

Time Series Analysis for Air Pollution Data of Beijing

Jiasheng Li[#], Huanshuo Liu^{*, #}

School of Mathematics, Sun Yat-Sen University, Guangdong, China, 510275

*Corresponding author: liuhsh35@mail2.sysu.edu.cn

[#]These authors contributed equally.

Abstract. For a long time, air pollution has been a serious problem encountered Beijing due to the city's rapid economic development, high level social production and living standard, and unfavorable climate. This paper will use hour-by-hour data for six air pollutants and eight weather variables collected from January 1, 2014, to December 31, 2015, at an air pollution monitoring site in Beijing. We mainly analyze the data on SO₂ pollutants with the ARIMA model which is used to make predictions on SO₂ data. Since the periodicity of hour-by-hour data is too long and fluctuates drastically, using the ARIMA model will face problems such as lacking valid forecasting time and significant forecast bias. Besides, we have data on many climate indicators, including temperature, moderation, and concentration of many pollutants, which may be correlated. Therefore, with proper treatment, variable correlation analysis can assist us in predicting specific indicators. We use two different time series decomposition methods in the SO₂ variable. The relationships between the periodic terms and air temperature and humidity were compared separately to verify the conclusions of some previous papers. Moreover, SCINet was used to predict and get excellent results.

Key Words: SCINet, air pollution, SO₂, ARIMA, Forecasting, Correlation Study, Robust-STL, STL with Loess

1. Introduction

As a classic time series model, ARIMA is usually used to make Time Series Forecasting as a baseline. However, since ARIMA is only a univariate forecasting model, it is less effective in complex time series. For example, ARIMA cannot utilize the correlation between data to make predictions, so it performs poorly in the data set of this paper, which is shown in 4.2.2. With the development of deep learning, novel methods continue to emerge from analyzing time series. Therefore we introduce novel methods to make more precise predictions. Then, we outline the relevant applications of time series models in the direction of air pollution. We will briefly describe the model chosen for this paper and its rationale. In Chapter 4, we present the experimental results of these methods. In Chapter 5, we summarize the results in our research and give a conclusion: SCINet can give a better prediction than ARIMA, which means it may be an effective model in our dataset.

2. Method

2.1. ARIMA

In statistics and econometrics, particularly in time series analysis, autoregressive integrated moving average (ARIMA) model is a generalization of autoregressive moving average (ARMA) model [3]. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is a linear combination of error terms whose values occurred contemporaneously and at various times in the past.

2.2. SCINet

SCINet proposed by Liu et al. 2021a is a binary tree-based deep learning model. It consists of 3 levels, which are SCI-Block, SCINet, and Stacked SCINet [1], enhancing the predictability of the original time series by capturing temporal dependencies at multiple temporal resolutions Mallat 1989[2]. The basic block, SCI-Block, decomposes the input data into odd sequences and even

sequences and then processes them using different convolution filters to extract homogeneous and heterogeneous information from each part. SCINet arranges SCI-Block with binary-tree structure and then realigns the child series. Moreover, it concatenates all the low-resolution components into a new sequence representation and adds it to the original time series for forecasting. A stacked SCINet is composed of multiple SCINets with probable intermediate supervision. SCI-Block is shown as Figure1. SCINet is shown as Figure2.

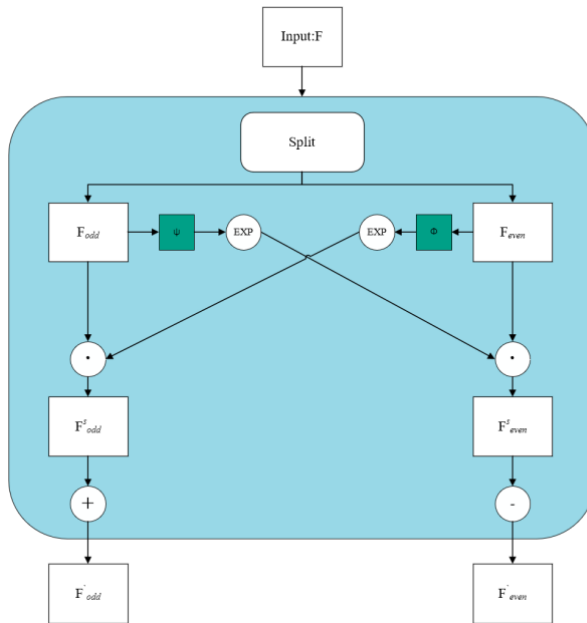


Figure 1. SCI-Block

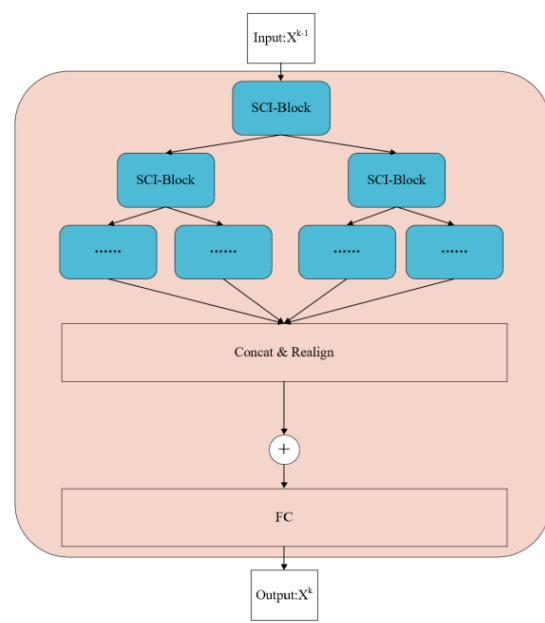


Figure 2. SCINet

SCI-Block (Figure1) is based on operations *Splitting* and *Interactive-Learning*. The Splitting procedure down samples the original sequence $\mathbf{F}$ into two sub-sequences $\mathbf{F}_{even}$ and $\mathbf{F}_{odd}$ by the different indexes of the elements [3]. Then different convolutional kernels are used to extract information from the two sequences. Then, *Interactive Learning* consisting of two steps is used to compensate for the loss of information. Stacked SCINet is shown as figure3.

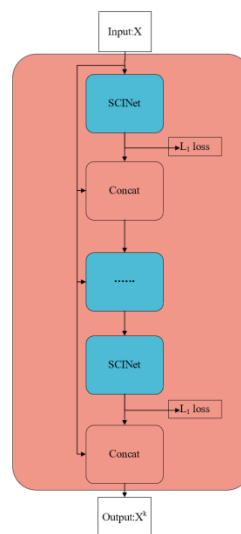


Figure 3. Stacked SCINet

Compared to the dilated convolutions used in the TCN architecture, the proposed downsample-convolve- interact architecture achieves an even larger receptive field. More importantly, unlike TCN, that employs a single shared convolutional filter in each layer, significantly restricting its feature extraction capabilities, SCI-Block aggregates essential information extracted from the two down sampled sub-sequences according to the given supervision for prediction.

With the SCI-Blocks presented above, we construct the SCINet by arranging multiple SCI-Blocks hierarchically and get a tree-structured framework shown in figure2.

When there are sufficient training samples, we could stack K layers of SCINets to achieve better prediction accuracy at the cost of a more complex model structure, as shown in figure1. Moreover, we apply intermediate supervision proposed by Bai et al. 2018b.[11] We concatenate the output of the k -th intermediate SCINet, X_k with length τ and part of the input $X_{t-1(T-\tau)+1:t}$ [3].

3. Experiments

In this section, we show the quantitative and qualitative analysis results of *Air Pollution data of Beijing* dataset. We predict the time series data at both daily and hourly scales and compare the effects between different models.

3.1. Exploratory Data Analysis

The data set consists of six air pollutants and eight weather variables collected at an air pollution monitoring site in Beijing from January 1, 2014, to December 31, 2015. It includes hourly data of PM2.5, PM10, SO2, O3, NO2, and CO and hourly data of wind speed, wind direction, rainfall, relative humidity, temperature, atmospheric pressure, and dew point temperature. The data set contains 17,520 hours which is treated as 17,520 time points. When we make daily predictions, we ignore all censored values and average the remaining data. On the other hand, if we make hourly predictions, the effect of missing values will be retained. On the other hand, if we make hourly predictions, regardless of the model used, we will simply use forward filling to avoid data snooping.

Meteorological conditions strongly influence the dispersion and dilution of SO2 pollutants. Research in Elazığ, Turkey, found that SO2 concentrations are strongly related to cold temperatures and high relative humidity Akpınar et al. 2008[4]. Another study of air pollution using Mongolian meteorological data concluded that SO2 concentrations increased with the decrease in temperature. Moreover, SO2 concentrations increased with the increase of relative humidity Luvsan et al. 2012 [5]. We wish to verify these conclusions using meteorological data from Beijing. We used this method to obtain the results about the time series decomposition of SO2 concentration during these 240 hours. Also, the periodic curves were used to compare with the temperature and humidity curves of the same period. The comparison of the images shows that the peaks and valleys of the SO2 curve and the temperature curve is staggered each other, while the peaks and valleys of the SO2 curve and the humidity curve overlap each other. It can be corroborated that higher SO2 levels depend on a lower temperature and higher humidity.

In the decomposition of SO2, it is evident that Robust-STL performs better than STL with Loess in tracking the trend term of the data and the seasonal term to the temperature bias[6-7]. In the seasonal term, it is evident that the amplitude of the vibration becomes smaller and shifts downward, corresponding to the similar changes in temperature at that time. Also, comparing the seasonal terms obtained by this method, we plot SO2 and the relative humidity curve. We still obtain the same conclusions as when using STL with Loess, which can corroborate our initial point. Therefore, we conclude that during the propagation of SO2 pollutants, the SO2 concentration tends to increase as the temperature decreases and the humidity increases, and vice versa. Humidity and temperature period plot is shown as figure4.

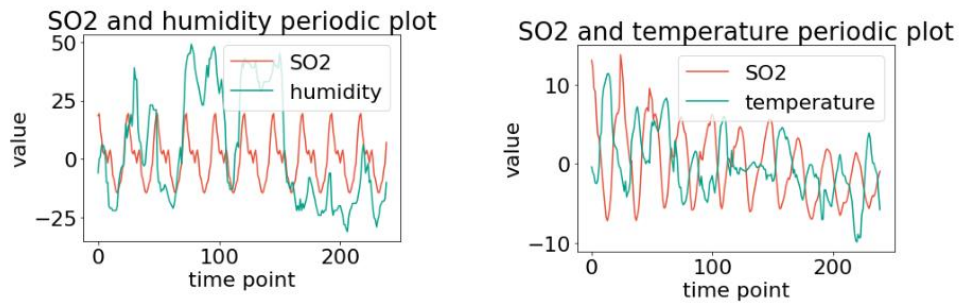


Figure 4. humidity and temperature period plot

In figure4, the left plot shows the curve of SO2 and temperature, and the right plot shows the curve of SO2 and relative humidity. Note: the relative humidity is in the range of 0 % to 100 %. To put the two curves together to compare their peak-to-valley distribution better, the humidity curve has been shifted down by 40 units. Trend and observations are shown as figure5. The trending term generated by STL with Loess compares with the SO2 data, Moreover, the trend term generated by Robust-STL compares with the SO2 data.

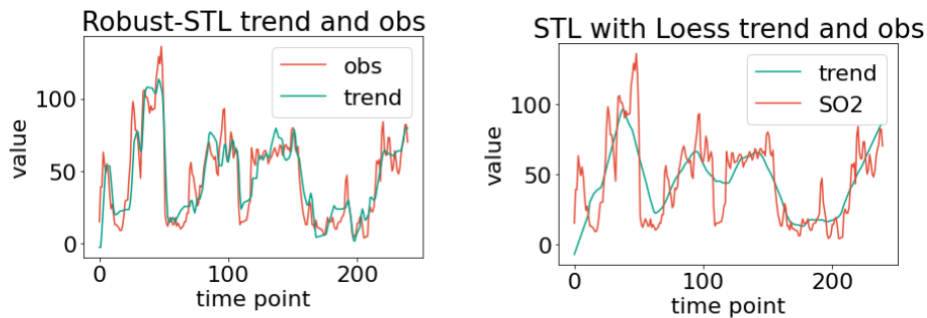


Figure 5. trend and observation

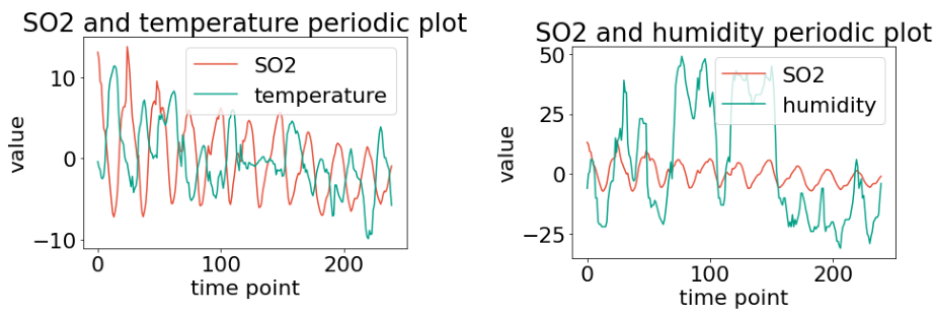


Figure 6. temperature and humidity period plot

Temperature and humidity are shown in figure6. The left plot shows the curve of SO2 and temperature, and the right plot shows the curve of SO2 and relative humidity. Note: the relative humidity is in the range of 0 % to 100 %. To put the two curves together to compare their peak-to-valley distribution better, the humidity curve was shifted down by 40 units.

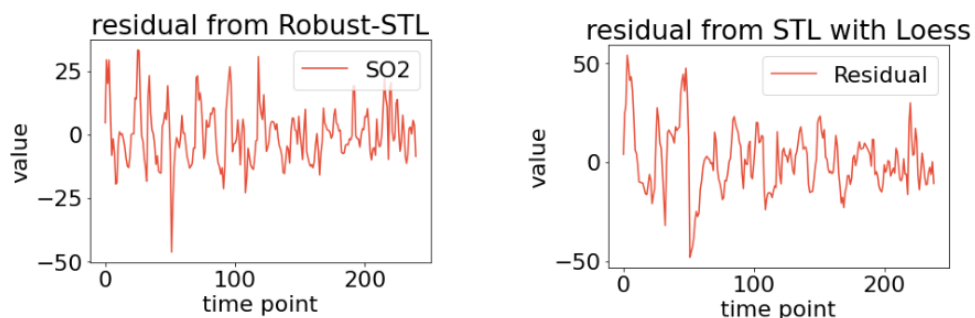


Figure 7. residual plots of Robust-STL and STL with Loess

In figure7, the residual term generated with STL with Loess compares with the SO2 data, and the residual term generated with Robust-STL.

3.2. Correlation of Dataset

This paper focuses on the correlation of SO2 series with temperature and relative humidity by time series decomposition. Over a shorter period (10 periods in all), a more stable period term can be decomposed and compared with the temperature and relative humidity images of the same period to visualize the correlation. Through these images, the peaks and valleys of the SO2 period series and the peaks and valleys of the temperature series intersect and overlap more with the peaks and valleys of the relative humidity. This leads to the conclusion that SO2 negatively correlates with temperature and positively correlates with relative humidity.

3.3. ARIMA

ARIMA model is a univariate forecasting model, so we can only use the information of variable SO2 to forecast its trend.

We first divide the training and testing sets from the original data. The ARIMA model has three coefficient terms, p , d , and q , that need to be determined artificially. We first verify the stationarity of the training set and find that the data are stationary at 99.9% confidence level. Then, we calculate the AIC and BIC of the models with different p and q , respectively, to find the models with the smallest AIC and BIC and record their parameters as our final choice. For hourly data, we select 2 groups of coefficients: (2, 0, 3), (1, 1, 3). After that, we will use the model fitted by the training set to make a 24-hour forecast for the testing set and calculate the Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Mean Squared Percentage Error (MSPE).

Then, we average the hourly data to obtain the daily data and perform the above process on the daily data to give the ARIMA model forecast results in two groups of coefficients: (2, 1, 4), (7, 1, 7), which perform best according to AIC and BIC

On the hourly scale, the ARIMA model is valid for one-step or two-step forecasts due to the long training set. However, ARIMA model is limited in long term series forecasting because its output will converge in several steps.

On the daily scale, convergence mentioned above still exists. Besides, noticing SO2 has a strong seasonal correlation while it is correlated with other variables. As a result, ARIMA fails to give valid results.

3.4. SCINet

Since ARIMA, which cannot utilize information from other series, is less effective in forecasting daily scale data and fails to forecast hourly scale data, we need to seek a model that is able to do multivariate time-series forecasting. Based on thorough research, we choose SCINet as the final model.

We use dataset in 3.1 to evaluate the performance of short-sequence TSF tasks, which are conducted on multivariate time-series forecasting at *hourly* scale. For a fair comparison, we keep all input lengths T the same as 4.2.2. One should be noticed that when we use ARIMA to predict the SO2 variable at a certain time, all data before this time are used to fit this ARIMA model. SCINet uses the data in the former 48 hours to forecast the data in the following 24 hours. If the ARIMA model is fitted with only the first 80% of the data, the model's rolling prediction values will converge rapidly. As a result, the prediction will be very poor.

The results in the test set are shown in Table 4.2.3. The MAE, MSE, MAPE, MSPE of the ARIMA model are calculated by averaging the results from all ARIMA models at each t moment, which was fitted with $x_1, \dots, x_{t-1}$.

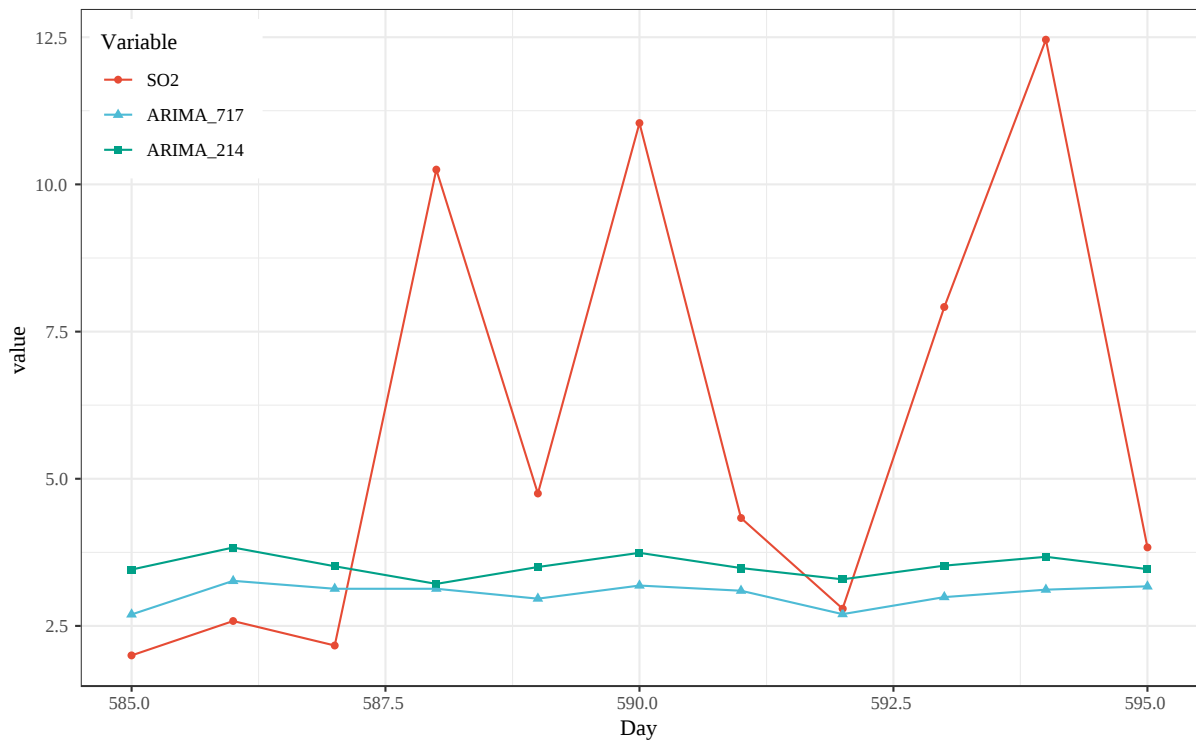


Figure 8. ARIMA using daily SO2 variable

It is worth noticing that the value of the SO2 indicator stays at 2 for a long time when the pollution is not severe. Given this feature of the dataset and the convergence of ARIMA, ARIMA model can obtain better MAE and MSE for hourly scale forecasts but fails to predict fluctuations in the data. The high error of ARIMA at MAPE and MASE illustrate this situation. The prediction results of ARIMA model are shown in Figure8 [8].

As can be seen from Table 1, comparing with the ARIMA model, SCINet is better at capturing the long-term patterns in the actual historical data for predicting the future, leading to lower prediction errors. Figure7 and Figure8 presents the qualitative results on some randomly selected sequences, which demonstrate the capability of SCINet in obtaining the trend and seasonality of time series for TSF [9].

Table 1. Results Comparison

Methods	MAE	MSE	MAPE	MSPE
ARIMA (2, 0, 3)	6.323228	96.744547	1.007135	2.786087
ARIMA (1, 1, 3)	6.0415143	96.8701061	0.8311094	2.1586272
SCINet	6.1667885	114.9615836	0.5962400	0.7452795

We attribute the significant performance improvements of SCINet on the dataset to this reason: SCINet effectively captures temporal dependencies from multiple temporal resolutions while ARIMA is limited to a nearly fixed lookback window which focuses on partial information. Two examples of comparison between ARIMA and SCINet are shown in figure9 and figure 10.

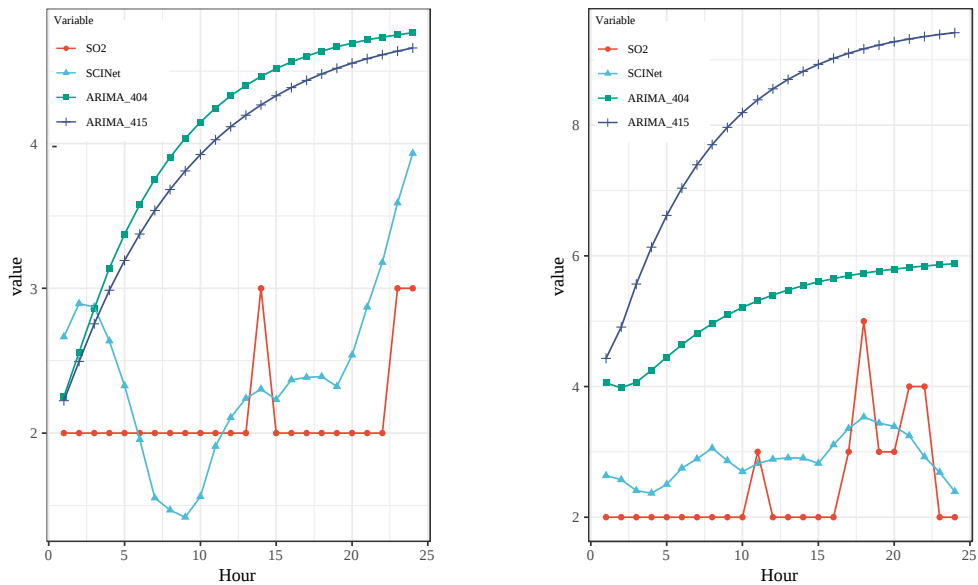


Figure 9. Comparison between ARIMA and SCINet (1)

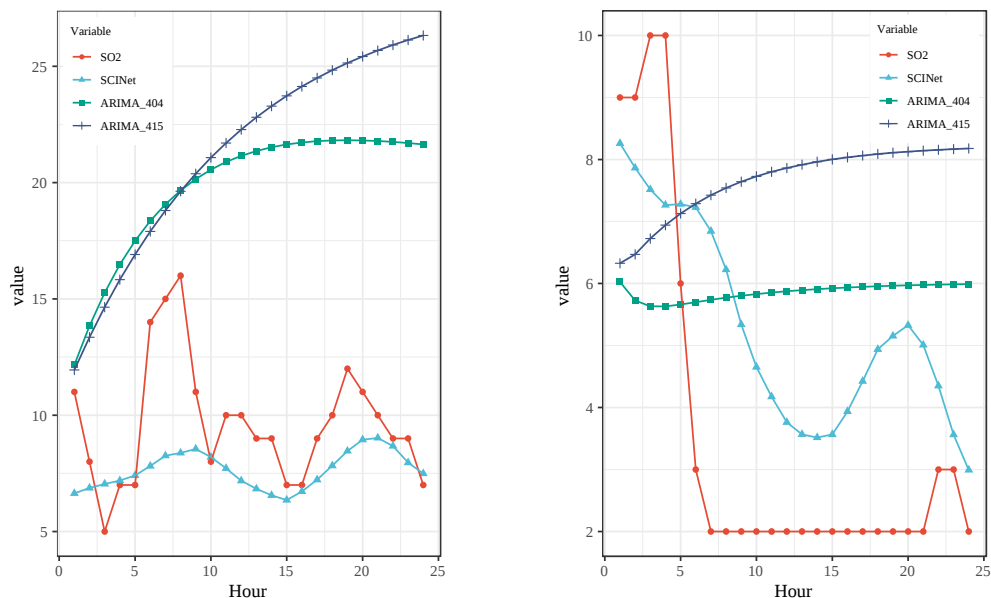


Figure 10. Comparison between ARIMA and SCINet (2)

Hyperparameters of SCINet in 4.2.3 is shown in Table2.

Table 2. Hyperparameters of SCINet in 4.2.3

Hyper Parameters	Value
Learning Rate	0.0007810976188611802
Dropout	0.837065169749385
Levels	2
Stacks	1
Loss	mape

4. Conclusion

First, we compare and verify the correlation between SO2 series and temperature as well as the correlation between SO2 series and humidity. Second, we use the ARIMA model to make a prediction of SO2 and summarize the shortcomings. Third, we predict SO2 with SCINet and improve the shortcomings of ARIMA. To achieve the first goal, we refer to several similar papers written before.

Thus, an initial conclusion is obtained and what we need to accomplish is to verify these conclusions. During this process, we use two-time series decomposition methods to treat the SO₂ series and compare the differences between these two methods [10]. Finally, we verify that SO₂ is negatively correlated with temperature, and SO₂ is positively correlated with related humidity. The second goal is the basis for the third goal. We treat AIC as a rule for filtering ARIMA parameters. Then we use the best model to predict the SO₂ series. However, even the best ARIMA models fail to give a good prediction of hour-by-hour data. Besides, the predicted value will converge after some time, making it rapidly lose its predictive significance. To improve these shortcomings, which is the third goal, we bring in SCINet. SCINet uses more information which leads to a more accurate, flexible, and closer prediction than ARIMA.

References

- [1] Liu M, Zeng A, Xu Z, et al. Time series is a special sequence: Forecasting with sample convolution and interaction[J]. arXiv preprint arXiv:2106.09305, 2021.
- [2] Mallat S G. A theory for multiresolution signal decomposition: the wavelet representation[J]. IEEE transactions on pattern analysis and machine intelligence, 1989, 11(7): 674-693.
- [3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [4] Akpınar S, Oztıp H F, Kavak Akpınar E. Evaluation of relationship between meteorological parameters and air pollutant concentrations during winter season in Elazığ, Turkey[J]. Environmental monitoring and assessment, 2008, 146(1): 211-224.
- [5] Luvsan M E, Shie R H, Purevdorj T, et al. The influence of emission sources and meteorological conditions on SO₂ pollution in Mongolia[J]. Atmospheric Environment, 2012, 61: 542-549.
- [6] Wen Q, Gao J, Song X, et al. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 5409-5416.
- [7] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A seasonal-trend decomposition[J]. J. Off. Stat, 1990, 6(1): 3-73.
- [8] Fan Feng, Xuebin Chi, Zifa Wang, Jie Li, Jinrong Jiang, Wenyi Yang. A nonnegativity preserved efficient chemical solver applied to the air pollution forecast[J]. Applied Mathematics and Computation, 2017, 314.
- [9] Sudhir Nigam, Rashmi Nigam. Performance Evaluation of Air Pollution Forecast Models[J]. Current Environmental Engineering, 2017, 4(3).
- [10] Lidia Contreras, Cèsar Ferri. Wind-sensitive Interpolation of Urban Air Pollution Forecasts[J]. Procedia Computer Science, 2016, 80.
- [11] Bai, S.; Kolter, J. Z.; and Koltun, V. 2018b. Trellis Networks for Sequence Modeling. In ICLR.