air pollution with increased incidence of ventricular tachyarrhythmias recorded by implanted cardioverter defibrillators. *Environmental Health Perspectives*, *113*(6), 670–674.

Doe, J., Smith, D., JaneFowler, & Williams, M. L. (2020). Global air quality, past present and future: An introduction. *Philosophical Transactions of the Royal Society A*, *378*(2183), 20190323.

Jin, A., Y., & Zhang, S. (2016). Air pollution control policies in china: A retrospective and prospects. *International Journal of Environmental Research and Public Health*, *13*(12), 1219.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., ... Chen, S. X. (2015). Assessing beijing's PM2. 5 pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *471*.

Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air and Waste Management Association*, *56*(6), 709–742.

Pu, W., Zhao, X., Zhang, X., & Ma, Z. (2011). Effect of meteorological factors on PM2. 5 during july to september of beijing. *Procedia Earth and Planetary Science*, *2*, 272–277.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., ... Cribb, M. (2021). Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in china: Spatiotemporal variations and policy implications. *Remote Sensing of Environment*, *252*, 112136.