Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
JUL. 2023 VOL.2, NO.4

Time Series Analysis — Analysis and Prediction of Monthly Mean Temperature in Baotou City

Linlin Gong[1]

[1] Ocean University of China

Correspondence: Linlin Gong, Ocean University of China.

doi:10.56397/IST.2023.07.05

Abstract

This paper analyzes the monthly average temperature of Baotou City from January 1, 1973, to December 31, 2022. The data comes from the R language worldmet dataset, with 600 data. Observing the time series chart shows the data have obvious periodicity and seasonality, and there is no trend. The order ARIMA(1,0,1)$\times(1,0,1)_{12}$s determined by observing the ACF and PACF diagrams, but since there is a trend term of AR, a difference is added to get P=1, D =1, Q=1. By comparing the AIC, BIC and σ^2 values of ARIMA(1,0,1)$\times(1,0,1)_{12}$ and ARIMA(1,0,1)$\times(1,1,1)_{12}$, the residual analysis diagrams are observed. The residual analysis diagrams are not very different, which reflects that the residual is white noise. The AIC, BIC and σ^2 values of the former are all smaller than those of the latter, so the time series model is determined as ARIMA(1,0,1)$\times(1,1,1)_{12}$, and the expression is, $x_t = 0.2859x_{t-1} - 0.1122w_{t-1} - 1.0000x_{t-12} - w_{t-12} + 0.0037$.

Where x_t represents the predicted value at time point t, x_{t-1} and x_{t-12} represent the original observed value at time point t-1 and t-12, respectively, and w_{t-1} and w_{t-12} represent the residual term at time point t-1 and t-12, respectively. The model takes into account the seasonality and trend of meteorological data and can fit the data well and make future temperature predictions. After residual analysis and model selection, the prediction effect of this model is good, the error is small, and it can provide a certain reference value. However, there may be shortcomings such as seasonal effects on forecast accuracy, which raises the need to improve models and study more meteorological data features. In general, meteorological data prediction based on time series analysis is an important research field, and more in-depth research and exploration are needed in the future to improve the prediction accuracy and provide better support for decision-making in the field of meteorology and climate.

Keywords: time series analysis, seasonal ARIMA model, R language, monthly mean temperature in Baotou City

1. Introduction

1.1 Background and Significance of Research

Meteorological change is a kind of time series data, and its change rules are time-dependent and continuous. Meteorological data time series analysis refers to the statistical and computational analysis of meteorological data to reveal the laws and trends in the time series, and to predict and forecast the future meteorological changes. It can be used to establish meteorological forecasting models, improve the accuracy and precision of meteorological forecasting, provide reliable meteorological forecasting services for agricultural production, energy development and other fields, help in-depth understanding of climate change, assessment of meteorological disaster risk, forecast meteorological change trends, etc., and have important reference value for formulating policies, planning, and decision-making in response to climate change.

In short, time series analysis of meteorological data is of great significance for us to understand meteorological change, predict the trend of meteorological change, and provide an effective scientific basis for formulating policies to deal with climate change.

1.2 Data Sources and Description

The data analyzed in this paper is the monthly average temperature of Baotou from January 1, 1973, to December 31, 2022. There are a total of historical weather data of Baotou obtained from R language worldmet database. The specific data is obtained by import (NOAA) function, incoming station number, and time range. Obtain daily hour-by-hour weather data for Baotou City from 1973 to 2022. The data source is from the National Oceanic and Atmospheric Administration (NOAA) weather station data, which saves various meteorological data including temperature, humidity, pressure, precipitation, and wind speed. This paper mainly analyzes the temperature of Baotou City.

1.3 Data Preprocessing

As this paper only analyzes the time series of air temperature in Baotou City, it only needs to select air_temp and date columns from weather data to form the data set and check whether there are missing values. It can be seen that missing values exist in the data set, and na.omit() function is used to remove rows containing NA values.

Since the downloaded data is hour-by-hour air temperature data, it takes a long time to fit the SARIMA model, and the model fitting and merging with hour-by-hour air temperature data will not significantly improve the accuracy of predicting future air temperature. In order to better establish the seasonal model, it is conducive to predicting the future air temperature of Baotou City. It is necessary to average the hourly temperature data given by the original data on a monthly basis. In order to facilitate the monthly average value of hourly temperature data, lubridate function and mutate function are used to split the date column and the month column into the data set. Then group_by is used to group the temperature data box by Year and Month to calculate the average temperature in each month. Next, use the summarise function to calculate the average temperature of a month, with only the year, month and average temperature remaining in the data box. Finally, use the unite function to combine the year and month into a Date, separated by a hyphen "-". The unite function adds a new date column to the data box and stores the result in a new data box named yearly_average.

Next, check whether there are outliers in the data by drawing a box plot. It can be seen from Figure 1 that there are no outliers in the data.

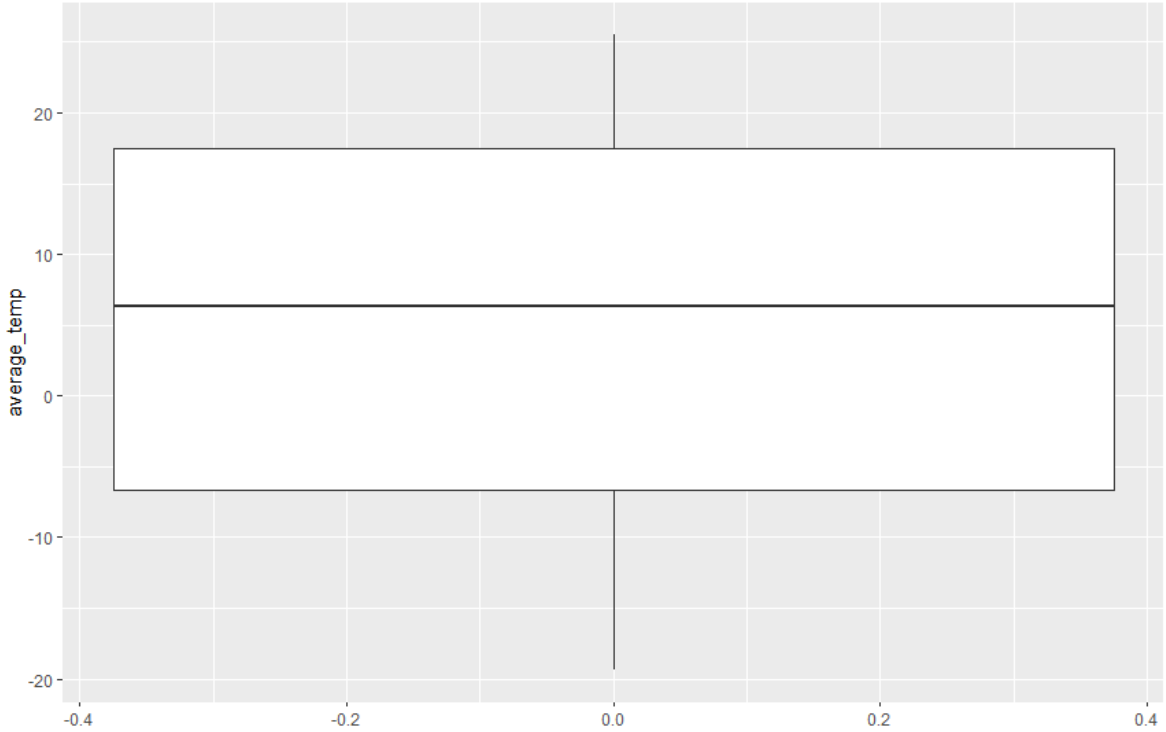

Figure 1. boxplot of average_temp

The time series diagram of the data is drawn below to check whether there are trends and seasonality in the series. As can be seen from Figure 2, the time series image of the data is relatively stable, and does not show a steady upward or downward trend in one direction, so there is no trend. The time series image of the data has a certain periodicity, so the data has seasonality. In the subsequent modeling, it is necessary to carry out seasonal difference

of the data to find the seasonal term and establish the seasonal model.

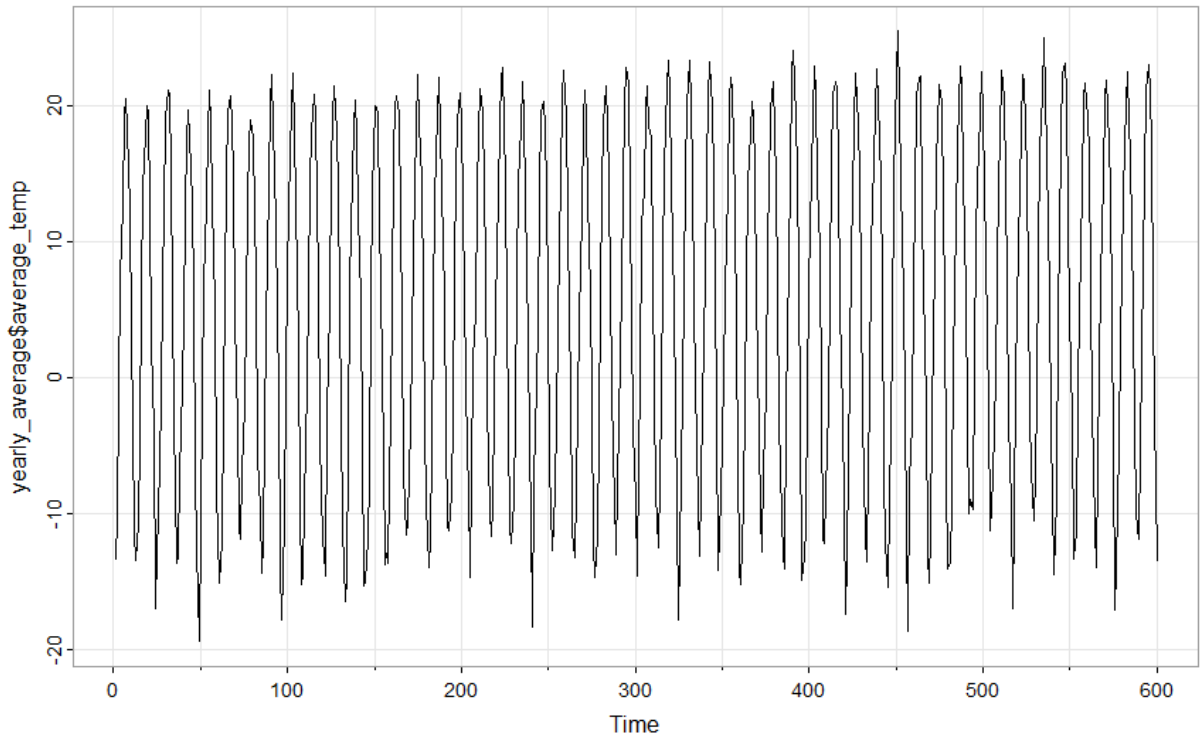

Figure 2. Time series diagram of average_temp

2. Introduction of Model and Theories

2.1 Introduction of Models

2.1.1 AR Model

AR (autoregression) model is a linear prediction model based on past observations. The AR model assumes that the value of the current time is a linear combination of observations at previous points in time, where past observations decay with decreasing weight coefficients. An autoregressive model is an extended form of a linear regression model, an autoregressive model of order p can be abbreviated as AR(p), the formula is as follows

$x_t = \alpha + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + w_t$, where x_t is stationary sequences, w_t is white noise.

2.1.2 MA Model

MA (Moving average) model is a linear prediction model based on the error term. The MA model assumes that the error at the current time is a linear combination of the error at the previous k time points, and has a mean of 0. The MA model can be represented by MA(q), and the representation error is the linear combination of the first q errors.

The prediction formula of the Q order MA model is as follows $x_t = w_t + \vartheta_1 w_{t-1} + \vartheta_2 w_{t-2} + \cdots + \vartheta_p w_{t-q}$, where w_t is white noise.

2.1.3 ARMA Model

ARMA (autoregressive moving Average) model is a time series prediction model that takes into account both the observed value of the first p time points and the error of the first q time points. The ARMA model can be represented by ARMA(p, q), that is, both AR and MA models are considered. The prediction formula of ARMA model is as follows.

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

2.1.4 ARIMA Model

An x_t process is called ARIMA(p, d, q)，if $\Delta^d x_t = (1-B)^d x_t$ is ARMA(p, q). Normally this model is written as $\Phi(B)(1-B)^d x_t = \alpha + \theta(B) w_t$, where $\alpha = \delta\ (1 - \Phi_1 - \cdots - \Phi_p)$ and $\delta = E(\Delta^d x_t)$.

2.1.5 Seasonal ARIMA Model

Seasonal ARIMA model is a common time series forecasting model, which is widely used to deal with seasonal and periodic time series data. It is an extension of the ARIMA model, in contrast to the ARIMA model, the seasonal ARIMA model adds seasonal terms to account for the seasonal effects of the data.

2.2 Order Determination

The ACF index measures the correlation between the time series and its corresponding lag value, and represents the correlation between the previous p time points and the current time point. PACF is the calculation of the degree of correlation between the sequence and the current value at a given point in time, after removing the influence of other intermediate orders. They are implemented by calculating the statistical characteristics and order of the differences between the previous p time points and the current time point. For AR(p) model, the PACF image is truncated after the p order, while the ACF image is trailed; for MA(q) model, the ACF image is truncated after the q order, while the PACF image is trailed. Both ACF and PACF images of the ARMA model are trailing.

2.3 Model Selection and Testing

The residual analysis diagram of sarima function is used for white noise test to determine whether the model is suitable and select the best model.

2.3.1 Standardized Residuals

This graph is mainly used to judge whether the residual term of the time series satisfies the two basic assumptions of “random” and “linear”. If the residual term shows a tendency to fluctuate randomly around 0, then the randomness hypothesis can be considered valid. If the fluctuation range of the whole residual series is large, then the hypothesis is not valid; If the residual term has a significant nonlinear tendency, then the linearity hypothesis is also not valid.

2.3.2 ACF of Residuals

Autocorrelation function (ACF) graphs can be used to determine whether the residual terms have autocorrelation or partial autocorrelation characteristics. If the residual term is characterized by random fluctuations near 0, then ACF should oscillate at 0 and all peaks are insignificant. If there is a significant peak, it indicates that there is an autocorrelation in the residuals, which means that the model may have captured some features of the time series in the modeling and have not been fully fitted.

2.3.3 Normal QQ Plot of Std Residuals

The normal QQ graph is often used to check whether a set of data follows a normal distribution, and if the residual term is random and it follows a Gaussian distribution, then the points in the graph should be distributed on a straight line. If the graph curve has asymmetric, non-linear features, it may mean that the model is not fully capturing the distribution characteristics of the data.

2.3.4 P value for Ljung-Box Statistic

In time series analysis, Ljung-Box statistics can be used to test whether there is an autocorrelation in the residual term of the time series. If the p-value ratio is smaller than the significance level (such as 0.05), the null hypothesis can be rejected and the data set is considered to have autocorrelation.

2.4 Model Prediction

A well-fitted SARIMA model is used to predict some of the data, and the predicted results are compared with the actual values. Some common indicators, such as mean square error (MSE), are used to evaluate the accuracy of the forecast results.

3. Empirical Analysis

3.1 Order Determination

First, the monthly mean temperature data of Baotou City from 1973 to 2022 were seasonally differentiated, and then the time series diagram (Figure 3), ACF and PACF diagram (Figure 4) were drawn. It can be seen from Figure 3 that the data after the difference is stable, with no trend or cycle.

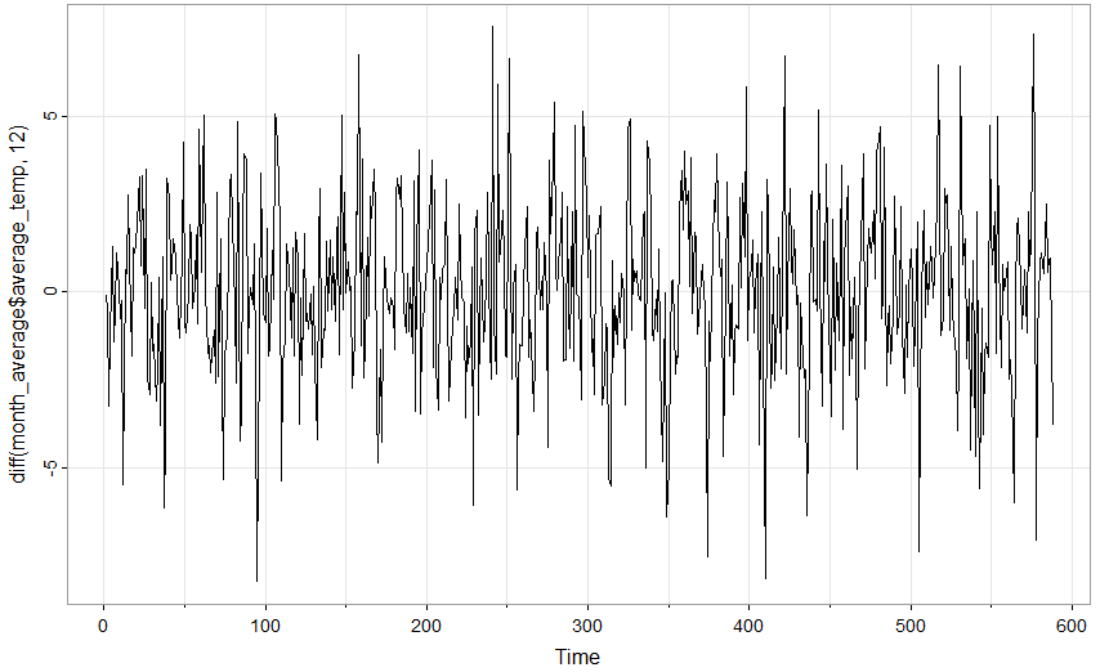

Figure 3. Time series diagram of monthly mean temperature data of Baotou City from 1973 to 2022 after seasonal difference

In this paper, the seasonal ARIMA model will be fitted based on the monthly average temperature data of Baotou City from 1973 to 2022, and the order of the model needs to be determined by looking at the ACF and PACF charts of the data. It can be seen from Figure 4 that both ARIMA and seasonal terms are lagging behind by a large order. Considering the cases of P=1, Q=1, P=1, Q=1, model fitting is carried out, and it is found that the AR part of non-stationary seasonality exists in the cases of P=1, Q=1, and the data needs to be stabilized by a difference. Therefore, consider making another difference in the seasonal term, that is, the seasonal component can consider ARIMA(1,1,1), P=1, D=1, Q=1.

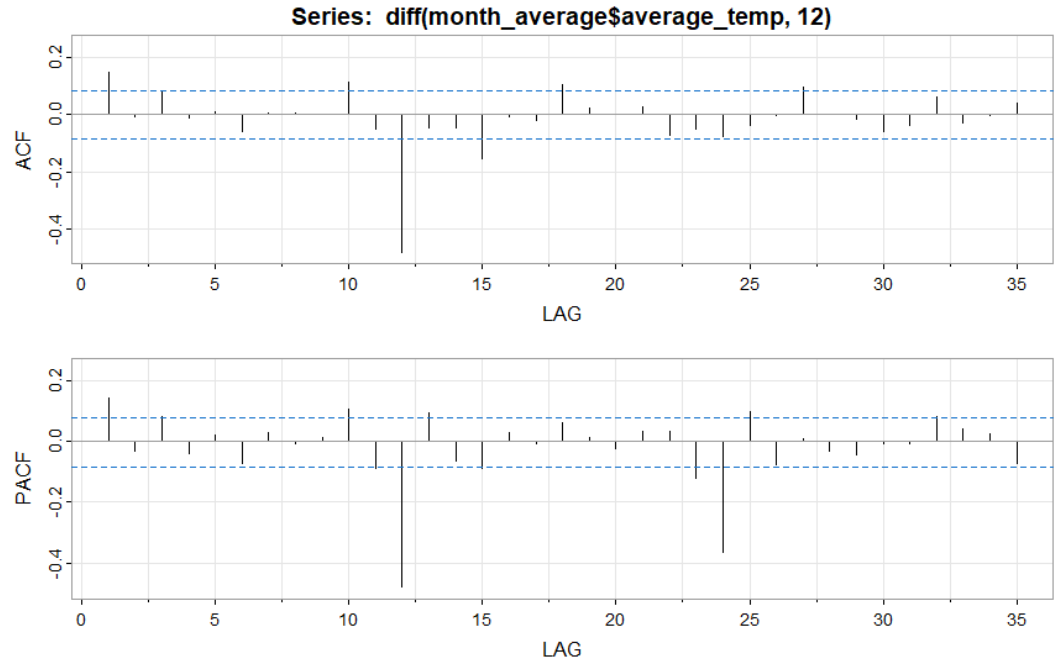

Figure 4. ACF and PACF charts of the monthly mean temperature data of Baotou City from 1973 to 2022 after seasonal difference

According to the order of the seasonal term in the above article, first try the data on ARIMA(1,0,1)× $(1,1,1)_{12}$

```
$ttable
         Estimate     SE  t.value p.value
ar1        0.2859 0.2572   1.1116  0.2668
ma1       -0.1122 0.2677  -0.4191  0.6753
sar1      -0.0310 0.0420  -0.7387  0.4604
sma1      -1.0000 0.0705 -14.1797  0.0000
constant   0.0037 0.0005   7.4629  0.0000

$AIC
[1] 4.023792

$AICc
[1] 4.023967

$BIC
[1] 4.068452
```

```
Coefficients:
          ar1      ma1     sar1     sma1  constant
      0.2859  -0.1122  -0.031  -1.0000    0.0037
s.e.  0.2572   0.2677   0.042   0.0705    0.0005

sigma^2 estimated as 2.958:  log likelihood = -1176.99,  aic = 2365.99
```

The residual analysis is shown in Figure 5. According to the p value for Ljung-Box statistic, it can be seen that most of the points are distributed above the dotted line, that is, for most of the residual p-value is greater than 0.05, but a small number of points are still distributed on the dotted line, and a small amount of autocorrelation may remain in the residual. It can be seen from the Normal Q-Q Plot of Std Residuals that the distribution of points is approximately a straight line, indicating that the residuals have good normality.

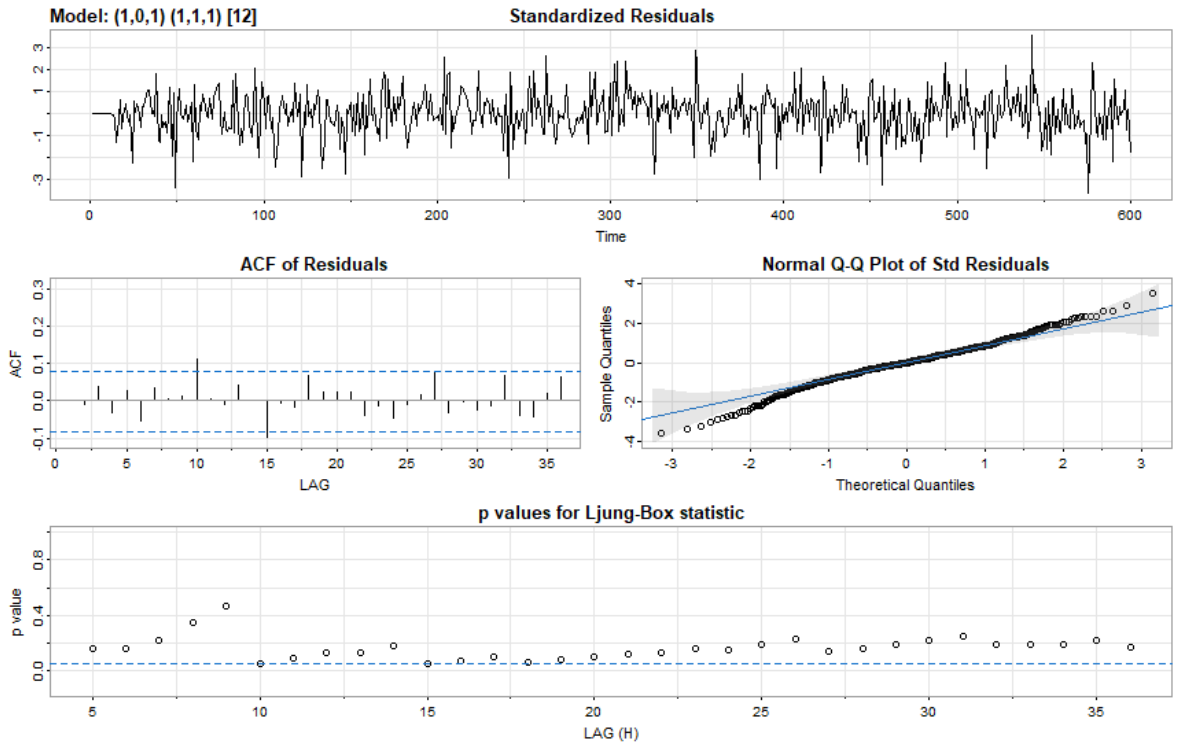

Figure 5. ARIMA(1,0,1)× $(1,1,1)_{12}$ Residual analysis diagram

3.2 Adding Parameters to Fit the Model in the Non-seasonal Terms of the ARIMA Model

Since adding an AR parameter and an MA parameter has little effect on the model, only adding a difference parameter (d=1) is considered, and the model becomes ARIMA(1,1,1)× $(1,1,1)_{12}$

```
$ttable
     Estimate     SE  t.value p.value
ar1    0.1720 0.0423   4.0684  0.0001
ma1   -0.9866 0.0143 -68.9142  0.0000
sar1  -0.0335 0.0425  -0.7875  0.4313
sma1  -1.0000 0.0886 -11.2816  0.0000

$AIC
[1] 4.040876

$AICc
[1] 4.040993

$BIC
[1] 4.078142
```

```
Coefficients:
          ar1      ma1     sar1     sma1
      0.1720  -0.9866  -0.0335  -1.0000
s.e.  0.0423   0.0143   0.0425   0.0886

sigma^2 estimated as 2.983:  log likelihood = -1181,  aic = 2371.99
```

Compared with ARIMA(1,0,1) × $(1,1,1)_{12}$, the residual analysis diagram shows that the fit degree is not significantly improved, and AIC, AICc, BIC, σ^2 are more inclined to ARIMA(1,0,1)× $(1,1,1)_{12}$.

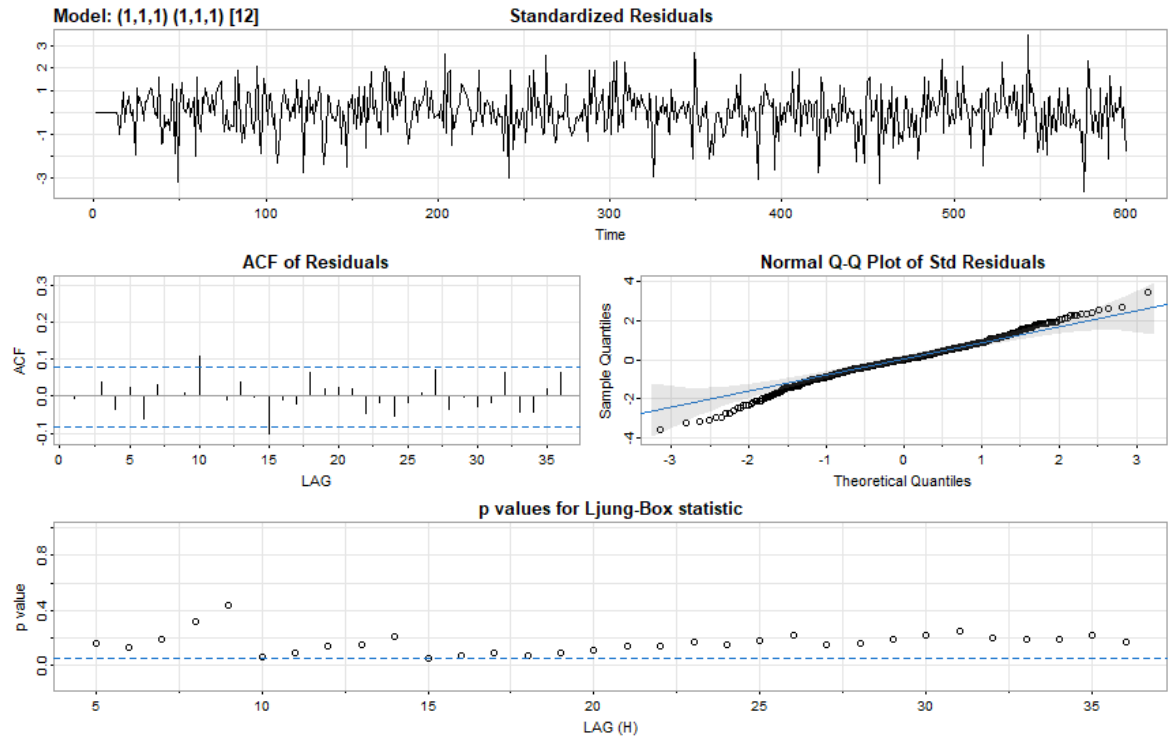

Figure 6. ARIMA(1,1,1)× $(1,1,1)_{12}$ Residuals Analysis Diagrams

3.3 Model Prediction

ARIMA(1,0,1)× $(1,1,1)_{12}$ is used to forecast the monthly average temperature of Baotou City from January to May 2023 to verify the accuracy of the model.

First, the monthly average temperature data of Baotou City from January to May 2023 were downloaded and processed in the same way as the data download and processing mentioned above, as shown in Figure 7:

	Date	average_temp
1	2023-1	-11.841475
2	2023-2	-5.611340
3	2023-3	1.879621
4	2023-4	7.815228
5	2023-5	15.078037

Figure 7. Monthly average temperature data from January to May 2023 in Baotou City

The sarmi.for () function is used to predict the average temperature for the next five months, with the following results:

```
> sarima.for(month_average$average_temp,5,1,0,1,1,1,1,12)
$pred
Time Series:
Start = 601
End = 605
Frequency = 1
[1] -13.1706388  -8.3352931  -0.4575202   8.2737385  15.4183632
```

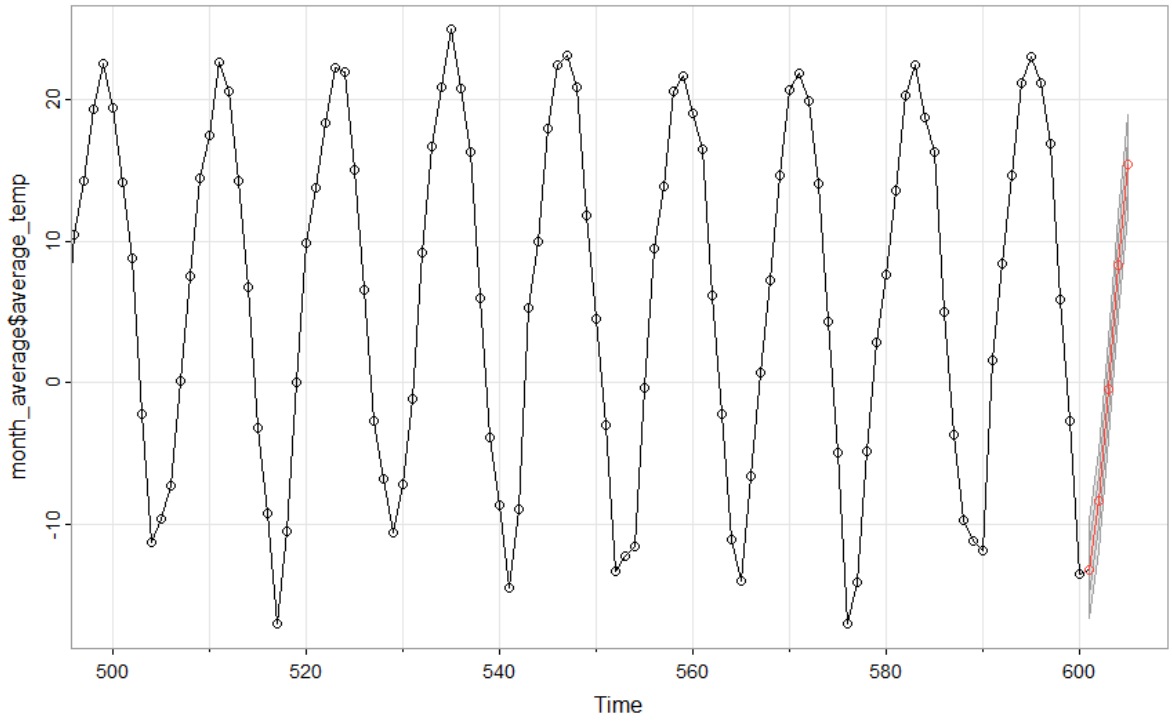

Figure 8. ARIMA(1,0,1)× $(1,1,1)_{12}$ model is used to forecast the monthly mean temperature data of Baotou City in the past five months

Compared with the original data, it can be seen that the fitting error of air temperature data is small. It can be seen from the figure that the predicted future air temperature data can basically maintain the same seasonality as before.

It is found that MSE is a relatively small value, so it can be considered that the prediction effect of the model is better.

```
> mean((test$average_temp - pred_value$pred)^2)
[1] 2.994976

> pred_value<-sarima.for(month_average$average_temp,5,1,0,1,1,1,1,12)
> mean((test$average_temp - pred_value$pred)^2)
[1] 2.994976
> pred_value2<-sarima.for(month_average$average_temp,5,1,1,1,1,1,1,12)
> mean((test$average_temp - pred_value2$pred)^2)
[1] 3.194572
```

Figure 9. ARIMA(1,1,1)× $(1,1,1)_{12}$ and ARIMA(1,0,1)× $(1,1,1)_{12}$ MSE value contrast

As can be seen from Figure 9, the MSE of ARIMA(1,1,1)× $(1,1,1)_{12}$ is larger than that of ARIMA(1,0,1)× $(1,1,1)_{12}$, which further confirms that ARIMA(1,0,1)× $(1,1,1)_{12}$ is a better model.

4. Conclusion and Prospect

In this paper, the monthly mean temperature of Baotou City from January 1, 1973, to December 31, 2022, is analyzed and modeled in time series. Through the analysis in this paper, the monthly mean temperature data of Baotou City can be predicted by the following models:

$$x_t = 0.2859x_{t-1} - 0.1122w_{t-1} - 1.0000x_{t-12} - w_{t-12} + 0.0037$$

Where x_t represents the predicted value at time point t, x_{t-1} and x_{t-12} represent the original observations at time points t−1 and t−12, respectively. w_{t-1} and w_{t-12} represent the residual term at time points t-1 and t-12, respectively.

The analysis in this paper still has some shortcomings, such as only roughly determining the order of the model without in-depth exploration of the trend of the data, and the prediction of future temperature is not accurate enough, and there are still errors. It can be seen from the coefficients of the model that the coefficients of AR term and MA term are relatively small, while the coefficients of seasonal MA term are close to -1, indicating strong seasonal changes. At the same time, the coefficient of the constant term is also relatively small, but the p-value is 0, indicating that the influence of the constant term is significant. The defect of this model is that the prediction accuracy of the model may be affected by seasonal changes, because the seasonal MA term coefficient is close to -1, and if there is a large deviation in seasonal changes, the accuracy of the prediction results may be affected. Of course, this also needs to be assessed according to the actual situation. In addition, there may be other unknown factors affecting the model, and more data and feature engineering are needed to further improve the accuracy of the model.

In view of these shortcomings and deficiencies, more rigorous methods can be selected to determine the order of the model, and other feature variables can be considered, such as increasing the weather data involved in training the model. In addition, more complex models, such as deep learning models, can be explored to predict time series data. At the same time, it is also possible to try different seasonal parameter combinations to obtain better fitting results, discuss the trend problem in the data in detail instead of ignoring it, and choose a better model for prediction to reduce the error.

References

Robert H. Shumway, (2022). *Time Series: A Data Analysis Approach Using R.* China Machine Press.

Wang Yan, (2015). *Time Series Analysis — Based on R.* China Renmin University Press.

Appendix

code

```
library (worldmet)

library (ggplot2)
```

```
library (astsa)
library (tidyverse)
library (forecast)
library (tseries)
#getMeta (site = “BAOTOU”)
getMeta (lat = 41, lon = 110)
data<-importNOAA (code = “533520-99999”, year = c (1973,1974:2022))
write.csv (data, file = “Linlin Gong_11_baotou.csv”, row.names = FALSE)
data<-read_csv (“Linlin Gong_11_baotou.csv”)
#temperature
temperature<-data[,c(1,9)]
skimr::skim (temperature)
temperature<-na.omit (temperature)
skimr::skim (temperature)
temperature <- temperature %>%
  mutate (Month = lubridate::month (date))
temperature <- temperature %>%
  mutate (Year = lubridate::year (date))
month_average <- temperature %>%
  group_by (Year, Month) %>%
  summarise (average_temp = mean (air_temp, na.rm = TRUE))
month_average <- month_average %>%
  unite (Date, Year, Month, sep = “-”)
ggplot (month_average, aes(,average_temp))+
  geom_boxplot()
tsplot (month_average$average_temp)
acf2 (month_average$average_temp)

tsplot (diff(month_average$average_temp,12))
acf2 (diff(month_average$average_temp,12))
sarima (month_average$average_temp, p=1, d=0, q=1, P=1, D=0, Q=1, S=12)
sarima (month_average$average_temp, p=1, d=0, q=1, P=1, D=1, Q=1, S=12)
sarima (month_average$average_temp, p=1, d=1, q=1, P=1, D=1, Q=1, S=12)

#test
test<-importNOAA (code = “533520-99999”, year = 2023)
test<-test[,c(1,9)]
test<-na.omit(test)
test <- test %>%
  mutate (Month = lubridate::month(date))
test <- test %>%
  mutate (Year = lubridate::year(date))
test <- test %>%
  group_by (Year, Month) %>%
```

```
  summarise (average_temp = mean(air_temp, na.rm=TRUE))
test <- test %>%
  unite (Date, Year, Month, sep = "-")
test<-test[1:5,]
write.csv (test, file="test.csv", row.names = FALSE)
pred_value1<-sarima.for (month_average$average_temp,5,1,0,1,1,1,1,12)
mean ((test$average_temp - pred_value1$pred)^2)
pred_value2<-sarima.for (month_average$average_temp,5,1,1,1,1,1,1,12)
mean ((test$average_temp - pred_value2$pred)^2
```

The data of the training model and the test data are detailed in Annex 1 and Annex 2 respectively.

Annex 1: Linlin Gong_11_baotou.csv

Annex 2: test.csv