Dean&Francis

ISSN 2959-6157

Research on the Weather Forecasting using ARIMA and SARIMA Model

Yichen Yang[1, *]

[1]Department of Mathematics and Physics, Xi'an Jiaotong Liverpool University, Suzhou, 215123, China

*Corresponding author: Yichen.Yang22@student.xjtlu.edu.cn

Abstract:

Weather predicting is important nowadays as it can not only significantly diminish the uncertainty regarding the future, but also provide valuable information for people to make decisions in advance in different areas. ARIMA, SARIMA and SARIMAX models are typical predicting models, which are favored and used by scholars from various nations. This essay aims to compare the predicting performance of them through data processing, parameters selecting, performance measuring and so on. The conclusion is as follows: the seasonal factors incorporated by SARIMA model can greatly improve the accuracy of prediction with smaller errors compared with ARIMA model. Besides, when considering the exogenous variables like longitude and latitude, the predicting performance of models can be enhanced as well. These findings can provide good suggestions for further research. Therefore, research in the future is recommended to exploit the full potential of SARIMAX model in other areas with more exogenous variables and also can attempt to find the best method for selecting exogenous variables.

Keywords: Weather prediction; ARIMA model; SARIMA model; SARIMAX model.

1. Introduction

Forecasting is the process that make forecasts or projections about the events or conditions in the future based on the historical and current data [1]. Since the industrial revolution, there has been a sharp increase in average temperatures, reflecting the impact of increased industrial activity on global warming. In light of the significance in the current context of climate change (global warming), there has been a growing focus on the prediction of seasonal and subseasonal climate patterns [2]. Therefore, weather or temperature prediction remained to be one of the most significant scientific and technological challenges globally during the last century [3].

Weather predicting is important as it can function as a useful tool to diminish the uncertainty regarding the future and proffer reliable information for people which can guide strategic planning and decision-making [3]. Also, weather forecasting can be beneficial to various fields. Firstly, weather prediction can be valuable for agriculture. By forecasting weather patterns, farmers can make well-informed decisions about when to plant crops and receive recommendations about how to secure their crops and resources. Weather prediction can also support the energy industry through providing meaningful information for planning energy demand and supply. Some of the uses of weather predicting including forecasting short-term power demand for power utilities and developing air conditioning and solar energy systems [4].

The main goal of analyzing the relationship between multiple variables and the target variable (temperature) is to enhance the accuracy and validity of temperature prediction. In the research conducted by numerous scholars during the past years, various models have been applied to analyze weather temperature. Machine Learning provides both unsupervised and supervised learning methods for predicting weather with a small error [5]. Dadhich et al. used linear regression, decision tress (DT) and other regression models to forecast weather [5]. However, linear regression models are suitable, but they may not be effective in indicating 'non-linear' relationships [6]. When predicting weather temperature, the relationship between temperature and other variables may be non-linear. If regression models are used, they may not fit this relationship well. Lee and Cho used Radial Basis Function (RBF) for modeling and prediction based on the information of location [7]. Hamzaçebi forecasted seasonal time series data by using a novel artificial neural network (ANN) structure [8]. This model demonstrates lower prediction error compared with other traditional statistical models [8]. Based on the time-series data from the local weather station, Hewage used the Temporal convolutional neural (TCN) network to predict the weather effectively and accurately [9]. Geogre and Gwilym introduced a new model named Auto regressive integrated moving average (ARIMA), which is a time series prediction method [10]. This model is a traditional approach for time series analysis that is widely applied and the Seasonal Autoregressive Integrated Moving Average (SARIMA) model developed from ARIMA model. The SARIMA model combines the ARIMA model with a Stochastic Seasonal Model, abbreviated as SARIMA(p, d, q) (P, D, Q)S [10]. Therefore, SARIMA model can be considered if the time series data demonstrated notable time fluctuation trends and seasonal characteristics [11]. Besides, a new model for prediction named SARIMAX can be applied if the dataset contains exogenous variable (X) like weather conditions (sunny, cloudy, rainy, etc.), humidity, wind speed.

In summary, predicting temperature has important meanings and attracted many scholars to do. This essay will make weather forecasts for a specific city by using ARIMA model and SARIMA model, compare their accuracy and predict the temperature for next few years. After that, some exogenous variables will be introduced to test the accuracy of SARIMAX model.

2. Methods

2.1 Data Sources

Time series data: The data is the average temperature of the city Alexandria in Egypt from May 1791 to September 2013, the total number is 2668. From Figure 1, this dataset has obvious seasonal characteristics, which is shown graphically in Figure 1. It demonstrates that there is an increasing trend in temperature of each season because of the human activities since industrial revolution.

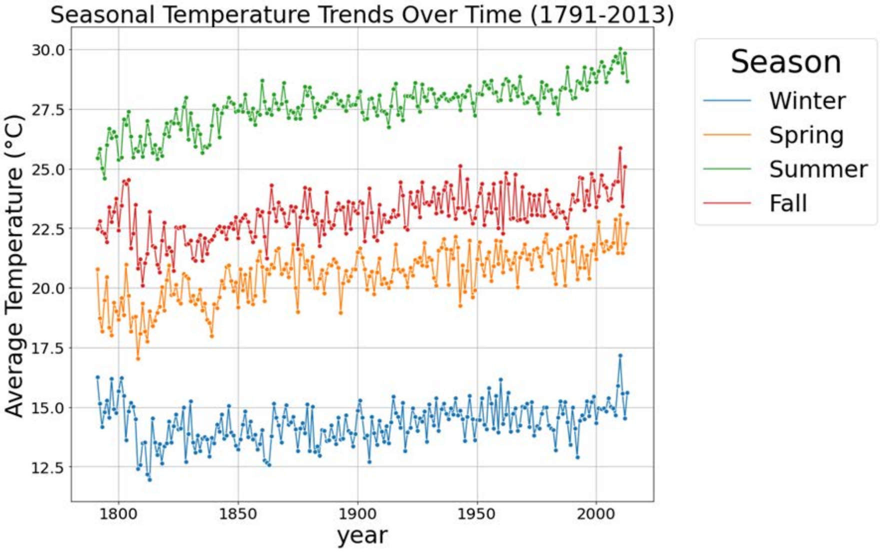

Fig. 1 Seasonal temperature trends over time (1791-2013) in Alexandria.

ISSN 2959-6157

2.2 Methods Introduction

2.1.1 Model Selection

This essay chooses Autoregressive Integrated Moving Average (ARIMA) model, Seasonal Autoregressive Integrated Moving Average (SARIMA) model and Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX) model to do prediction as they are typical and traditional forecasting model based on historical data.

The parameters of ARIMA models are commonly represented by (p, d, q), where p, d and q are the autoregressive order, difference times and moving average order respectively. In order to determine the parameters of ARIMA model, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of time series data should be drawn and observed.

Through observing these graphs, the stationarity and trends of data can be preliminarily determined and thus determine the p, d and q. The SARIMA model, abbreviated as SARIMA(p, d, q) (P, D, Q, S), where p, q and d are the same as ARIMA, P and Q represent the seasonal autoregressive and moving average orders, D and S represent seasonal difference times and seasonal period and cycle length respectively [10]. The parameters P, D and Q can also be determined by observing the ACF and PACF.

The SARIMAX model is typically represented as SARIMAX (p, d, q) (P, D, Q, S)[X], where X represents exogenous variables. This model considers the influence of external factors, which can improve the performance of the model.

2.2.2 Measures of performance

In order to evaluate the performance of the forecasting model, tests for mean square error (RMSE) are performed between the predicted values and the actual values. This performance metrics is commonly used in time series analysis and it can be represented by the following equations:

$$RMSE = \sqrt{(1/n)\sum_{t=1}^{n}(E_t - F_t)^2} \qquad (1)$$

E_t and F_t represent the predicted values and actual values respectively at time t and n represents the estimate values. Generally, the smaller the RMSE, the smaller the difference between the actual values and the predicted values. Therefore, it can be concluded that the model has better ability of prediction.

3. Results and Discussion

3.1 Data Processing

Firstly, data required to be preprocessed before analysis. This includes selecting data for a specific time range, filling in missing values, etc. After that, the stability of the data used for ARIMA must be ensured. Therefore, the autocorrelation function (ACF) plot and the partial autocorrelation function (PACF) plot need to be drawn. From the autocorrelation function (ACF) image (Figure 2), the x-axis represents the time lag and the y-axis represents the value of the autocorrelation coefficient. Each point represents the autocorrelation coefficient for the corresponding lag value, the blue line represents the value of the autocorrelation coefficient at different lag values, and the blue shaded area represents the boundary of statistical significance, usually the 95% confidence interval. If the autocorrelation coefficient exceeds this shaded area, then the autocorrelation can be considered statistically significant at that lag. Also, it can be observed that these data demonstrate an obvious seasonal pattern as the presence of a peak or trough appears every specific number of lags. Therefore, it may be necessary to consider seasonal differentials or use a seasonal model (SARIMA). In light of this, it can be guessed that SARIMA will be more advantageous in dealing with and predicting these data.

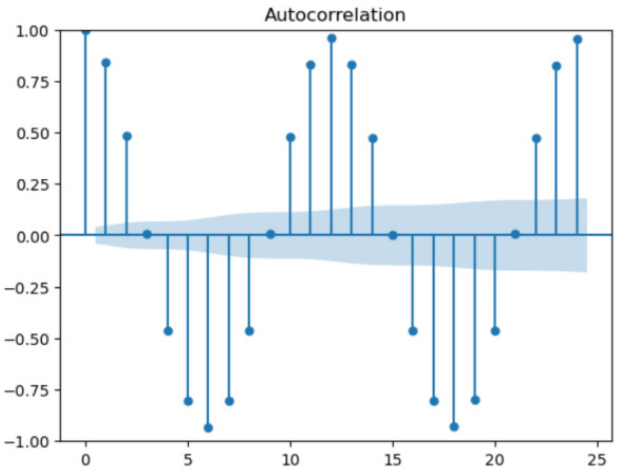

Fig. 2 The ACF plot of average temperature

According to the partial autocorrelation function (PACF) image (Figure 3), it can be concluded that the partial autocorrelation coefficient is very high at lag 1. After that, a few points after lag 1, the partial autocorrelation coefficient drops rapidly to near 0 and fluctuates around 0 in the following lags. These are exactly the typical characteristics of stable data. But in order to determine if the data is stable, ADF test need to be performed on the data.

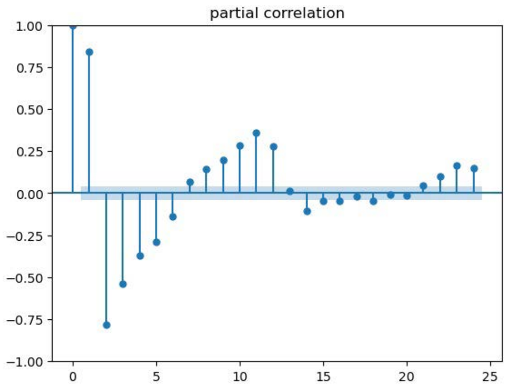

Fig. 3 The PACF plot of average temperature

The p-value of the ADF (Augmented Dickey-Fuller) test is used to determine the stationarity of time series data. Null hypothesis can be rejected if the P-value is less than 0.05 and vice versa [12]. In other words, if P-value is less than 0.05, the time series data will be stable. According to Table 1, the P-value is around 0.0007, which is obviously less than 0.05. Therefore, the data is stable and there is no need to do a first-order difference and the parameter d in the time series models ARIMA and SARIMA should be 0.

Table 1. The ADF test

ADF statistics	p-value
-4.184	0.001

3.2 Model Evaluation and Prediction

From Figure 2 and 3, some parameters can be determined for SARIMA model. According to the PACF, at the first lag (lag 1), there is a significant positive partial autocorrelation value, which implies that the model may require a seasonal autoregressive term P=1. Moreover, there is also a significant peak at the 12th lag, which is the same as the 1th lag, which suggests that the data may have an annual seasonal cycle. Therefore, S should be 12. From the ACF plot, seasonal difference may be required due to the seasonal pattern and hence D need to be 1. Besides, the ACF shows a significant tail at seasonal lags (the 12th lag), then the moving average orders Q > 0 may be required. Here, the value of Q is 1.

When considering to choose the parameters p and s for ARIMA and SARIMA model, sometimes it is necessary to take into account the rule of thumb and the results of model diagnosis. From rule of thumb, the values of p and s should not be more than 3 or 4 in order to avoid overfitting. Besides, RMSE can be a good indicator to measure the deviation between fitting values and actual values and Akaike information criterion (AIC) can be utilized for choosing models that have the smallest number of free parameters in the dataset to prevent situations of overfitting. Using these two metrics and comparing their values can make it possible to choose the optimal combination of parameters.

Table 2. Measures of performance for ARIMA

ARIMA Model	RMSE	AIC
(0,0,0)	4.731	13410.92
(0,0,1)	2.596	10720.70
(0,0,2)	1.984	9930.24
(1,0,0)	2.655	10990.20
(1,0,1)	1.801	9527.68
(1,0,2)	1.761	10542.96
(2,0,0)	1.625	9531.32
(2,0,1)	1.239	8100.57
(2,0,2)	1.228	7733.06

Table 3. Measures of performance for SARIMA

SRIMA Model	RMSE	AIC
(0,0,0)	0.970	7067.74
(0,0,1)	0.965	6632.25
(0,0,2)	0.973	5872.80
(1,0,0)	0.968	7023.96
(1,0,1)	0.972	6273.46
(1,0,2)	0.965	5716.31
(2,0,0)	0.967	6452.73
(2,0,1)	0.976	6149.54
(2,0,2)	0.970	5709.10

From the Table 2 and Table 3, it can be concluded that the parameters (2,0,2) is optimal for ARIMA and SARIMA model with the lowest RMSE and AIC. Also, the SARIMA model has better predicting performance than the ARIMA model. In light of this, the ARIMA and SARIMA models are used to make predictions and two pictures of actual values and predicted values (Figure 4 and 5) are drawn based on temperature data up to 2014.

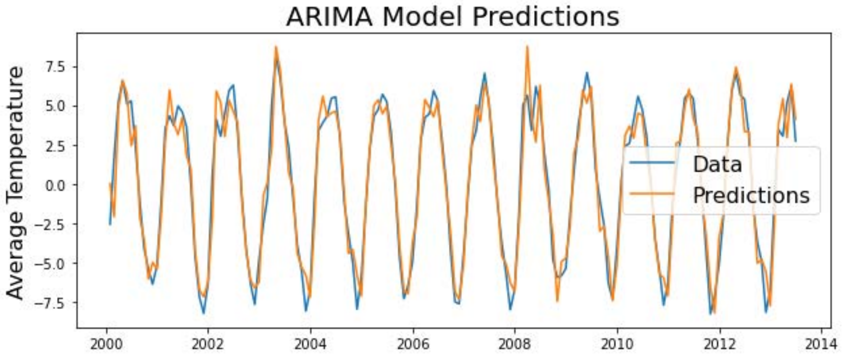

Fig. 4 Predictions from ARIMA

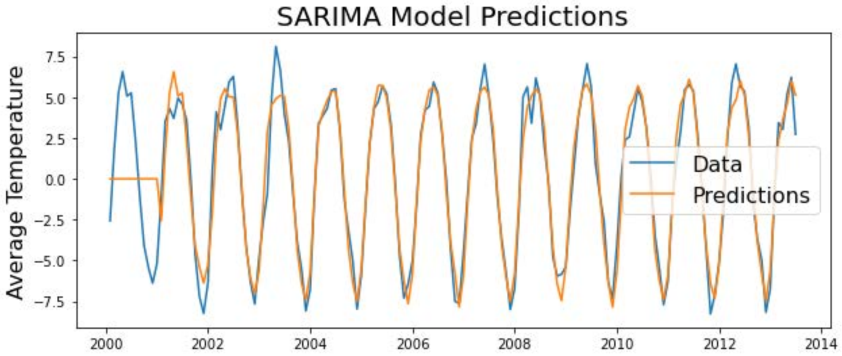

Fig. 5 Predictions from SARIMA

It can be seen that the two models successfully completed the prediction task. After that, some exogenous variables X are tried to be introduced into the SARIMA model and they can bring additional information which can help to improve the performance of predicting and the accuracy of the model. Based on the dataset, the latitude and longitude information of the city Alexandria can be obtained, which are 31.35N and 30.16E respectively. They can function as exogenous variables and be introduced to the SARIMA model.

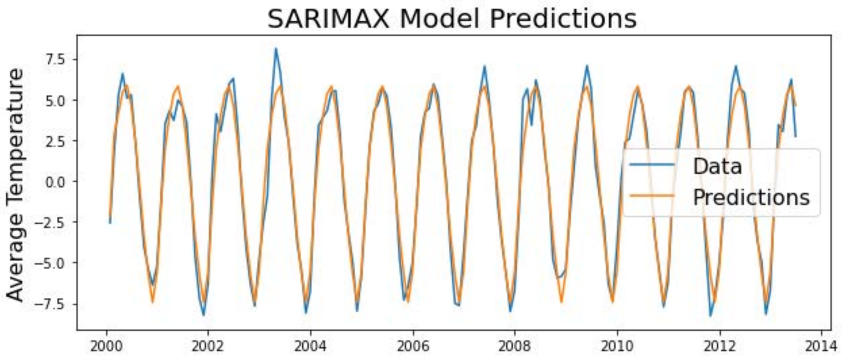

Fig. 6 Predictions from SARIMAX

Table 4. Comparsion among different models

Model	RMSE
ARIMA	1.228
SARIMA	0.970
SARIMAX	0.965

According to Figure 6 and Table 4, SARIMAX has the best performance with the lowest RMSE. Compared with ARIMA, it has a better forecasting ability, and at the same time, compared with the SARIMA model, there is no large deviation between the predicted and actual values in some years (e.g. 2000).

4. Conclusion

Through the research in this essay, it can be concluded that ARIMA model and SARIMA model both have successfully accomplished the task of prediction and demonstrate good performance of predicting with small errors. The result shows that the SARIMA model performed slightly better than the ARIMA model. This suggests that seasonal factors contained in the SARIMA model can provide additional valuable information that enhance the predicting power. Besides, another important improvement in predicting accuracy is to introduce some exogenous variables into the SARIMA model and this creates the SARIMAX model. The superior predicting performance of the SARIMAX model shows that exogenous variables can improve the predicting power. Despite these findings, this study has some limitations as well. One of them is the lack of exogenous variables in the dataset, which contains only longitude and latitude information. This limitation may not exploit the potential of the SARIMAX model.

These findings can have important implications for predicting in areas where seasonal factors and exogenous variables can play an important role. Therefore, further research with more exogenous variables is necessary to exploit the potential of the SARIMAX model. Besides, research in the future can attempt to explore the ability of SARIMAX model in other domains and examine the best methods for selecting exogenous variables.

References

[1] Kumari S, Muthulakshmi P. SARIMA Model: An Efficient Machine Learning Technique for Weather Forecasting. Procedia Computer Science, 2024, 235: 656-670.

[2] Pérez-Aracil J, Fister D, Marina CM, et al. Long-term temperature prediction with hybrid autoencoder algorithms. Applied Computing and Geosciences, 2024, 23: 100185.

[3] Jain G, Mallick B. A review on weather forecasting techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2016, 5(12): 177-180.

[4] MCifuentes J, Marulanda G, Bello A, et al. Air temperature forecasting using machine learning techniques: a review. Energies, 2020, 13(16): 4215.

[5] Dadhich S, Pathak V, Mittal R, et al. Machine learning for weather forecasting. Machine Learning for sustainable Development, 2021, 10.

[6] Manigandan P, Alam M D S, Alharthi M, et al. Forecasting natural gas production and consumption in United States-evidence from SARMA and SARIMAX models. Energies, 2021, 14(19): 6021.

[7] Lee S, Cho S, Wong P M. Rainfall prediction using artificial neural networks. journal of geographic information and Decision Analysis, 1998, 2(2): 233-242.

[8] Hamzacebi C. Improving artificial neural networks' performance in seasonal time series forecasting. Information

Sciences, 2008, 178(23): 4550-4559.

[9] Hewage P, Behera A, Trovati M, et al. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Computing, 2020, 24: 16453-16482.

[10] Ma S, Liu Q, Zhang Y. A prediction method of fire frequency: Based on the optimization of SARIMA model. PLoS one, 2021, 16(8): e0255857.

[11] Ren S, Cui H B. Application of SARIMA time series analysis in tax forecast: take Guizhou Province as an example. Journal of Hubei University (Natural Science), 2021, 43(1): 80-85.

[12] Mushtaq R. Augmented dickey fuller test. SSRN electronic Journal, 2011.