Temperature Forecasts in New York City Exploiting ARIMA Model

Shaofan Du *

Department of Statistics, Case Western Reserve University, Cleveland, 44106, United States of America

* Corresponding Author Email: sxd659@case.edu

Abstract. Efforts to estimate air temperatures play an important role in predicting climate impact on ecological, industrial, and agricultural fields. Accurate temperature prediction benefits the government, industry, and the public for planning activities. Moreover, the temperature forecasts provide people with safeguard life and property. This paper conducts an exploratory data analysis of the temperature time series data, taking New York City as a representative. The time series data contains the daily temperature in New York City from 1995 to May 2020. The considerable time scope is so comprehensive that it makes a forecast based on the periodic performance similar in approximately 20 years. In this work, the autoregressive integrated moving average (ARIMA) model is leveraged to forecast future values based on past periods. Its hyperparameters are determined by autoregressive correlation function, partial autoregressive correlation function, and Bayesian Information Criterion. The ARIMA model assumes that the future performance will resemble the past, and the temperature in different time stamps is similar, which makes it a suitable model in this case.

Keywords: Temperature forecasting, ARIMA modeling, New York City, time series data.

1. Introduction

This research is necessary because big data is the most familiar friend to people nowadays, providing people with immeasurable information. Indeed, it is ambiguous for people to see a large quantity of data presented in a table and sometimes may make them dizzy to see those bunch numbers. However, achieving data analysis and visualizing the data will clarify everything. For several years, research has shown that the world temperature has been growing because of artificial destruction to the environment, such as fossil fuel combustion, deforestation, and automobile exhaust. Those damage to the atmosphere causes greenhouse gases exhausting to the air day by day and increases the world temperature [1]. Given the data that has shown that the ground temperature is continuously growing, this research chooses the temperature of New York City specifically to verify the negative effects of global warming on a metropolis, and more importantly it forecasts the future temperature of New York City using machine learning models. This research collects the time series temperature data in New York City from 1995 to 2020. It also presents the change in temperature in those years by visualizing the data. Besides that, the primary method of forecasting the future temperature is the autoregressive integrated moving average (ARIMA) model that predicts the value based on historical data. By training the model several times and modifying the hyperparameters, the final predicted temperature data has a 7.8 mean squared error, which is the least in a different tested model.

2. Theorem and Methodology

2.1. Dataset and Preprocessing

The temperature predicting analysis and forecasting uses Python programming. The temperature data comes from and is authorized by the open-source website Kaggle [2]. The data collect the city temperature around the world approximately from 1995 to 2020. Machine learning tasks of analyzing and forecasting the temperature uses several Python tools and packages: Jupyter Notebook for coding, Pandas for processing the data, Seaborn and Matplotlib for plotting, Statsmodels for building the

ARIMA models, NumPy for processing working with arrays, Math and Scikit-learn for math calculating [3].

First, use Pandas to read the downloaded data in a CSV file by dropping the index column and using the standard date time format because it will become easy to analyze by indexing the DateTime. The current raw data frame displays the average temperature around cities, see Table 1.

Table 1. Temperature data of the major cities in the world.

Region	Country	State	City	Month	Day	Year	AvgTemperature
Africa	Algeria	NaN	Algiers	1	1	1995	64.2
Africa	Algeria	NaN	Algiers	1	2	1995	49.4
Africa	Algerai	NaN	Algiers	1	3	1995	48.8
…	…	…	…	…	…	…	…

2.2. Data Cleaning and Sets Partitioning

Given the disorganized raw data, data cleaning is the first task. First, the project's mission is to forecast only the temperature of New York City, so extracting the data of New York City by choosing the rows containing the string "New York City". Secondly, dropping the abnormal values of temperature data lower than 0 because there is no evidence proving that the temperature of New York City was lower than 0 Fahrenheit for several decades. Furthermore, deleting the useless columns: "Region", "Country", "State", and "City", and renaming the column "AvgTemperature" to "AvgTemp". A short column name makes the coding easier. The final thing of cleaning data is transferring the date to DateTime format in Python. There are 9245 rows and 1 column in total. See the cleaned data in Table 2.

In addition, partitioning the train set, validation set, and test set are essential procedures in machine learning. A train set is used for training and evaluating the forecasting model, a validation set is used for controlling the sophistication of the model parameters, and a test set is used for testing the performance of the forecasting model and helping determine the final model [4]. Because the parameters of the forecasting model are easy to decide on, the validation set is not used. The train set and test set data are partitioned by date time, separately from the start of 1995 to the end of 2019 and from the start of 2020 to May 2020. The final goal is to build the predicting model by the train set and test the accuracy of the prediction by the test set.

Table 2. Cleaned temperature data of New York City from 1995 to 2020.

Datetime	AvgTemp
1995-01-01	44.0
1995-01-02	41.8
...	...
2020-05-12	51.1
2020-05-13	51.8

2.3. Data Visualization and Analysis

One of the best and easiest ways of understanding the data is data visualization, which presents what the data looks like through plotting. Figure 1 presents the visualization of the temperature data using Matplotlib and Seaborn.

Time series forecasting requires the data to be regular. Otherwise, it is not possible to forecast the data. Therefore, one of the most important tasks is determining whether the time series is stationary or not.

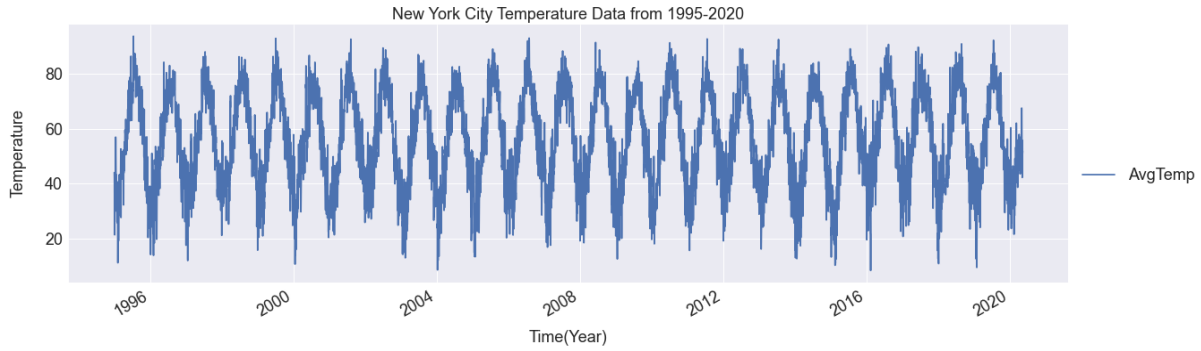


Fig 1. Temperature Data Visualization.

The trend and regularity of the temperature data shown above look accessible but still need to check its stationarity using a statistical method. In statistics, the Dicky-Fuller test tests whether the data is stationary or not. In this test, the null hypothesis is that that an an ARIMA model has unit root. The alternative hypothesis is usually stationarity or trend-stationarity. Autoregression is one component of the ARIMA model [5]. An autoregression model with a constant can be written as:

$$y_t = a_0 + \delta y_{t-1} + u_t \tag{1}$$

Where $y_t$the temperature as a variable is, a_0 is the constant, u_t is the error term. Whether a unit root exist or not could be exploited for estimating the model, equivalent to testing δ=0. It is easy to achieve the Dicky-Fuller test in Python by importing "adfuller" from Python Statsmodels and setting the rolling mean and rolling standard deviation by seven days to remove errors and absoluteness in the data as much as possible [6].

2.4. Modeling and Hyperparameters Selection

ARIMA model is constituted by the autoregressive model (AR model) and moving average model (MA model). AR model explores the relationship between the current and previous values, and further forecast the variable according to its previous values. It is necessary to satisfy the prerequisite of data stationarity. The autoregressive function of p order is denoted as:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{2}$$

Where y_t denotes the current value of the variable, which is the average temperature in this case, μ is the constant, p is order value, γi is the autoregressive coefficient, and ϵ_t denotes the error. The autoregressive function describes the current value y_t by accumulating the historic value y_{t-i} times the coefficient for p orders of data in the time series. On the other hand, the MA model focuses more on the summation of the error term in the model, whose function of q order is denoted as:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{3}$$

Where ϵ_t is the error of the current value, ϵ_{t-i} is the error of the previous values. The moving average function can effectively eliminate the random fluctuation of the time series. Finally, if integrating the two models after using the difference method, it will become the ARIMA model, whose function can be written as:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{4}$$

The principle of the ARIMA model is now apparent. It changes the series into stationary series and then determines the dependent variable based on the current, previous values and the error.

The next step is determining the hyperparameters, which is the orders in the model. It could be achieved by the partial autoregressive correlation function (PACF) and autoregressive correlation function (ACF), denoted as:

$$Corr(y_t, y_{t-1}) \tag{5}$$

The two fucntions determine the hyperparameters p, q, and d in ARIMA model. When the points of PACF locate in the confidence level after p order but those of ACF do not, so choose AR model. When the points of ACF locate in the confidence level after q order but those of PACF do not, so choose MA model. Therefore, p order can be seen in the PACF, and q order can be seen in the ACF. The order of differencing d can be determined by the time of differencing to the time series [7].

3. Results and Discussion

3.1. Results of Data Visualization

Figure 2 displayes the the result of the visualization in that there is no obvious and intense fluctuation between the periods in those years. The average daily temperature conspicuously increases to approximately 80 during summer and decreases to 20 to 30 in winter. Therefore, there is no obvious global warming sign according to the time series data, at least in New York City.

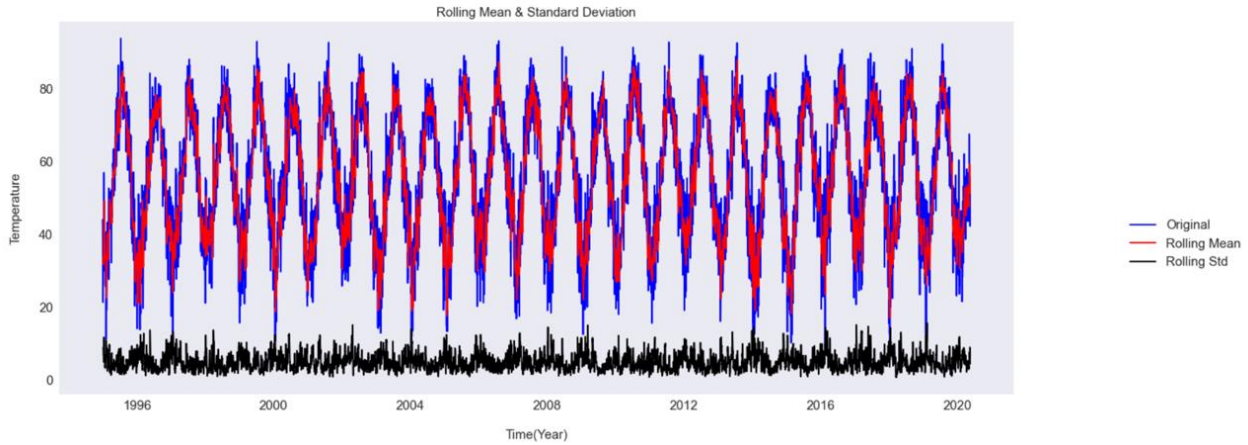


Fig 2. Dicky-Fuller time series stationary test plot.

3.2. Results of Dicky-Fuller Test

The rolling mean and rolling standard deviation is plotted in Figure 2. In the summary of results, P-value is much lower than 0.05 for 95% confidence level, and the Dicky-Fuller t statistic is -6.9 much lower than the 1%, 5%, and 10% critical values in different degrees of rejecting the null hypothesis, which indicates that it is reasonable to reject the null hypothesis, which means there is a unit root in the model. Considering the time series is stationary, it is not necessary to stabilize the data using the difference method but building the ARIMA model using it directly.

Table 3. Results of Dicky-Fuller Test.

Test Statistics	-6.900383E+00
P-value	1.286578E-09
# lags used	3.800000E+01
NO. of observation used	9.206000E+03
Critical value (1%)	-3.431061E+00
Critical value (5%)	-2.861854E+00
Critical value (10%)	-2.566937E+00
dtype: float 64	
Reject H0, that the data has no unit root, the data is stationary.	

3.3. Model Hyperparameters Selection

In this case, d is 0 because the data is stationary without any differencing. The plots of ACF is demonstrated in Figure 3 and the PACF is shown in Figure 4. In ACF, the points will never locate in the confidence level (represented by the deep blue area in the figures). However, the points locate in the confidence level after order 2 or maybe 10. As far as so, the hyperparameters are partially determined: p is between 2 and 10, both q and d are 0. The suitable model is AR model.

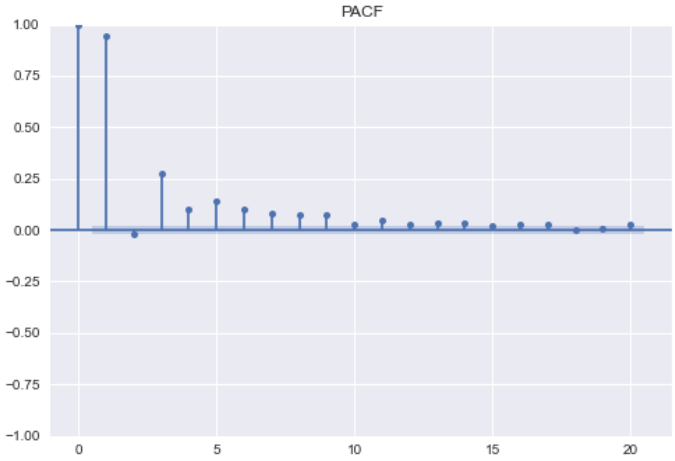


Figure 3. Autoregressive correlation function plot.

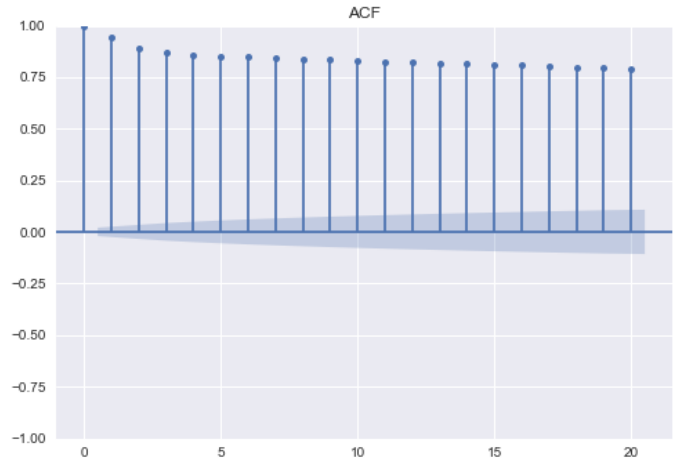


Figure 4. Autoregressive correlation function plot.

But right now, what is the best value for p? Bayesian Information Criterion (BIC) [8] can help us to determine that, whose function is denoted by:

$$BIC = kln(n) - 2\ln(L) \tag{6}$$

Where k, n and L respectively denotes the number of hyperparameters, the number of data samples, and the likelihood function. BIC helps to find the best hyperparameters in the model by weighing the likelihood function and the number of data samples, and models with lower BIC is preferred. Due to the bigger of the likelihood function the better, the lower k makes BIC lower. A histogram is plotted presenting the BIC in different p order in Figure 5.



Figure 5. Bayesian Information Criterion values in different parameters.

The result shows that BIC is lower when p is equal to 9 or 10. The next step is building two models with two different p order values, which are 9 and 10, and training the data. The two training summary results are shown in both the Table 4 and the Table 5.

Table 4. SARIMAX summary results of ARIMA model when p is 9.

	Dependent Var:	AvgTemp		# Observations:	9245	
	Method:	ARIMA(9,0,0)		Log Likelihood	-28411.27	
	Sample:	9245		AIC	56844.54	
	Covariance:	opg		BIC	56922.99	
				HQIC	56871.19	
	coefficient	std err	z	P<|z|	0.025	0.975
const				0.000	50.86	61.087
L1	0.8921	0.009	98.68	0.000	0.87	0.910
L2	-0.2802	0.012	-22.98	0.002	0.01	0.064
L3	0.1789	0.013	13.30	0.089	-0.04	0.003
…	…	…	…	…	…	…
sigma2	27.3342	0.335	81.56	0.000	26.677	27.991
Ljung-Box (Q) :		0.03		Jarque-Bera (JB):		618.70
Prob (Q) :		0.86		Probability (JB) :		0.00
Heteroskedasticity (H) :		1.04				
Probability (H) :		0.24				

Table 5. SARIMAX summary results of ARIMA model when p is 10

	Dependent Var:	AvgTemp		# Observations:	9245	
	Method:	ARIMA(10,0,0)		Log Likelihood	-28407.618	
	Sample:	9245		AIC	56839.237	
	Covariance:	opg		BIC	56924.819	
				HQIC	56868.319	
	coefficient	std error	z	P<|z|	[0.025	0.975]
const	55.975	2.680	20.887	0.000	50.723	61.227
L1	0.8901	0.009	98.308	0.000	0.872	0.908
L2	-0.2805	0.012	-23.040	0.002	0.015	0.064
L3	0.1781	0.013	14.242	0.074	-0.048	0.002
…	…	…	…	…	…	…
sigma2	27.3131	0.335	81.638	0.000	26.657	27.969
Ljung-Box (Q) :		0.01		Jarque-Bera (JB):		616.15
Prob (Q) :		0.91		Probability (JB) :		0.00
Heteroskedasticity (H) :		1.04				
Probability (H) :		0.26				

3.4. Interpreting the SARIMAX Results

AIC represents Akchi's information standard [9]. It is an indicator that helps assess model strength, like BIC. The maximum likelihood values and number of parameters are also taken into consideration. The large number of parameters is likely to boost the maximum likeluhood value. The ACI is helpful for decreasing the overfitting problem, by penalizing the number of parameters. By using this technique, the model with few parameter but good performance is obtained. HQIC denotes the Hannan–Quinn information criterion [10], which is not used for parameter selection as frequently as AIC or BIC.

The "stderr" column estimates the error in the predicted value. It shows how much the residuals affect the estimated parameters. "z" equals the coefficient value divided by the standard error. The p-values are below the given threshold (usually 0.05 significant level), so the model is good.

The Ljung-Box test is proposed for examing whether time series autocorrelations is 0, which tests for overall randomness based on multiple lags [11]. The null hypothesis and alternative hypothesis are usually given by: H0: The data are independently distributed. H1: The data are not independently distributed; they exhibit serial correlation.

The test could be:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{7}$$

Where n, $\hat{\rho}_k^2$, and h are the sample size, the sample autocorrelation at lag k, and the number of lags being tested. Usually, the rejection threshold level α is:

$$Q > \chi_{1-\alpha,h}^2 \tag{8}$$

Where $\chi_{1-\alpha,h}^2$ is the$(1-\alpha)$ quantile for the chi-squared distribution with h degrees of freedom. In ARIMA model, the $h-p-q$ should be considereed as the degree of freedom.

Therefore, for p is equal to 9, the degrees of freedom is 9236; and for p is equal to 10, the degrees of freedom is 9235. The calculated chi-squared value is 36028.78 and 36028.80 separately, which are both larger than the calculated Ljung-Box in summary table 0.03 and 0.01 so that it is reasonable to not reject the hypothesis of randomness. The data are independently distributed.

3.5. Forecasting

Now, let's predict the temperature data and compare each of them to the data in test set and plot the results. See the plotting results in Figure 6 with p equals to 9 and in Figure 7 with p equals to 10.

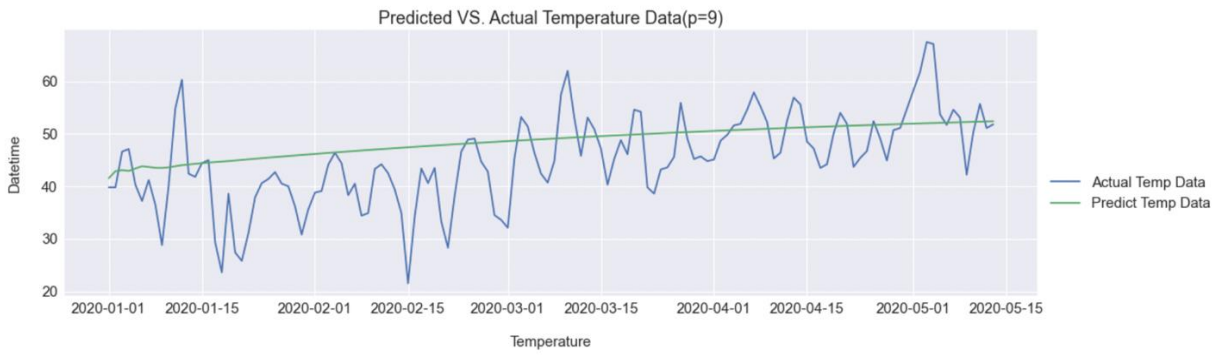


Figure 6. Predicted versus actual temperature data when p equals to 9.

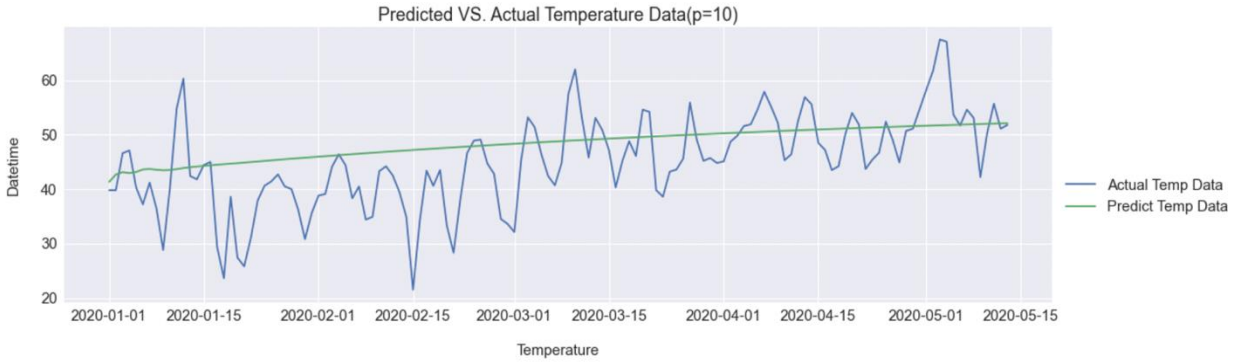


Figure 7. Predicted versus actual temperature data when p equals to 10.

Mean squared error evaluates the goodness of fit [12]. Therefore, the lower MSE is preferred. By calculating that of the two predicting models, the MSE of the model with p equals 9 is 7.9583, and the MSE of the model with p equals 10 is 7.8687. The AR model with hyperparameters of p is 10 and both of q and d are 0 is more accurate for forecasting. Finally, the low median value of the difference between the predicted and actual temperature in the Excel sheet shows the forecasting is precise thanks to the AR model.

Table 6. Actual and predicted temperature value.

Datetime	AvgTemp	PredTemp	Absolute Difference	Absolute Deviation from the Actual Value (σ)
2020/1/1	39.80	41.32	1.52	3.78
2020/1/2	39.80	42.68	2.88	2.42
2020/1/3	46.60	43.11	3.49	1.99
...	...	…	…	…
Min	21.50	41.32	0.13	0.008507463
Max	67.50	52.11	25.69	7.011492537
Median	45.05	48.79	4.57	3.75
Mean	45.10	48.31	6.05	3.631427935

4. Conclusion

ARIMA model, in this case, accurately predicts the average temperature on specific days. The minimum difference of the forecasted temperature is 0.13, and the absolute deviation from the actual value is very low as 0.0085. The mean absolute deviation from the actual value is 3.63, indicating that the data, including the predicted temperature distribution values, is relatively close to the actual values but not as close as possible. So that achieve more perfect and accurate temperature forecasting, it is not reasonable to merely analyze the historical temperature data using the ARIMA model. The temperature changes due to plenty of variables, and it often needs satellites to observe. Another method of predicting temperature but not mentioned in this paper is multiple linear regression. Still, it is the prerequisite that determining the correlation between the dependent and independent variables and effectively observing them. More advanced methods could be using artificial intelligence or other deep learning algorithms. Even though the ARIMA model cannot consider the factors affecting temperature, it predicts the future temperature based on the historical time series.

References

[1] IPCC 2001 Climate Change 2001: Impacts, Adaptation and Vulnerability ed J J McCarthy et al (Cambridge: Cambridge University Press)

[2] SRK, Daily Temperature of Major Cities, Kaggle, https: // www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities.

[3] H. F. Tayeb, M. Karabatak and C. Varol, "Time Series Database Preprocessing for Data Mining Using Python," 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1-4, doi: 10.1109/ISDFS49300.2020.9116260.

[4] Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. Annual review of computer science, 4(1), 417-433.

[5] Shumway, R.H., Stoffer, D.S. (2017). ARIMA Models. In: Time Series Analysis and Its Applications. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-52452-8_3.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[6] Kyung So Im, M.Hashem Pesaran, Yongcheol Shin, Testing for unit roots in heterogeneous panels, Journal of Econometrics, Volume 115, Issue 1, 2003, Pages 53-74, ISSN 0304-4076, https://doi.org/10.1016/S0304-4076(03)00092-7.

[7] Shamsnia, Seyed Amir, et al. "Modeling of weather parameters using stochastic methods (ARIMA model) (case study: Abadeh Region, Iran)." International Conference on Environment and Industrial Innovation, IPCBEE. Vol. 12. 2011.

[8] Neath, Andrew A., and Joseph E. Cavanaugh. "The Bayesian information criterion: background, derivation, and applications." Wiley Interdisciplinary Reviews: Computational Statistics 4.2 (2012): 199-203.

[9] Dixon, Matthew, and Tyler Ward. "Takeuchi's Information Criteria as a form of Regularization." arXiv preprint arXiv: 1803.04947 (2018).

[10] Sin, Chor-Yiu, and Halbert White. "Information criteria for selecting possibly misspecified parametric models." Journal of Econometrics 71.1-2 (1996): 207-225.

[11] Hassani, Hossein, and Mohammad Reza Yeganegi. "Sum of squared ACF and the Ljung–Box statistics." Physica A: Statistical Mechanics and Its Applications 520 (2019): 81-86.

[12] Wallach, Daniel, and Bruno Goffinet. "Mean squared error of prediction as a criterion for evaluating and comparing system models." Ecological modelling 44.3-4 (1989): 299-306.

[13] Marill, Keith A. "Advanced statistics: linear regression, part II: multiple linear regression." Academic emergency medicine 11.1 (2004): 94-102.