A Comparative Study on the Prediction Effects of Multiple Time - series Models

Xu Feng, Guzalnur Abdukadir *, Guldana Ramazan, Yutong Fan
School of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, China, 830052
* Corresponding Author Email: Guzalnur212@xjau.edu.cn

Abstract. With the continuous improvement of urbanization, the pollution of NO_2 in urban atmosphere has become increasingly severe. In order to accurately predict the concentration of NO_2 in the atmosphere of Urumqi, this study utilized the SARIMA model, Holt - Winters model, and the intervention model to forecast the NO_2 concentration in the atmosphere of Urumqi from April 2024 to March 2025. Additionally, the fitting performance and prediction accuracy of the three models were compared. Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were selected as the evaluation indicators for model fitting and accuracy. The results showed that for the optimal SARIMA model, the values of MAE, RMSE (Root Mean Squared Error) and MAPE were 5.874, 7.880, and 15.7%, respectively. For the optimal Holt - Winters model, the values of MAE, RMSE, and MAPE were 4.295, 5.460, and 12.0% respectively. And for the intervention model, the values of MAE, RMSE, and MAPE were 4.730, 5.615, and 13.8%. By comparing the prediction results of the three models, it was concluded that the modified optimal Holt - Winters model had the best prediction performance, followed by the intervention model, while the optimal SARIMA model had the worst prediction performance. Finally, the optimal Holt - Winters model was used to predict the NO_2 concentration in the atmosphere of Urumqi in the coming year, providing rational suggestions for the local government's policy - making and residents' travel arrangements.

Keywords: NO_2 concentration, SARIMA intervention model, Holt-Winters model, Prediction accuracy.

1. Introduction

With the rapid development of industrialization in recent years, the phenomenon of air pollution has become increasingly severe [1] . Urumqi, as the capital of Xinjiang, has witnessed rapid economic growth in recent years, generating a large amount of NO_2 gas, which has brought numerous troubles to local residents and the government.

In order to accurately predict the future content of pollutants in the atmosphere, many scholars have conducted relevant research. Chen Peidi et al. used the $ARIMA(0,0,1)(1,1,0)_{12}$ model to model the monthly average concentration of particulate matter PM10 in Urumqi from 2016 to 2022, and found that the monthly average concentration of particulate matter PM10 in Urumqi was high in autumn and winter each year and showed a decreasing trend year by year [2]. Xu Tao et al. used the $ARIMA(0,0,1)(1,1,0)_{12}$ model to model the air pollutant data in Changji from 2018 to 2022, and concluded that six common pollutants all had seasonality. The concentration values of PM10, PM2.5, SO_2, NO_2, and CO reached their peaks in winter, and the concentration of O_3 was the highest in summer [3]. Ouyang Di et al. used the Holt - Winters model to model the data in Harbin from January 2019 to March 2022, and found that the concentrations of four major air pollutants were high in January and low from June to July [4]. In these studies, only the ARIMA model was selected to predict the change in the content of pollutants in the atmosphere, without considering the influence of external variables. Regarding the Holt - winters three - parameter exponential model, its parameters are basically automatically generated based on the software's optimal fitting principle, which has certain limitations.

To address the above problems, this paper adds a policy factor, an external variable, to the ARIMA model to construct an ARIMA intervention model for prediction. For the Holt - winters three -

parameter exponential model, the R traversal loop algorithm is used to screen out the optimal parameter combination model. Finally, the prediction accuracies of various models are compared. Through multi - dimensional modeling, the content of NO_2 in the atmosphere can be predicted from different perspectives, which can greatly improve the prediction accuracy of the NO_2 content in the atmosphere of Urumqi and provide theoretical guidance for the local government's decision - making and residents' travel.

2. Method

2.1. Data source

The monthly data of NO_2 content in the atmosphere of Urumqi from December 2013 to March 2024 were sourced from the China Air Quality Online Monitoring and Analysis Platform. A total of 124 sample data were collected, with a sampling interval of one month. The data from December 2013 to March 2023 were used as the training set, and the data from April 2023 to March 2024 were used as the test set. Differentiate the original sequence data by 1 - order and 12 - step to transform it into a stationary series. Subsequently, build a model for the differenced and stationary sequence.

2.2. Model introduction

The SARIMA model is an extension of the ARIMA model and is suitable for predicting data with strong seasonality. Its parameter representation is $SARIMA(p,d,q)(P,D,Q)_m$, where (p,d,q) represents the non - seasonal part, (P,D,Q) represents the seasonal part, and m represents the periodicity of the time series. The data in this paper are monthly data, so $m=12$ [5].

The intervention model is mainly used to analyze the impact of external interventions on time - series data. Evaluating the influence of external events on the series is called intervention analysis. The intervention model is a special case of the ARIMAX model. The key to the intervention model is to introduce intervention events into the corresponding sequence analysis in the form of dummy variables. According to the mechanism of action, the most common types are step intervention and impulse intervention [6].

$$x_t = \begin{cases} 0, \text{Before the intervention event occurs}(x < T) \\ 1, \text{After the intervention event occurs}(x \geq T) \end{cases} \tag{1}$$

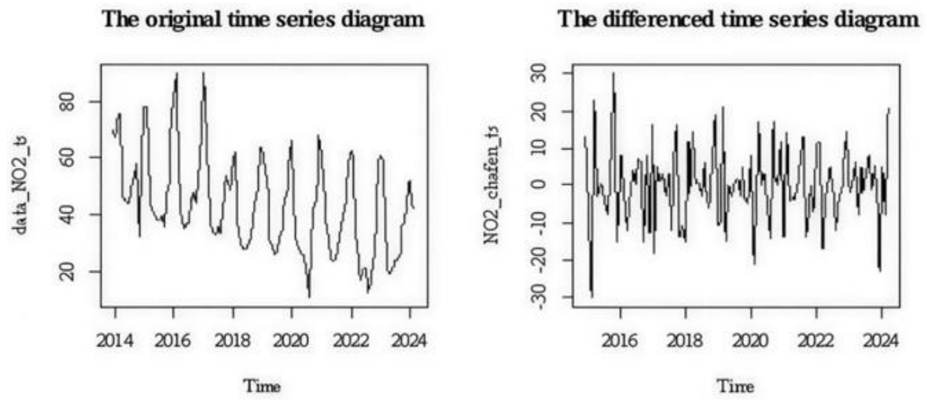
$$x_t = \begin{cases} 0, \text{When the intervention event occurs}(x = T) \\ 1, \text{Other times}(x \neq T) \end{cases} \tag{2}$$

The Holt - winters method is used to analyze time series and is applicable to non - stationary series with a linear trend and a fixed cycle. Its basic idea is to decompose the time series by introducing three smoothing parameters alpha, beta, and gamma to estimate the long - term trend, the increment of the trend, and the seasonal variation respectively, so as to make predictions. Therefore, when determining which model to use, factor decomposition should be carried out first. Usually, the X11 model is used for factor decomposition, which decomposes the factors into three parts: the seasonal index, the trend effect, and the random effect. According to the situations of these three parts, an additive or multiplicative exponential smoothing model is selected for modeling [7].

2.3. Model establishment

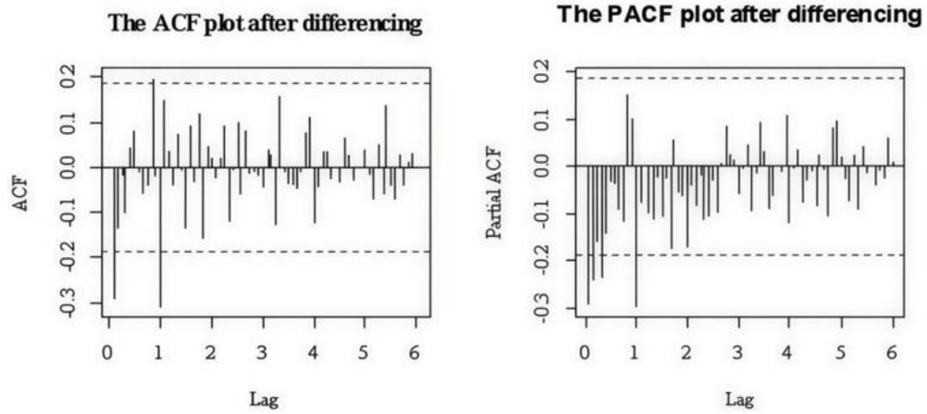
2.3.1. Establishment of SARIMA Model

First, draw the time - series plot of the original data. The data exhibits an obvious decreasing trend and periodicity, showing typical non - stationary characteristics. Both the time - series plot of the first - order 12 - step difference sequence (Figure 1) and the ADF test of the difference sequence indicate that the sequence becomes stationary after differencing. The white - noise test shows that the differenced sequence is a stationary non - white - noise sequence.



(a) NO2 time sequence diagram. (b) Differenced Time Sequence Diagram.

Figure 1. Time series plots of NO2 before and after differencing.



(a) ACF Diagram of NO2. (b) PACF Diagram of NO2.

Figure 2. ACF and PACF plots after differencing.

The autocorrelation plot and partial autocorrelation plot of the differenced sequence (Figure 2) show that only the first - order autocorrelation coefficient is significantly greater than twice the standard deviation, presenting a characteristic of truncation. In the partial autocorrelation plot, only the first, second, and fourth - order partial autocorrelation coefficients are all significantly greater than twice the standard deviation, which can also be considered as a characteristic of fourth - order truncation. We can try $AR(1)$ and $MA(4)$. Considering that a first - order 12 - step difference was performed earlier, if the additive model is used, we can use $ARIMA(0,1,1)(0,1,0)_{12}$ and $ARIMA(4,1,0)(0,1,0)_{12}$. If the multiplicative model is used, by observing the characteristics of the autocorrelation and partial autocorrelation coefficients with the cycle length as the unit, the 12 - step - lagged autocorrelation and partial autocorrelation coefficients are both 0 . By regarding the autocorrelation and partial autocorrelation coefficients as truncated at the cycle respectively, the fitting multiplicative models are obtained as follows.

$$\begin{gathered} ARIMA(0,1,1)(1,1,0)_{12} \text{、} ARIMA(0,1,1)(0,1,1)_{12} \\ ARIMA(4,1,0)(1,1,0)_{12} \text{、} ARIMA(4,1,0)(1,1,0)_{12} \end{gathered} \tag{3}$$

Through the parameter significance test, it can be obtained that the absolute values of all parameter estimates are greater than twice their standard deviations. Therefore, all parameters are significantly non - zero. Through the model test, it can be seen that only the residual sequence of ARIMA (4, 1, 0) (1, 1, 0) is a white - noise sequence, and the residual sequence approximately follows a normal distribution. The optimal model obtained is

$$ARIMA(4,1,0)(1,1,0)_{12} \tag{4}$$

2.3.2. Establishment of SARIMA Intervention Model

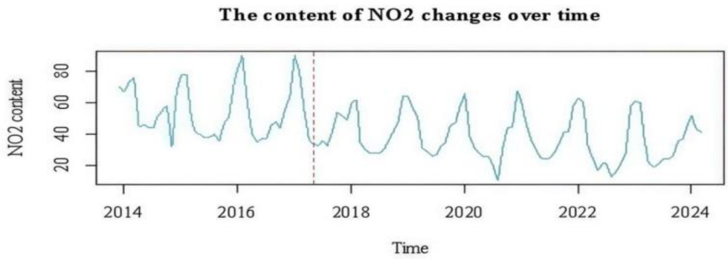


Figure 3. The content of NO2changesr over time.

As can be seen from Figure 3, the sequence experienced a significant decline from 2017 to 2018. Considering the implementation of policies such as "Implementation Plan for the Prevention and Control of Air Pollution in Urumqi during the Winter of 2017 and the Spring of 2018", "Implementation Plan for the 'Blue Sky' Project in Urumqi", and "Implementation Plan for the Joint Prevention and Control of the Environment in Urumqi in 2017", the fluctuation range of the sequence decreased significantly after 2017. This indicates that these policies played a crucial role during this period. Therefore, taking April 2017 as the reference line, a step - type intervention policy variable x_t (policy intervention measure) was constructed for re - analysis.

$$x_t = \begin{cases} 0, t < April, 2017 \\ 1, t \geq April, 2017 \end{cases} \tag{5}$$

To investigate whether there is a significant correlation between the atmospheric NO_2 content in Urumqi and the intervention variable, the cross - correlation diagram of the NO_2 sequence and the intervention variable (Figure 4) was used to study the intervention mechanism of the intervention variable on the sequence. Figure 4 shows that the correlation is the strongest when the intervention variable has a 0 - order lag. It can be assumed that the intervention variable only has a horizontal impact without delay. After performing 12 - step differencing on the NO_2 content sequence in Urumqi, stationarity and white - noise tests were carried out. Then, the intervention variable was introduced, and a regression model was established based on the differenced NO_2 content sequence data.

$$\nabla_{12NO2} = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{6}$$

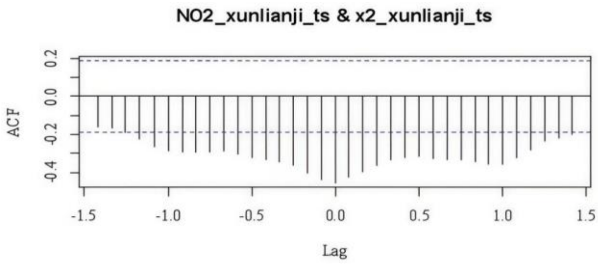
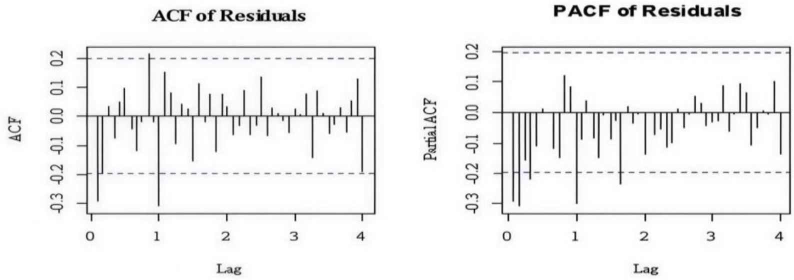


Figure 4. Cross-correlation Diagram of the NO_2 Sequence and the Intervention Variable.



(a) ACF Diagram (b) ACF Diagram

Figure 5. Autocorrelation Diagram and Partial Autocorrelation Diagram of the Regression Residual Sequence.

By examining the characteristics of the autocorrelation coefficients and partial autocorrelation coefficients of the residual sequence (Figure 5), only the first - order autocorrelation coefficient is significantly greater than twice the standard deviation, showing a truncation feature. In the partial autocorrelation plot, only the first, second, and fourth - order partial autocorrelation coefficients are all significantly greater than twice the standard deviation. Therefore, the autocorrelation coefficients and partial autocorrelation coefficients can be regarded as truncated respectively, or both can be regarded as having a trailing feature. The seasonal autocorrelation coefficients of both are significantly non - zero at a lag of 12 orders, and the coefficients at a lag of 24 orders are within twice the standard deviation. So both can be considered as truncated or trailing simultaneously. For the differenced data, all possible intervention models can be established.

$$\begin{aligned} &SARIMA(4,0,0)(1,0,0)_{12}、SARIMA(4,0,0)(0,0,1)_{12}、SARIMA(4,0,0)(1,0,1)_{12} \\ &SARIMA(4,0,1)(1,0,0)_{12}、SARIMA(4,0,1)(0,0,1)_{12}、SARIMA(4,0,1)(1,0,1)_{12} \\ &SARIMA(0,0,1)(1,0,0)_{12}、SARIMA(0,0,1)(0,0,1)_{12}、SARIMA(0,0,1)(1,0,1)_{12} \end{aligned} \tag{7}$$

After parameter testing, the models that passed the test are only $SARIMA(0,0,1)(1,0,0)_{12}$ and $SARIMA(0,0,1)(0,0,1)_{12}$. Finally, according to the principle of minimizing AIC and BIC, the optimal intervention model is...

$$SARIMA(0,0,1)(0,0,1)_{12} \tag{8}$$

2.3.3. Establishment of Holt - Winters Model

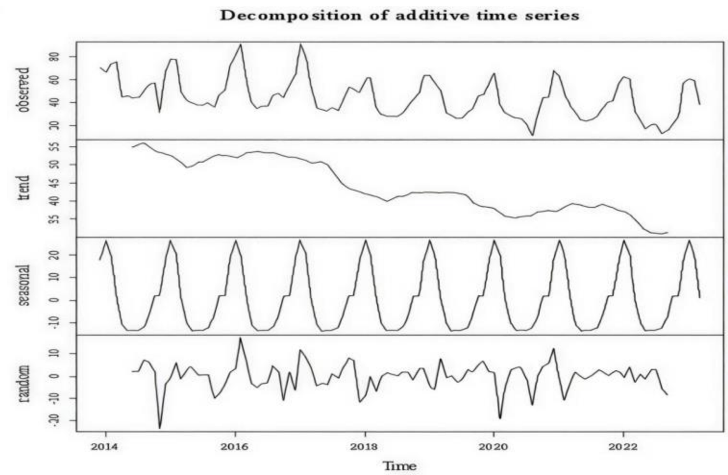


Figure 6. Decomposition Diagram of NO_2 Time Series Data.

From the decomposition graph of the NO_2 sequence (Figure 6), it can be seen that the NO_2 sequence has obvious periodicity and trend. Therefore, the three - parameter exponential smoothing model is selected for modeling. However, it is difficult to determine whether the multiplicative model or the additive model has a better prediction effect. In order to find the optimal fitting model, the additive and multiplicative models are used for modeling respectively. The optimal prediction model is selected by comparing the RMSE, MAE, and MAPE values.

Using R software, the parameter combinations are automatically selected based on the principle of optimal fitting as follows:

$$alpha = 0.1974584, beta = 0.01280326, gamma = 0.613219 \tag{9}$$

To quantitatively compare the prediction performance of these two models, the $RMSE$, MAE and $MAPE$ values of both are calculated respectively. The main reason for the relatively large error is the unreasonable selection of the three parameters. To find more suitable parameter combinations, the loop traversal method in R language is used to search for parameters. The three parameters, alpha, beta, and gamma, are traversed and take values from 0.1 too 0.9 respectively. Through the R loop traversal, the optimal parameter combinations of the additive and multiplicative models are obtained as follows.

$$\begin{gathered} alpha = 0.2, beta = 0.1, gamma = 0.2 \\ alpha = 0.1, beta = 0.1, gamma = 0.2 \end{gathered} \tag{10}$$

By calculating the $RMSE$, MAE and $MAPE$ values of these two models, it is found that the $RMSE$, MAE and $MAPE$ values of the models obtained through R loop traversal are significantly lower than those of the models screened by the optimal fitting principle in R. Therefore, this method improves the prediction accuracy of the models. The MAE, $RMSE$ and $MAPE$ values of the modified additive model are all smaller than those of the multiplicative model. Thus, the prediction performance of the additive model is better, and the modified additive model has a higher prediction accuracy. The $MAPE$ value of the modified additive model is 12.0%. The main reason for the remaining error is that the selection of the step size is not precise enough. Therefore, to further improve the prediction accuracy of the model, the step size can be reduced, but this will increase the computational load. When the required prediction accuracy is not extremely high, this step size can be used to select parameters. In addition, the future content of NO_2 in the atmosphere is also affected by other factors (such as coal consumption and seasonal wind speed), which is also the cause of prediction errors in the model. Subsequently, the model can be continuously adjusted according to the required prediction accuracy to obtain the desired prediction results.

3. Comparison of the Precision of Three Models.

To compare the prediction performance of the selected models, the $RMSE$, MAE and $MAPE$ values of the three selected models were obtained (Table 1). According to the principle of minimizing the MAE, RMSE, and MAPE values, the model with the best performance is the additive model with $alpha = 0.2, beta = 0.1, gamma = 0.2$. To visualize the prediction performance of these three models, a comparison chart of the fitting effects of the three models was drawn simultaneously in Figure 7.

Table 1. $RMSE$, MAE and $MAPE$ Values of Holt - Winters Model.

	MAE	$RMSE$	$MAPE$
Holt - Winters Additive Model Based on R's Optimal Fitting Principle	6.665	7.709	0.201
Holt - Winters Multiplicative Model based on R's Optimal Fitting Principle	6.222	7.546	0.179
Additive model with $alpha = 0.2, beta = 0.1, gamma = 0.2$	4.295	5.460	0.120
Multiplicative Model with $alpha = 0.1, beta = 0.1, gamma = 0.2$	4.943	5.888	0.149

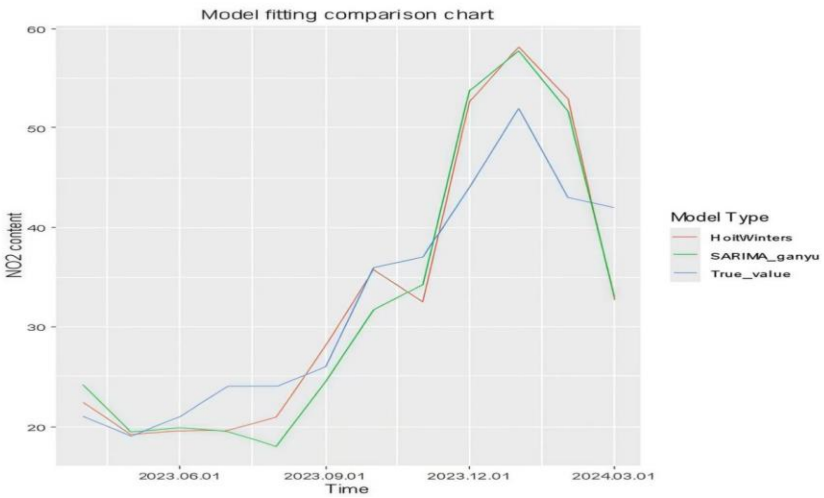


Figure 7. Comparison of Fitting Effects of Three Models in the Training Set.

By observing Figure 7, it can be seen that in the coming year, the predicted values of the three models do not differ significantly from the actual values. The maximum values predicted by these three models occur in December, January, and February (winter), and the minimum values are around June, July, and August (summer), which is consistent with the actual situation. All three models can reflect this pattern of change. In order to accurately predict the NO_2 content in the atmosphere of Urumqi, based on the principle of minimizing the $RMSE$, MAE, and $MAPE$ values, the modified Holt - Winters additive model is finally selected to predict the NO_2 content in the atmosphere of Urumqi from April 2024 to March 2025.

4. Future Outlook and Suggestions

Prediction of NO_2 Content (unit: ug/m^3) in the Atmosphere of Urumqi from April 2024 to March 2025 by the Modified Holt - Winters Additive Model.

Table 2. Predicted Atmospheric NO_2 Values in Urumqi from April 2023 to March 2024.

Time	Predicted value	Time	Predicted value	Time	Predicted value
April,2024	20.76491	August,2024	19.42012	December,2024	48.74563
May,2024	17.66982	September,2024	25.61750	January,2025	54.96732
June,2024	18.24219	October,2024	33.62518	February,2025	49.27446
July,2024	18.57875	November,2024	34.99283	March,2025	32.21258

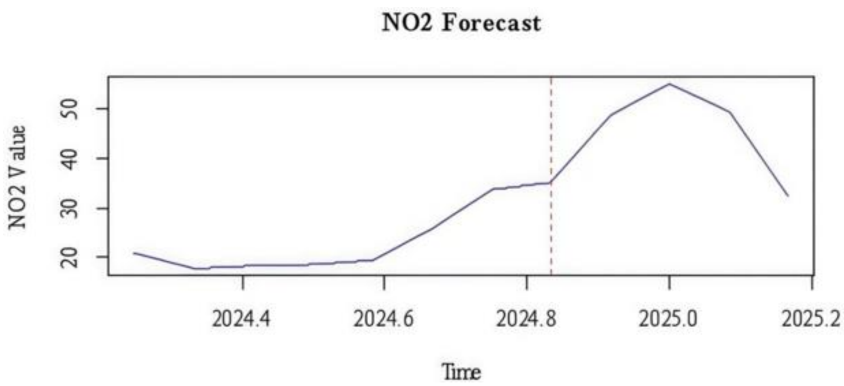


Figure 8. Predicted values of atmospheric NO2 content in Urumqi.

As can be seen from Table 2 and Figure 8, the NO_2 content in the atmosphere of Urumqi is relatively high in winter and relatively low in summer. The red dotted line in Figure 8 represents October 2024, and it can be directly observed that the NO_2 content in the atmosphere from October

2024 to March 2025 is significantly higher than that in other months. There are two main reasons for this. Firstly, the heating period in Urumqi is from October 10th to April 10th of the following year. During the heating period, a large amount of coal is burned, generating a large amount of NO_2 waste gas. The discharge of these NO_2 waste gases into the air leads to a sharp increase in the NO_2 content in the atmosphere [8]. Secondly, it is related to the terrain of Urumqi. Urumqi is surrounded by mountains on three sides, and there is less wind in winter, resulting in poor air circulation, which makes it easier for pollutants to accumulate, leading to a relatively high NO_2 content in Urumqi in winter [9]. The NO_2 content is lower in summer. The reasons are as follows. Firstly, the temperature in Urumqi is relatively high in summer, and heating stops, reducing the emission of NO_2. Secondly, the wind is relatively strong in summer in Urumqi, which is not conducive to the accumulation of NO_2. Thirdly, especially in July and August, the weather is hot and residents often suffer from heatstroke. Therefore, many residents with cars are reluctant to travel, which also avoids the emission of NO_2 waste gas. In view of this phenomenon, three suggestions are provided for the local government and residents: Since the NO_2 content in the atmosphere is relatively high in winter, residents should wear masks when traveling in winter, and try to travel in an environmentally - friendly way. High - efficiency heating equipment can be installed at home. The government should strengthen the management of winter heating equipment, promote natural gas heating or electric heating equipment, and gradually phase out highly - polluting coal - burning stoves. Moreover, through economic or tax preferential policies, it should encourage the introduction of low - carbon production technologies. In summer, although the NO_2 content in the atmosphere is relatively low, vigilance cannot be relaxed. The government should strengthen the supervision of industrial emissions, and residents should use high - energy - consuming electrical appliances less frequently.

5. Conclusion

In terms of air quality prediction, the SARIMA intervention model with external variables outperforms the traditional SARIMA model. Regarding the use of the Holt - winters model, the model with the optimal parameter combination obtained through manual iterative loop is superior to the one with the optimal parameter combination automatically selected by software. Air quality prediction is a very complex issue. It is difficult to achieve good prediction results by using only a single prediction model. In this paper, the Mean Absolute Percentage Error (MAPE) of the SARIMA intervention model incorporating policy factors is 1.9% lower than that of the single SARIMA model. The MAPE of the Holt - winters model obtained through an R traversal loop is 8.1% lower than that of the model selected according to the optimal fitting principle in R. These two models have significantly improved the prediction accuracy. Therefore, using multiple time - series prediction models for prediction and selecting the optimal prediction model from them is an important method to improve prediction accuracy.

Acknowledgements

The authors gratefully acknowledge the financial support Innovation Projects for College Students of Xinjiang Agricultural University (at the autonomous region level) s202410758070 funds.

References

[1] Shao Yingying, Wang Chen, Cui Anfeng, et al. A Time - Series Study on the Short - Term Impact of Atmospheric Fine Particulate Matter PM2.5 on the Lung Function of Elderly Patients with Chronic Obstructive Pulmonary Disease in Taiyuan [J]. Public Health and Preventive Medicine, 2025, 36(01): 18 - 22.

[2] Chen Peidi, Zhou Mingzhang, Xiao Tingting, et al. Analysis of Pollution Characteristics and Trends of Particulate Matter PM10 in Urumqi Based on the ARIMA Model [J]. Occupation and Health, 2024, 40(15): 2086 - 2090.

[3] Mulati Ainiwaer, Xu Tao, Xiaokaiti Yibulayin. Air Quality and Prediction in Changji City from 2018 to 2022 [J]. Occupation and Health, 2024, 40(10): 1370 - 1375.

[4] Ouyang Di, Cui Jia. Spatiotemporal Dynamics of Air Pollutants in Harbin under the Carbon Neutrality Strategy - Based on the Holt - Winters Time - Series Model [J]. Environmental Protection Science, 2023, 49(02): 112 - 119.

[5] Ma Fengbin. AQI Forecasting Model of Qingdao Based on MLP and SARIMA [J]. Science - Technology Innovation and Productivity, 2023, (01): 62 - 67.

[6] Na Yanping. Research on the Impact of Policy Intervention on Air Quality in Yinchuan [J]. Mathematics in Practice and Theory, 2020, 50(09): 243 - 249.

[7] Zhu Jie, Bi Dong. Analysis of the Development Trend of Rural Online Retail Sales Based on the Holt - Winters Model [J]. China Journal of Commerce, 2024, 33(16): 9 - 12.

[8] Wang Sirui. Atmospheric Environmental Conditions and Evolution Trends of Heavy Pollution in Urumqi from 2018 to 2022 [J]. Environmental Protection of Xinjiang, 2024, 46(04): 30 - 35.

[9] Xia Rongxiang, Adili Simaayi, Lin Qin, et al. Impact of Air Pollution on the Lung Function of Primary School Students in Urumqi [J]. Occupation and Health, 2024, 40(16): 2260 - 2264 + 2270.