Time Series Forecasting of Temperatures using SARIMA: An Example from Nanjing

Peng Chen[1, *], Aichen Niu[1, a], Duanyang Liu[2], Wei Jiang[3], Bin Ma[1]

[1]Jiangsu Meteorological Information Center, Nanjing, China
[2]Jiangsu Meteorological Observatory, Nanjing, China
[3]Jiangsu Meteorological Climate Center, China

*Corresponding author e-mail: 409856986@qq.com, [a]ai.nwork@hotmail.com

Abstract. Time series modelling and forecasting – a method that predicts future values by analysing past values – plays an important role in many practical fields. In this paper, we analyse the monthly mean temperature in Nanjing, China, from 1951 to 2017, using SARIMA (Seasonal Autoregressive Integrated Moving Average) techniques. Data from 1951 to 2014 are used as the training set, while data from 2015 to 2017 are used as the testing set. A detailed explanation of model selection and forecasting accuracy is presented. The results show that the proposed research approach obtains good forecasting accuracy.

1. Introduction

The main goal of time series modelling is to collect and analyze past values to develop appropriate models that describe the inherent structure and characteristics of the series [1]. Time series forecasting is the use of certain model to forecast future values based on past observed values, and thus can be understood as a method for predicting future values by understanding past values [2]. Numerical weather forecasts use atmospheric models to predict future weather conditions based on current weather conditions [3, 4, 5]. Unlike numerical weather prediction, time series forecasting uses a model to predict future values based on past values. Owing to the importance of time series forecasting in countless practical fields, researchers should pay proper attention to fitting an appropriate model to the time series. Over the year, many intelligent time series models have been developed in the literature to improve the accuracy and efficiency of time series forecasting. One of the most widely used and recognized statistical forecasting time series models is the Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model is well-known for notable forecasting accuracy and efficiency in representing various types of time series [6] with simplicity as well as the associated, Box–Jenkins methodology for optimal model construction. The basic assumption made in implementing this model is to assume the time series is linear and follows a statistical distribution, such as the normal distribution [1]. For seasonal time series forecasting, Box and Jenkins [7] proposed a quite successful variation of the ARIMA model called the Seasonal ARIMA (SARIMA) model.

Air temperature is a common meteorological variable indicative of how hot or cold the air is. It not only affects the growth and reproduction of plants and animals, but also has an influence on nearly all other meteorological variables, such as the rate of evaporation, the relative humidity, wind speed, wind direction and precipitation patterns. In this paper, we analyze the monthly mean temperature in Nanjing, a city in the southeast of China, during 1951–2017. The monthly mean temperature during 1951–2014 is used as the training set, while that during 2015–2017 is used as the testing set. To evaluate the forecast

accuracy, as well as to compare the results obtained from different models, the mean-square error (MSE) is calculated. Section 2 describes the data and briefly discusses the ARIMA method to time series modelling and forecasting. Section 3 reports and interprets the results obtained from this method, evaluates the accuracy of the fitted forecasting models and compares the different models fitted to the time series. Section 4 offers discussion and conclusions.

2. Data and methods

The data used in this study are the monthly mean temperature of Nanjing from January 1951 to December 2017. For the monthly mean temperature, data from 1951 to 2014 are used for training, while data from 2015 to 2017 are used for testing. The original temperature data are from the automatic weather station in Nanjing, collected on an hourly basis, and there are no missing values. Meanwhile, the monthly mean temperature data are from the original observed temperature data. The longitude and latitude of the automatic weather station is 118°54′00″ and 31°56′00″, respectively. The time series of the monthly mean temperature is plotted in Figure 1.

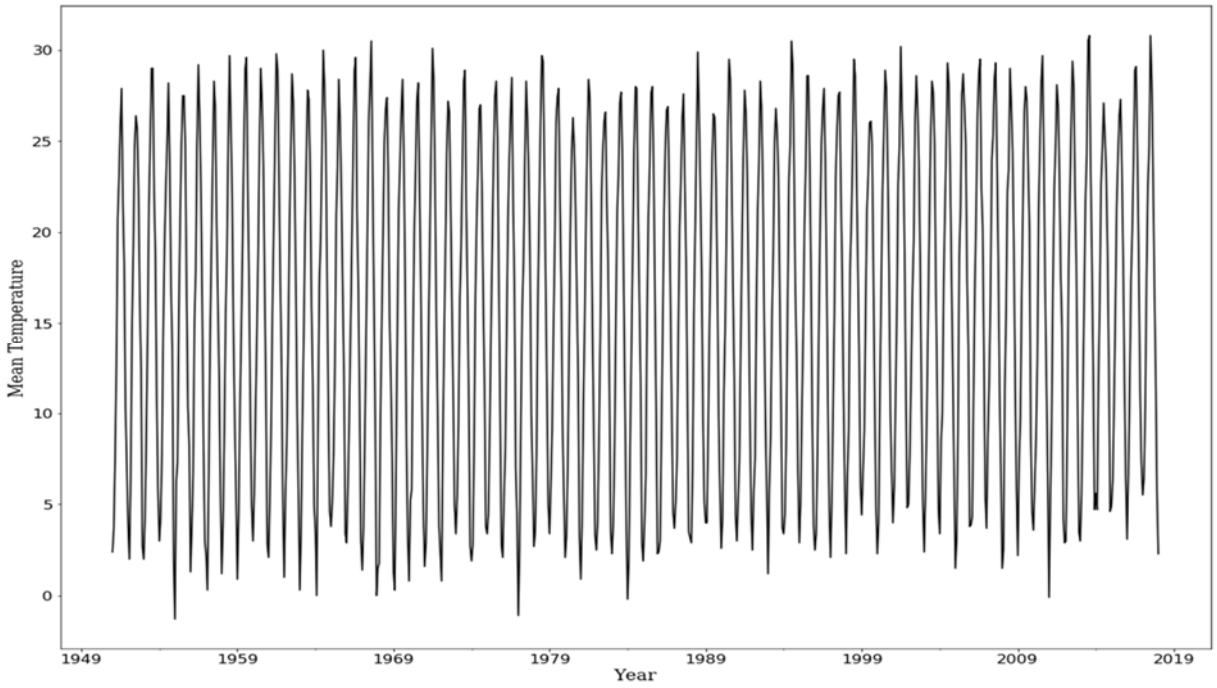

Figure 1. Time series of monthly mean temperature in Nanjing, China.

Seasonal ARIMA model (SARIMA) is formed by adding seasonal terms in the ARIMA models listed above. SARIMA models are written as

$$\text{ARIMA}\ (p, d, q)\ (P, D, Q)_{m} \tag{1}$$

Where (p, d, q) and $(P, D, Q)_{m}$ are the non-seasonal and seasonal part of the model, respectively. The parameter m is the number of periods per season. The seasonal part of the model is very similar to the non-seasonal part, but it is involved in backshifts of the seasonal period. Using available dataset, the ARIMA model is finalized by changing the values of p, d and q. To determine the parameters of an ARIMA model, Akaike's Information Criterion (AIC) is widely used. It is given by

$$\text{AIC}\ (p) = n\ln\ (\text{RSS} / n) + 2\text{K} \tag{2}$$

Where n is the number of data points and RSS is the residual sums of squares. The model with the minimum AIC value will be selected as the best forecasting model. Another method to determine appropriate parameters of an ARIMA model is to analyse auto correlation function (ACF) and partial autocorrelation function (PACF) plots.

3. ARIMA modelling of temperature time series and results

This section describes the proposed ARIMA model and presents the processes of model selection. The first step is to formulate a class of models and assume certain hypotheses. The next step is to estimate the parameters of this identified model. Sections 3.1–3.4 describe all the steps in detail.

3.1. Step 1

In this process, the data should be plotted to identify any unusual values. To stabilize the variance, data need to be rescaled if necessary. All data are rescaled using the formula

$$Vi = \frac{a_i - \min(a_i)}{\max(a_i) - \min(a_i)} \tag{3}$$

Where V_i is the rescaled value, a_i represents the original data, and min (a_i) and max (a_i) are the minimum and maximum values of the original data set.

3.2. Step 2

In this step, the ACF and PACF of the rescaled data are plotted, as shown in Figure 3. The ACF and PACF are used to determine if an AR (p) or MA (q) model is appropriated and determine possible candidate models.

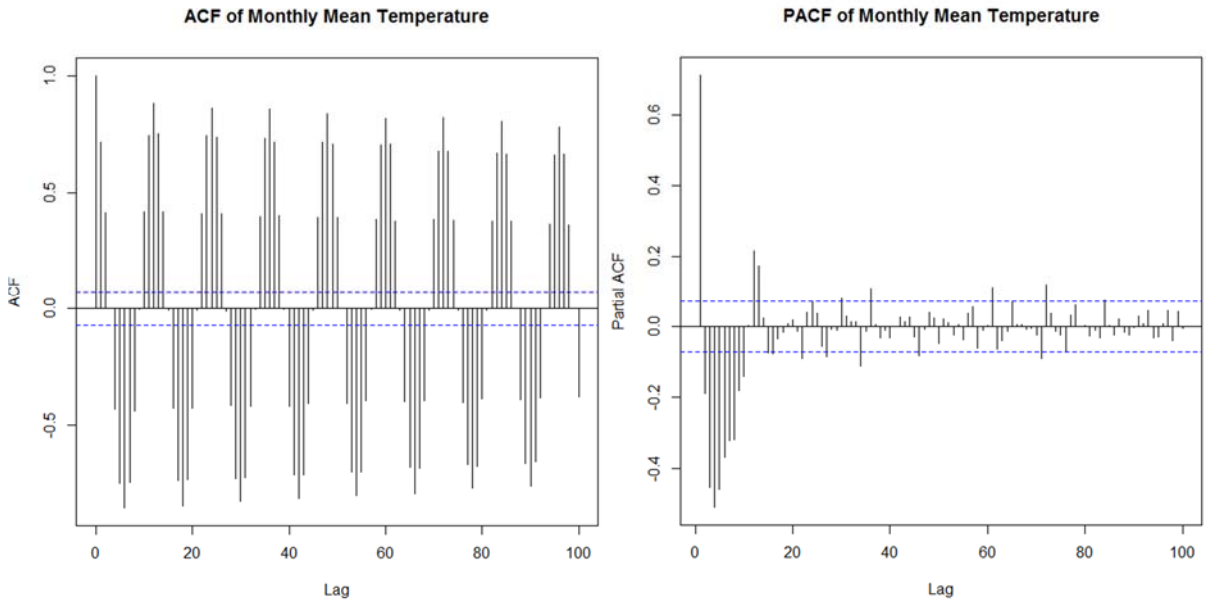

Figure 2. ACF and PACF of monthly mean temperature during 1953-2014.

3.3. Step 3

In this step, a SARIMA model is applied to forecast the temperature data. For the monthly mean temperature, observations one year apart in the time series x_t might be modelled as

$$\emptyset(B^{12})\Delta_{12}^{D} x_t = \theta(B^{12})\alpha_t \tag{4}$$

Where$\nabla_{12}x_t = (1 - B^{12})x_t = x_t - x_{t-12}$, and $\emptyset(B^{12})$ $\theta(B^{12})$ are polynomials in the B^{12} of p and q, respectively. Both terms satisfy appropriate stationarity and invertibility conditions [8]. Generally, the error component α_t would be expected to be correlated with the time series.

The method used in this study to search for the appropriate parameters of forecasting models is hyper parameter optimization. In this study, the ARIMA (p, d, q) (P, D, Q) $_m$ model requires six parameters: p, d, q, P, D and Q. The value of m is set as 12 because the data used are monthly data with a period of 12. The AIC values of selected models are shown in Table 1. According to Table 1, SARIMA (1, 1, 1) × $(1, 0, 1)_{12}$ shows the lowest AIC value. Thus, this model should be considered as the best forecasting model.

Table 1. AIC values of SARIMA models.

Parameters		AIC Value
p, d, q	P, D, Q, m	
0, 0, 0	0, 0, 1, 12	471.85
1, 1, 1	1, 0, 0, 12	-2280.76
1, 1, 1	1, 0, 1, 12	-2754.63
1, 4, 1	4, 1, 2, 12	-1183.66
2, 1, 3	4, 2, 3, 12	-2446.92

3.4. Step 4

The forecast accuracy of the selected model is validated by applying a diagnosis check. According to Table 1, the AIC value of SARIMA (1, 1, 1) × $(1, 0, 1)_{12}$ is the lowest.

Table 2. Results of the diagnostics test of the SARIMA (1, 1, 1) × (1, 0, 1,) 12 model.

	Coef.	Std err.	*z*	*P*> |*z*|	[0.025	0.975]
MA.L1	−0.8569	0.037	−23.186	0.000	−0.929	−0.784
MA.L2	−0.1098	0.036	−3.061	0.002	−0.180	−0.039
AR.S.L12	0.9990	0.000	3191.831	0.000	0.998	1.000
MA.S.L12	−0.9743	0.023	−43.269	0.000	−1.018	−0.930

Table 2 summarizes the results of the diagnostics test of the SARIMA (1, 1, 1) × $(1, 0, 1)_{12}$ model. The second column is the weight of the coefficients. The 'Coef.' column shows the weighting (i.e., importance) of each feature and how each one impacts the time series**.** Since all values of *P*> |*z*| are less than 0.05, the results are statistically significant.

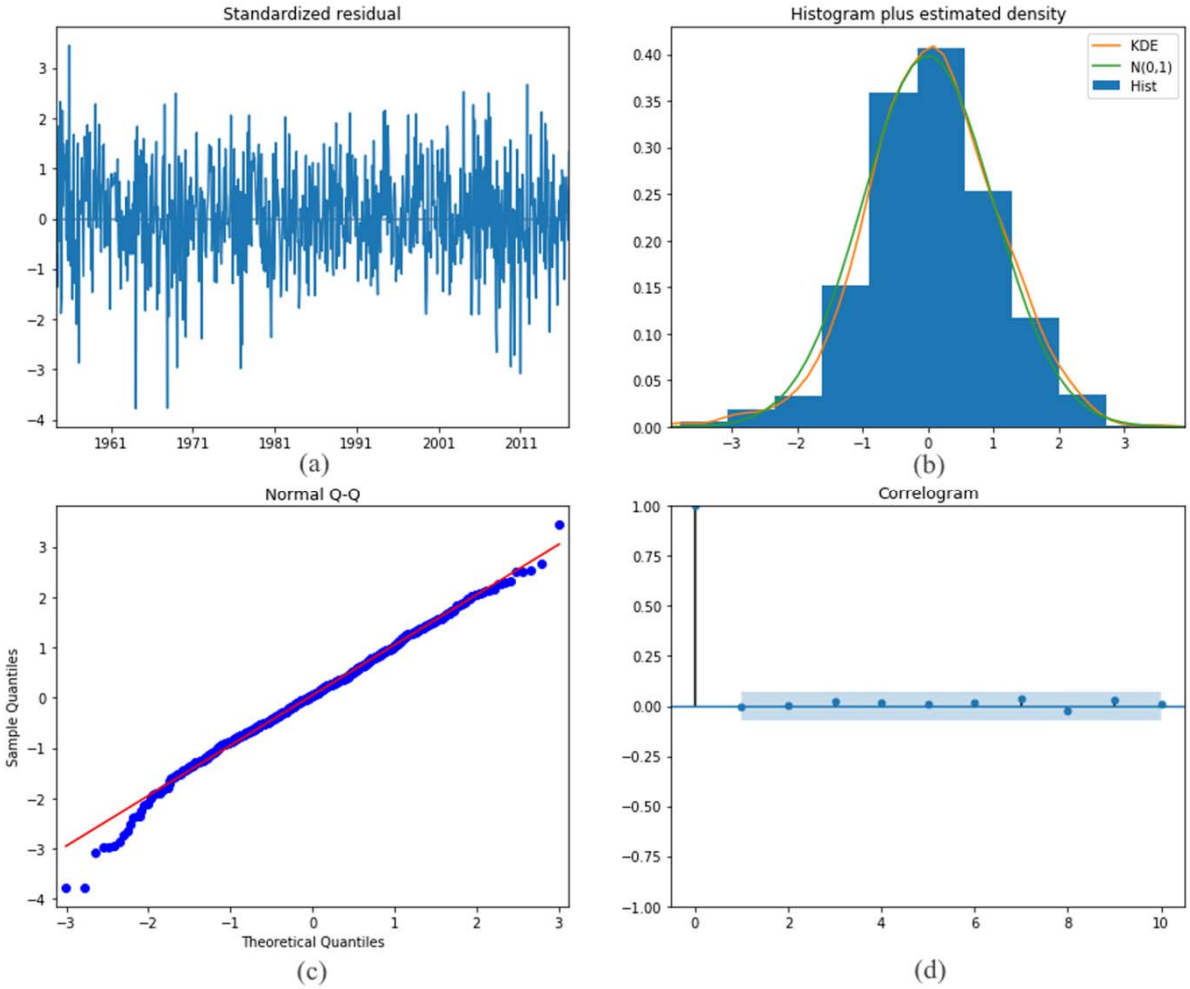

Figure 3. Plots of residuals: (a) residuals over time; (b) frequency distribution histogram; (c) Q-Q plot; (d) autocorrelation.

The residuals over time are shown in Figure 4a. The results imply the residuals show no obvious seasonality and appear to be white noise. Likewise, the autocorrelation shown in Figure 4d implies that the residuals of the original data have low correlation with the lagged data.

According to Figure 4b, the Kernel Density Estimation (KDE) (red curve) is nearly overlapped with the N (0, 1) (green curve). The results imply that the residual follows a normal distribution, with mean equal to 0 and standard deviation equal to 1. In Figure 4c, the red line stands for a normally distributed dataset, with mean equal to 0 and standard deviation equal to 1, while the blue dots represent the residuals. The Q-Q plot of the residuals implies that the residuals follow a linear trend. Thus, the residuals are normally distributed. In general, the model shows good forecasting accuracy and can be used to predict future values.

4. Discussion and conclusions

The selected model can now be used to make forecast time series. Due to the fundamental importance of forecast accuracy, a test should be performed to verify the forecasting accuracy, by comparing the forecast values with observational values. This test can also avoid under-fitting or over-fitting. The statistical tests of the forecast results are analysed in detail, as follows.

The model predicts the next 36 months' mean temperature at Nanjing station, based on the 35 years of past data. Data from January 1980 to December 2015 are used as the training set, while data from

January 2015 to December 2017 are used as the testing set. The MSEs of the predicted values from 2015 to 2017 are 0.84, 0.89 and 0.94, respectively. The MSEs are relatively low, with an increasing trend of 0.05 every year. Since the increasing trend is not obvious, the selected model shows good forecasting accuracy of the testing set and can be applied in future works.

Figure 5 shows the training set and a comparison between the testing set and forecast values. According to Figure 5, the forecast values (red line) are close to the real values (blue line), and are within the confidence intervals (grey shading). The MSE of the forecast values is 0.89, which is relatively low. In general, the forecast results are acceptable.

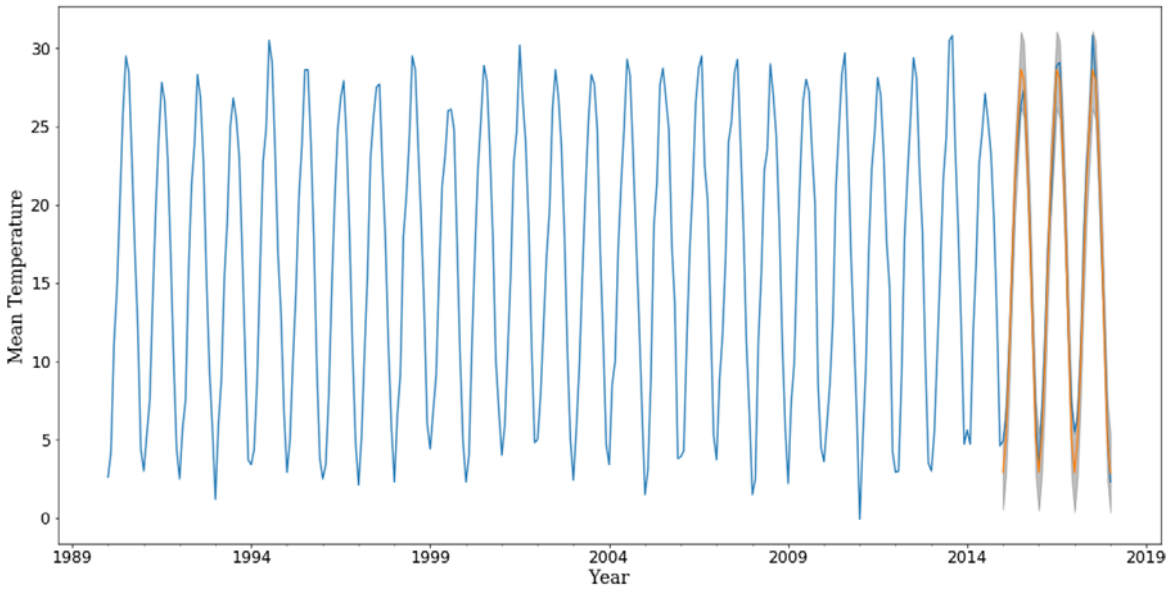

Figure 4. Comparison between the real values and forecast values.

According to the above discussion, the selected SARIMA model can be used to forecast future values because its forecasting accuracy is acceptable. In future work, we intend to widen the range of parameter combinations when carrying out the grid search. This process might help us to identify models with higher forecasting accuracy. Furthermore, forecasting accuracy might be related not only to the parameters of the SARIMA model, but also to the length of the training set. Both assumptions should be studied in a follow-up study.

References

[1] Adhikari, R., & Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting.

[2] Raicharoen, T., Lursinsap, C., & Sanguanbhokai, P. (2003). Application of critical support vector machine to time series prediction. International Symposium on Circuits and Systems (Vol.5, pp.V-741-V-744 vol.5). IEEE.

[3] Barker, D., Huang, X. Y., Liu, Z. Q., Auligné, T., Zhang, X., & Rugg, S., et al. (2012). The weather research and forecasting model's community variational/ensemble data assimilation system: wrfda. Bulletin of the American Meteorological Society, 93 (6), 831 - 843.

[4] Shen, F., Min, J., & Xu, D. (2016). Assimilation of radar radial velocity data with the wrf hybrid etkf–3dvar system for the prediction of hurricane ike (2008). Atmospheric Research, 169, 127-138.P.G. Clem, M. Rodriguez, J.A. Voigt and C.S. Ashley, U.S. Patent 6,231,666. (2001).

[5] Xu, D., Min, J., Shen, F., Ban, J., & Chen, P. (2016). Assimilation of mwhs radiance data from the fy‐3b satellite with the wrf hybrid‐3dvar system for the forecasting of binary typhoons. Journal of Advances in Modeling Earth Systems, 8 (2).

[6] Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on dwt decomposition ☆. Procedia Computer Science, 48, 173 - 179.
[7] Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis forecasting and control - rev. ed. Oakland, California, Holden-Day, 1976, 37 (2), 238 - 242.