

Analysis of Average Air Temperature in New York based on SARIMA Model

Yuning Dong^{1, *, †}, Zihua Li^{1, †} and Wentian Zhong^{1, †}

¹ School of Stony Brook Institute, Anhui University, Hefei, Province 230039(post), China

* Corresponding Author Email: r22014004@stu.ahu.edu.cn

[†]These authors contributed equally.

Abstract. Temperature prediction is one of the most vital part of weather forecasting. The temperature change has a great impact on agricultural, environmental evolution, human activities etc. The monthly average temperature of the Central Park Meteorological Observatory from 1869 to 2020 is the data to get sorted and analyzed. The SARIMA model is selected to forecast and analyze the monthly mean temperature of New York City. The first step is using the ADF test to check whether the original time series are basically stable, and through the test, the period was obviously 12. Then, by comparing with the data of first-order difference and first-order seasonal difference, the AIC value of the original time series was the smallest, it is no need to carry out the difference. The function in R is used to determine the order, and it is found that SARIMA (1,0,2) (2,1,0,12) model has the best fitting effect. Meanwhile, the residual is no autocorrelation of white noise. Finally, the average monthly temperature from 2010 to 2020 is predicted based on the original data. The difference between the predicted temperature and the actual temperature in most months is less than 5%, and the model is basically consistent with the reality. However, during the counting of the average annual temperature, the prediction results are almost unchanged after using the monthly data. This situation is quite different from the current state.

Keywords: Time series forecasting; SARIMA model; air temperature.

1. Introduction

Global warming has been an eye-catching topic. In recent years, people have realized the necessity to curb global warming with numbers of problems coming, such as the sea level rise caused by melting glaciers. At the Paris Climate Agreement, 178 Parties from all over the world signed the agreement on climate change [1]. According to incomplete statistics given by the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report, the average temperatures of the land and the sea ocean have risen by approximately 0.85 degrees Celsius [2]. Obviously, it is worth for the social society to pay more attentions.

Although the trend of global warming is inevitable, the exact degree of temperature increase will not be the same in different geographical regions, so it requires independent analysis and prediction. As the concern of the Climate Crisis, the average annual temperature in New York turns into an orientation of scientific research. Predictions will be given about the future average temperature trend at first. Besides, It is vital to confirm whether there is an affiliated connection between the increasing annual temperature and the human industrialization in the past century. The selection of New York meteorological data to analyze is mainly due to New York's international leading level of the climate recording data [3].

Temperature change is usually influenced by a combination of factors, such as the intensity of solar radiation [4], the effects of ocean currents [5], Urban or rural (affected by heat island effect) [6]. So it is important for the researchers to process the various data which have been mentioned or unmentioned. By searching data and relevant professional background, this is widely believed that the research of temperature change can be continued from the perspective of professional meteorology and the statistical knowledge to construct statistical models for prediction and analysis. With the advent of the information age, computer technology and meteorological remote sensing enable meteorologists to obtain more accurate and time-sensitive data, and also promote the progress

of meteorology. For instance, the knowledge of meteorology given by the institution: National Oceanic and Atmospheric Administration shows the detailed topic: NOAA Climate Data Record (CDR) of AVHRR Daily and Monthly Aerosol Optical Thickness over Global Oceans, and NOAA Climate Data Record (CDR) of Atmospheric Layer Temperatures [7]. These are all professionalism studies that rely on specific instruments and even specialized weather stations, which means that only professional meteorologists can do it correctly. Continually the statistical models will be emerged below.

Since it is affected by many factors, multiple linear regression problem and ARIMA model can be taken into account to deal with it. However, it is difficult to give a correlation equation between the variables affecting temperature. Comparatively speaking, the use of the time series method is more feasible [8]. In short, a time series is a list of valid data recorded in chronological order. By observing, testing, analyzing and studying these data, the law of change can be found and the predictions of its future trend can be viewed. The preprocessing of time series can be divided into autocorrelation and partial autocorrelation, both of which are research tools to study the characteristics of time series. When the characteristics of time series are analyzed, time series has randomness, stationarity and seasonality. Generally speaking, stationary time series can be divided into two categories: strict stationary and wide stationary. In practice, it is very difficult and impractical to achieve strict stability. Generally, wide stationary time series with more relaxed conditions are used. Given by the definition of wide-stable time series: as long as the lower moments of the series are stable, the basic properties of the series can be made to be stable. For the other two properties, seasonality and trend coexist in real problems. Only after eliminating the trend of sequence, its seasonality will be able to be determined, otherwise there will be errors. The stationary sequence obtained after these two steps is called differential stationary sequence.

After introducing the time series this scientific research will use, the specific research idea is briefly given: After testing and analyzing the stationarity of the annual average temperature in New York, the difference operation is used to eliminate the trend of the series in the case of instability. Secondly, the Autocorrelation and Partial Autocorrelation diagrams [9] are used to determine the model parameters, and the ARIMA model [10] will be performed by a fitting test, getting the residual sequence and white noise test [11]. The ARIMA model can be eventually obtained through the detection results. Finally, the certified ARIMA (p,d,q) is permitted to be used to predict the annual average temperature of New York City over the next five years. However, ARIMA model does not predict the periodic data well, so we make seasonal difference for the data with obvious seasonality and use the SARIMA model [12] which can be described as SARIMA(p,d,q)(P,D,Q,S).

2. Methods

In determining the predictive model of time series, a typical method, Seasonal Autoregressive Integrated Moving Average(SARIMA) model is selected because of its Feasibility and high accuracy. After the raw data is preprocessed, several tests are conducted and orders of this model is determined. The final results are produced by SARIMA model.

2.1. Time Series

As mentioned in the introduction above, Time Series is a list of valid data recorded in chronological order. By observing, testing, analyzing and studying these data, the law of change can be proved and its future trend is able to be predicted. Stationarity is the basis of time series. It can be distinguished by direct observation of graph changes, ACF graph and ADF test.

To represent it graphically, Here are some statistical definitions of autocorrelation and non-autocorrelation: Set the stationary time series $\{y_t, t \in T\}$, Arbitrary select the $t, t+k \in T$, define $\gamma(k) = \gamma(t, t+k)$, It can be seen that the $\gamma(k)$ as the Delayed K-order autocovariance function. And by nature, when $k = 0, \gamma(0) = \gamma(t, t) = Dy_t$ is the constant value.

Similarly there is the Delayed K-order autocorrelation coefficient ρ_k of the $\{y_t, t \in T\}$. If the $\{y_t, t \in T\}$ is stationary, $\gamma(k)$ is the Delayed K-order autocovariance function, so the $\rho_k = \frac{\gamma(t, t+k)}{\sqrt{Dy_t} \sqrt{Dy_{t+k}}} = \frac{\gamma(k)}{\gamma(0)}$.

Finally, according to the corresponding autocorrelation graph, considering in a long-term trend, when the autocorrelation function rapidly decays to near 0, it can be judged that it is stable. If there is a periodic decay or does not fade rapidly to zero, the data is non-stationary.

2.2. Model Construction

The model which has the structure of $y_t = \phi_0 + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$ is called ARMA model. This also means that it is composed of autoregressive and moving average. After adding the delay operator B, $ARMA(p, q)$ can be written as $\varphi(B)y_t = \Theta(b)\varepsilon_t$, the former part is the AR(p), and the latter part is MA(p). The stationarity of $ARMA(p, q)$ process depends only on its autoregressive part, the values of $\varphi(B)y_t = 0$ are all in the unit circle, while the values of $\Theta(b)\varepsilon_t = 0$ should be totally outside the unit circle. And $ARMA(p, q)$ model corresponds to time series data whose mean value is 0, if the average value is not 0, execute first difference to the whole data. Finally, the order of $ARMA(p, q)$ model is determined by observing autocorrelation coefficient diagrams and partial autocorrelation coefficient diagrams.

However, some time series data have obvious periodic changes, which are often caused by seasonal changes which include some weekly, monthly and quarterly inherent factors. Such series are called seasonal time series. For example, the series of temperature values in an area (which is observed every hour) contains annual as well as daily variations. If only ARIMA or ARMA model is used to deal with seasonal time series, it will fail to reflect the seasonality of the series and only reflect the general trend, and resulting in distorted prediction results. Therefore, seasonal time series model is needed, which is represented by SARIMA.

The seasonality of a time series is the recurring occurrence of certain features over a period of time. In practical problems, seasonality and trend coexist. After eliminating the trend of the sequence, seasonality of the time series can be obtained by testing whether there is a significant difference between the autocorrelation coefficient under different values of k and the number 0.

Firstly, time series with non-stationary can be changed into a stationary one through a method called d-order difference, then $ARMA(p, q)$ model is used for processing, and the whole model is denoted as $ARIMA(p, d, q)$ model:

$$\Phi_p(L)\Delta^d x_t = \Theta_q(L)\varepsilon_t \quad (1)$$

Let the variation period of seasonal time series be s . First, seasonal-difference method is used to eliminate periodic variations. The seasonal difference operator is defined as

$$\Delta_s^D y_t = (I - L^s)y_t = y_t - y_{t-s} \quad (2)$$

For seasonal time series with non-stationary, D seasonal differences are made before it can be converted into stationary series. On this basis, a model called P order autoregressive Q order Seasonal Autoregressive Integrated Moving Average model with period s can be established as (3).

$$A_p(L^s)\Delta_s^D y_t = B_q(L^s)u_t \quad (3)$$

Formula (3) is equivalent to assuming that u_t is non-autocorrelation and stationary. When u_t is not stationary and autocorrelation exists, u_t can be described as

$$\Phi_p(L)\Delta^d u_t = \Theta_q(L)\varepsilon_t \quad (4)$$

Among this, ε_t means a white noise process, p and q represents the maximum lag order of non-seasonal autoregressive and moving-average respectively. d represents the first order (non-seasonal) difference number of u_t . From the above equation

$$u_t = \Phi_p^{-1}(L) \Delta^{-d} \Theta_q(L) \varepsilon_t \quad (5)$$

By substituting equation (5) into equation (3), the general expression of seasonal time series model can be obtained:

$$\Phi_p(L) A_p(L^S) \Delta^d \Delta_S^D y_t = \Theta_q(L) B_Q(L^S) \varepsilon_t \quad (6)$$

Among this, q and q , P and Q express non-seasonal and seasonal autoregressive and moving-average with the maximum lag order respectively. d and D explain non-seasonal and seasonal difference number respectively. The above formula is called the $(p, d, q) \times (P, D, Q, S)$ order seasonal time series model.

3. Results and Discussion

3.1. Data Source and Preparing

To make air temperature predictions at Central Park, which is located in the middle of Manhattan, New York, USA, we gathered and selected data in the series of monthly means for each month and annual means from 1869 to 2019. In the process of gathering data, there are several official websites available for checking air temperature at Central Park and getting precise data series (<https://www.weather.gov/okx/CentralParkHistorical>). In this paper, considering the sheer volume of data, only part of specific values has been shown. The full annual average temperature and monthly average temperature are demonstrated in form of pictures. As shown in Figure 1, it can be seen from the diagram that the data is not stable but fluctuates and rises.

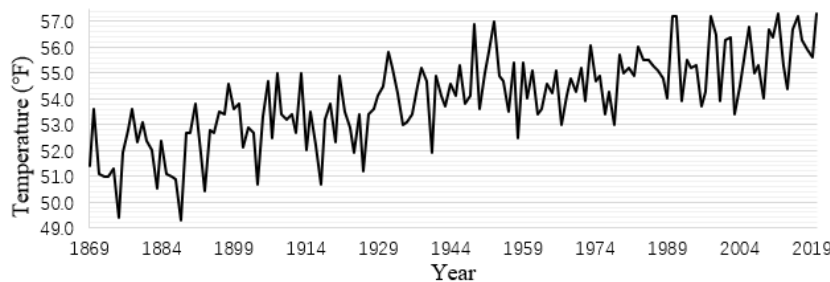


Fig. 1 Annual Average Air Temperature (°F) at Central Park

Without ignoring the influential factors of ecological environment, global climate warming and other uncertain factors, it can be found in the figure above that the change of temperature is characterized by volatility and seasonality. Moreover, from 1869 to 2019, the temperature obviously shows an overall upward trend, which as a whole shows a gradual increase in lowest temperature and highest temperature every seven years respectively, with troughs rising from 49.3 to 55.6 and peaks rising from 53.6 to 57.3.

Among 1869 and 2019, the air temperature means of every month at Central Park are shown in Figure 2. Different from annual means, the monthly average temperature is affected by the seasons and fluctuates within a certain range in each month. Different from annual average temperature, the seasonal fluctuation of monthly means are more regular, obvious and organized, because the effects, such as political factors, global climate change and human activities, have less influence on monthly means.

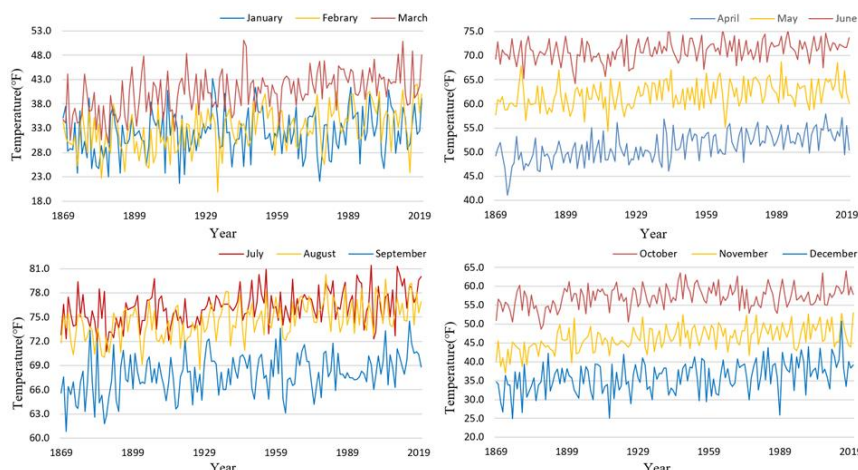


Fig. 2 Monthly Average Air Temperature(°F) at Central Park

By contrast, in order to demonstrate the general characteristics of air temperature and consider the characteristic of seasonal trend, monthly means air temperature are chosen finally as the basic data for modeling and prediction.

3.2. Modeling Process Overview

In order to make statistical prediction and decision on air temperature at Central Park, New York, as shown in the framework of Figure 3, we firstly run a stationary test on the unseasonal part of data depicted above. If it does not pass the test, the method of n-order difference on this unseasonal part is conducted until it passes the test. For the seasonal part which have a cyclical trend, seasonal stationary test is implemented until the data have no correlation with each other. Next, white noise test is checked to see if the data is suitable for modeling. Passing all these tests, a mathematical model called Seasonal Autoregressive Integrated Moving Average (SARIMA) is built to give a relatively precise prediction on the monthly average air temperature during 2010-2019. To order the model, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are two efficient methods that used to determine the orders. In the meantime, residual white noise test is utilized to check the normal distribution and correlation of the residuals. Predicted results output by the model is compared with the actual air temperature series collected to justify its accuracy, validity and reliability. Lastly our model is used to forecast both monthly and annual average air temperature in year 2021, 2022 or more.

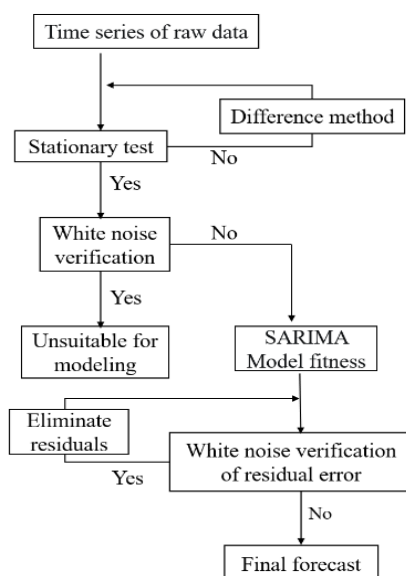


Fig. 3 flow chart of constructing SARIMA Model

3.3. Model Ordering

Before constructing mathematical model, firstly stationary test is conducted. If test shows that original data series is not stationary, n-order difference method is utilized until a stationary data series is ensured. Otherwise, the next step is proceeded. In the model of ARIMA (p, d, q)(P,D,Q,S), the value of these seven parameters is determined by ACF, PACF and minimization AIC, SBC criteria. When ordering, ACF and PACF is utilized and it is estimated that for unseasonal part, ACF is truncation in order 2 and PACF is truncation in order 1 with no need of difference processing. For the seasonal part, ACF is trailing and PACF is truncation in order 2 with seasonal difference once.

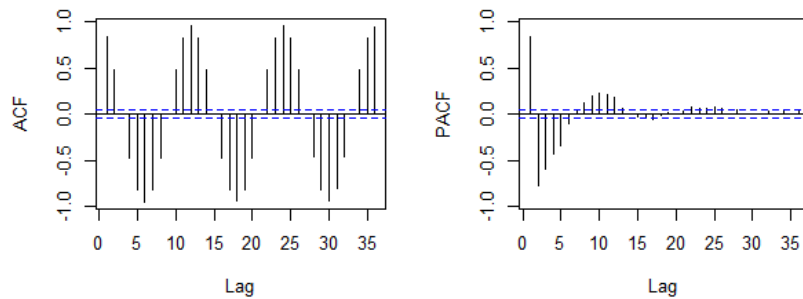


Fig. 4 Initial monthly temperature and ACF, PACF plots

After running Augmented Dickey-Fuller test, the p-value of primary data is 0.01, indicating a stationary time series. On the other hand, it can be seen in ACF and PACF showed in Figure 4 that the data series is stationary. Moreover, according to minimization AIC criteria, the best model (p,d,q) = (1,0,2) is selected with minimization AIC value 8837.75. It is set that value p, which is determined by PACF plot, is 1. Value q, which is determined by ACF plot, is 2 without differential processing.

However, It is demonstrated that the data series has a strong periodicity, with a lag of 12. For the seasonal part, seasonal difference is necessary to derive a stationary series and make predictions. It can be seen in Figure 5 that after seasonal 1st difference, PACF values fall outside the confidence limit at lags of 12, 24, and is basically stable at lag 36, so it is determined that $P = 2$. ACF value is large at lag 12, falling outside the confidence interval, and then is basically stable at 24 and 36. Taking the best fitness effect and minimization AIC criteria into account, $Q=0$ is selected.

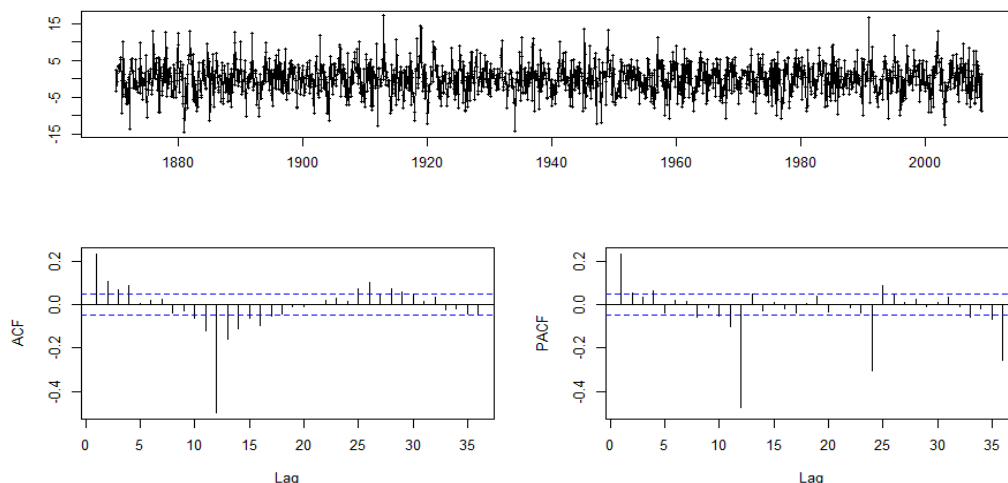


Fig. 5 Data Series after first-order difference

Then white noise test, which is also called pure randomness test, is implemented to see if the data series chosen is worth to make predictions. During this test, method called Box-Pierce test is conducted and it is resulted that p-value is less than 2.2×10^{-16} . It indicates that the data series passed white noise test and is suitable to fit the order of the model and predict the future value.

Finally, ARIMA model is ordered with (p,d,q)(P,D,Q,S)=(1,0,2)(2,1,0,12). It shows the best fitness result without approximations and its accuracy is shown in Table 1.

Table 1. accuracy indicators of ARIMA (0,1,1)

Indicators	ME	RMES	MAE	MPE	MAPE	MASE	ACF1
Results	0.030	3.386	2.614	-0.505	5.911	0.796	-0.001

In order to carry out the next model prediction, whether the residual errors are correlated is judged, because the model prediction can only be carried out when there is no correlation. As shown in Figure 6, it is indicated that the Q-Q plot has a high degree of fit with its addition line, so the residual errors follow a normal distribution. In addition, through the residual white noise test data, p-value is equal to 0.979, indicating that there is no correlation between residual errors.

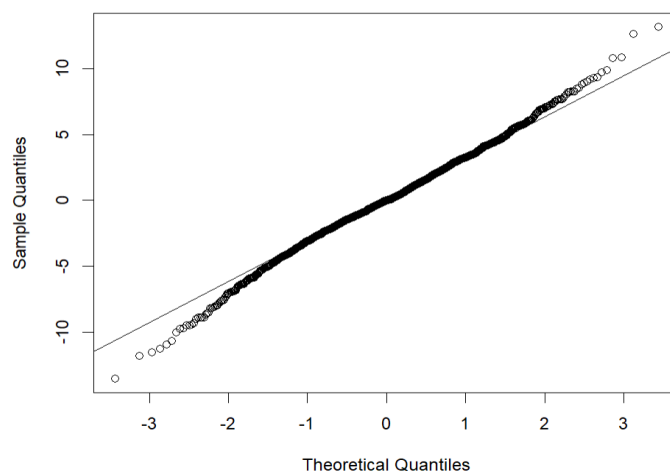


Fig. 6 Correlation test of residual errors

1.1 Forecasting Results

The best fitting model, ARIMA(1,0,2)(2,1,0,12) is utilized to make predictions on air temperature of every month during year 2010-2019. Part of the outputting results by this model is given in table 2. By comparing with actual air temperature in New York, the accuracies of every month are given below, which shows that predictions are close enough to the actual values. It is stipulated that

$$\text{Relative Error} = \frac{|\text{Actual Value} - \text{Predicted Value}|}{\text{Actual Value}} \times 100\%.$$

Table 2. Partial comparison between prediction results and actual value

Time(Month-Year)	Predictions(°F)	Actual Timperature	Relative Error
January-10	33.78	32.5	3.95%
February-10	31.57	33.1	4.61%
March-10	42.13	48.2	12.60%
April-10	52.72	57.9	8.95%
May-10	62.52	65.3	4.26%
June-10	72.39	74.7	3.09%
July-10	76.76	81.3	5.58%
August-10	74.06	77.4	4.31%
September-10	69.17	71.1	2.72%
October-10	58.94	58.1	1.45%
November-10	46.31	47.9	3.33%
December-10	38.23	32.8	16.57%
January-11	32.78	29.7	10.38%
February-11	32.74	36.0	9.05%
March-11	42.13	42.3	0.40%
April-11	53.53	54.3	1.43%
May-11	61.64	64.5	4.43%

In figure 7, the prediction results show an obvious seasonal trend, which fluctuates with the seasons. The predicted curve almost coincides with the actual temperature, which indicates a relatively successful forecast with high accuracy. However, the peaks and troughs don't vary much in this model forecasting, and the maximum mild and minimum cold temperatures are almost the same each year.

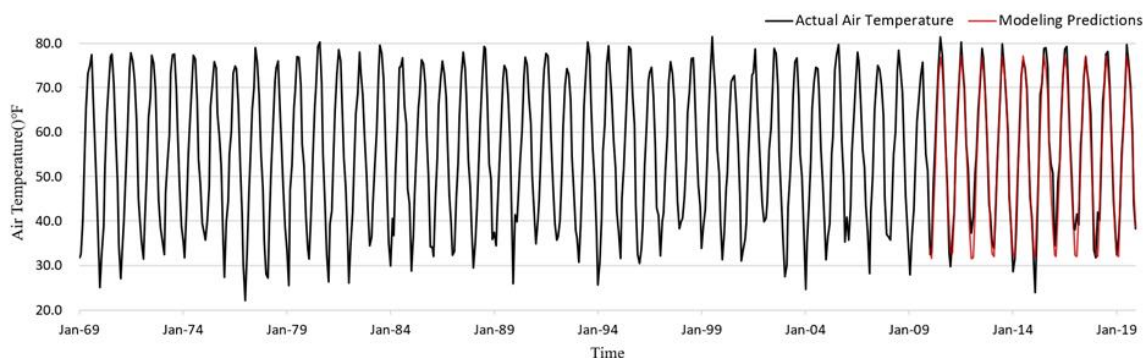


Fig. 7 air temperature forecasting of 2010-2019

Specifically, calculate an annual average using monthly data, and the results show that annual average air temperature at Central Park in 2019 is 56.26 degrees Fahrenheit. Compared with the actual history data, which is 55.6, the prediction given by the model gives a relatively accurate prediction with an accuracy of 0.95.

4. Conclusion

Ninety percent of the monthly forecasts were accurate to 0.9 or more of the actual temperature, which indicates that a successful model close to the truth is built. It is noticed that there are cycles in terms of seasons of data existence, so to eliminate seasonality, ARIMA (p,d,q) Model is upgraded to SARIMA (p,d,q)(P,D,Q,S) model, where seasonal difference is added to the previous model.

However, this model has some shortcomings in predicting the average annual temperature. In terms of these predictions by SARIMA, the fluctuation of annual mean temperature tends to be stable. The changes of predicted temperature are not significant, which range from 54.711 to 56.861, while the actual annual maximum mild minimum temperature are 57.3 and 54.4 respectively. In future studies, models such as vector autoregressive models, machine learning and neural networks can be combined with the SARIMA model to better predict the annual data.

References

- [1] Lyster R. Climate justice, adaptation and the Paris Agreement: a recipe for disasters. *Environmental Politics*, 2017, 26(3): 438-458.
- [2] Pachauri R K, Allen M R, Barros V R, et al. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Ipcc, 2014.
- [3] Edwards A, Gibson A, Kundmueller B, et al. NYCDEP Focuses on Recruitment and Talent Development//WEFTEC 2016. Water Environment Federation, 2016.
- [4] Gray L J, Beer J, Geller M, et al. Solar influences on climate. *Reviews of Geophysics*, 2010, 48(4): 1032–1047.
- [5] He J, Winton M, Vecchi G, et al. Transient climate sensitivity depends on base climate ocean circulation. *Journal of Climate*, 2017, 30(4): 1493-1504.
- [6] Huang Q, Huang J, Yang X, et al. Quantifying the seasonal contribution of coupling urban land use types on Urban Heat Island using Land Contribution Index: A case study in Wuhan, China. *Sustainable Cities and Society*, 2019, 44: 666-675.

- [7] Bates J J, Privette J L, Kearns E J, et al. Sustained production of multidecadal climate records: Lessons from the NOAA Climate Data Record program. *Bulletin of the American Meteorological Society*, 2016, 97(9): 1573-1581.
- [8] Sharpley C. Time-series analysis of counseling research. *Measurement and Evaluation in Guidance*, 1981, 14(3): 149-157.
- [9] Pincombe B. Anomaly detection in time series of graphs using arma processes. *Asor Bulletin*, 2005, 24(4):2.
- [10] Fattah J, Ezzine L, Aman Z, et al. Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*, 2018.
- [11] Bartlett M S. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 1946, 8(1): 27-41.
- [12] Zhang N, Zhang Y, Lu H. Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways. *Transportation Research Record*, 2011, 2215(1): 85-92.