Insert title of project here

Web address for GitHub repository

Name

Contents

List of Tables

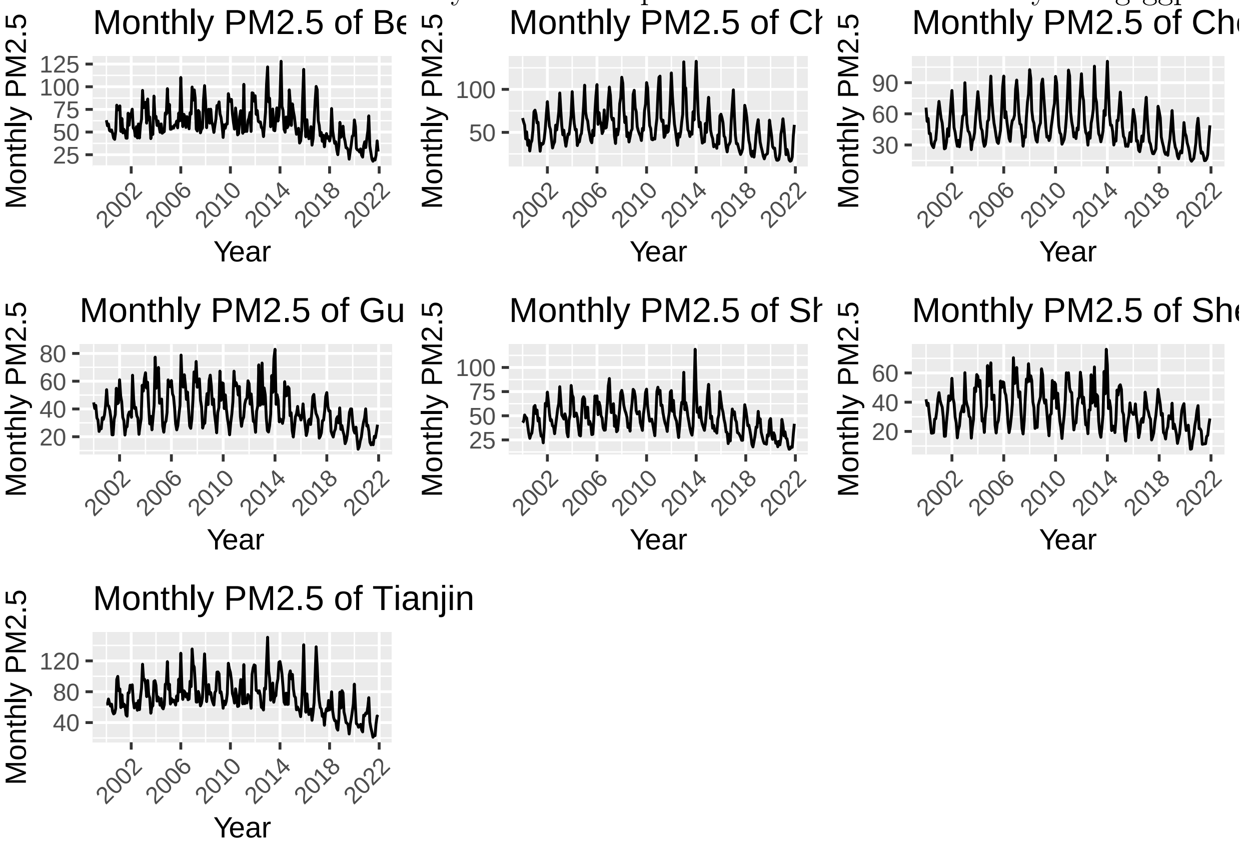
List of Figures

1 Rationale and Research Questions

2 Dataset Information

3 Exploratory Analysis

To explore the data, we wanted to create a visualization of the monthly PM2.5 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 PM2.5 value for each month. A monthly PM2.5 line plot was created for each city using ggplot2.



The plot showed that Beijing and Tianjin had the highest PM2.5 values among the seven cities, and Shenzhen had the lowest values. Most of the cities seemed to have a slow increase in PM2.5 from 2000 to 2006, then stayed relatively constant from 2006 until 2014. In 2014, most cities had a spike in PM2.5. From 2014, PM2.5 in all cities showed a decreasing trend. To better understand the changes in PM2.5, we conducted time-series analysis and predicted the monthly PM2.5 for the seven cities from 2022 to 2026.

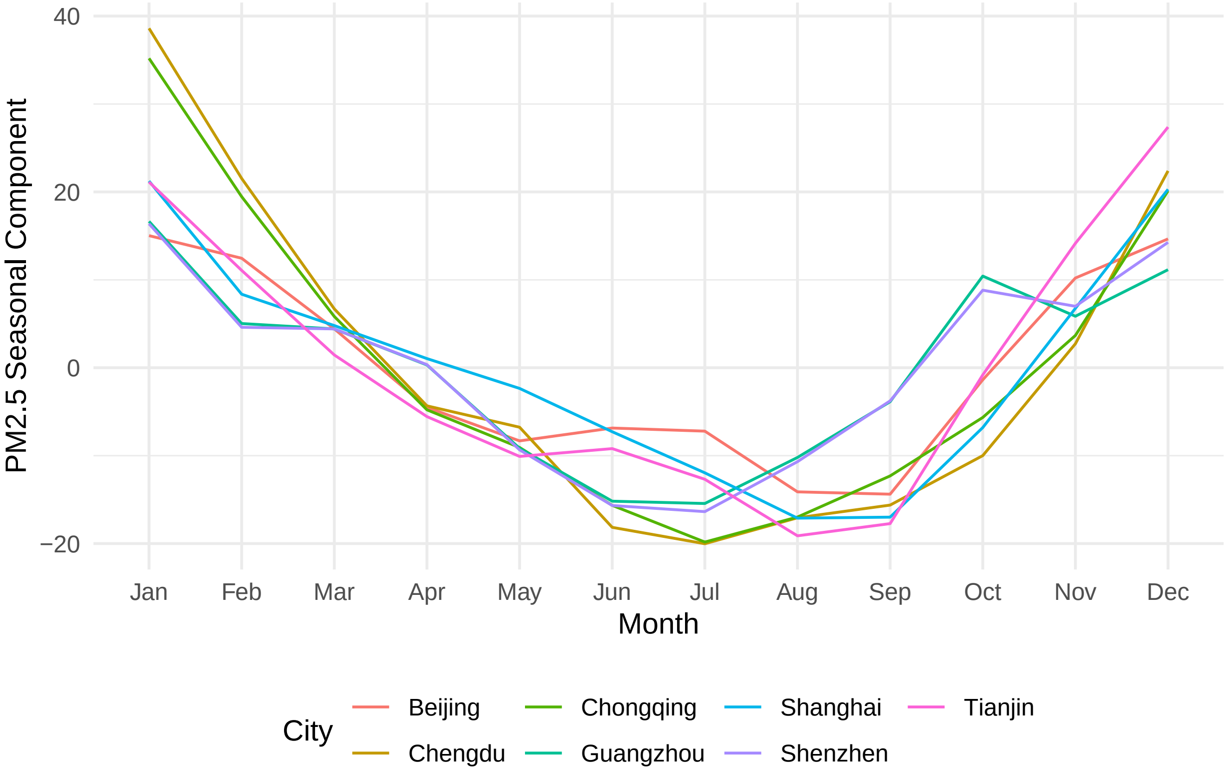
4 Analysis

4.1 Part1: Time Series Analysis

4.1.1 Seven cities monthly PM2.5 (2000~2021)

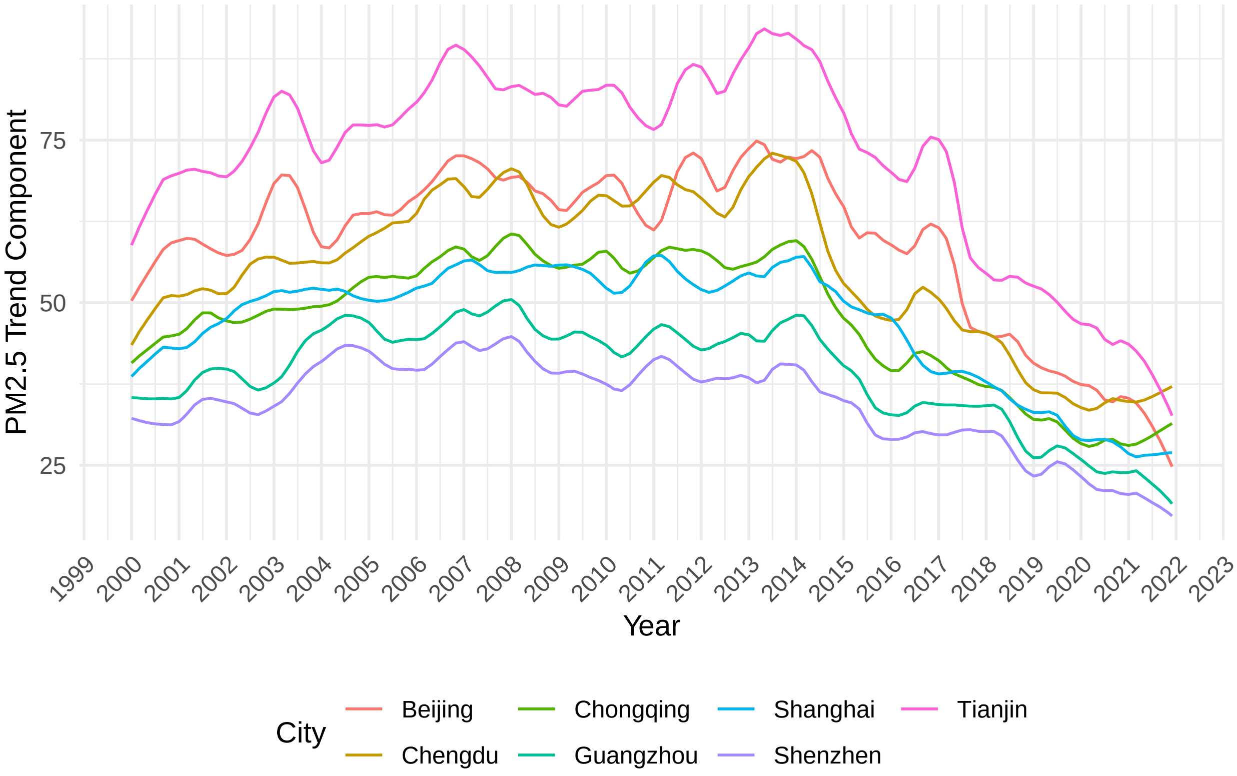
Using the STL function, we decomposed the PM2.5 time series data of the target cities into seasonal and trend components and performed a comparative analysis. The PM2.5 values in these cities are generally lower during the summer months (June, July, August, and September) and higher during the winter 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 PM2.5 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. This leads to increased emissions of coal-related pollutants, causing a rise in PM2.5 levels.

Figure 2. Seasonality in PM2.5 in 7 Chinese Cities



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 PM2.5 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, released in September 2013. This plan was the first time the Chinese government set explicit air quality improvement targets. It required that, by 2017, the annual average PM2.5 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 7 Chinese Cities



4.1.2 Seven cities monthly PM2.5 forecast (2022~2026)

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.
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.

Table 1: 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 2: 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

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 PM2.5 levels in the target cities.

4.2 Question 2:

To better visualize 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 PM2.5 distribution pattern in the country. The “Time Series Visualization by City” panel has a dropdown box that allows users to select the monthly PM2.5 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 PM2.5 prediction in the “Prediction” tab.

5 Summary and Conclusions

6 References

<add references here if relevant, otherwise delete this section>

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.