Project I – 'wwwusage' Data Analysis

Li Hantao, G2101725H, MSAI, hli038@e.ntu.edu.sg

The wwwusage time series data consist of the number of users connected to the internet through a server.

The data are collected at a time interval of one minute and there are 100 observations.

Please fit an appropriate ARIMA model for it and submit a short report including R codes, the fitted model, the diagnostic checking, AIC, etc.

I. Original Data Analysis

This project asks us to fit an appropriate ARIMA model for the 'wwwusage' dataset. First of all, we will look in the original dataset, trying to find some attributes about the time series data. The original plot of data is shown in Fig. 1. We can effortlessly know that the maximum and minimum of data are 228 and 83, with a mean value of 137.08.

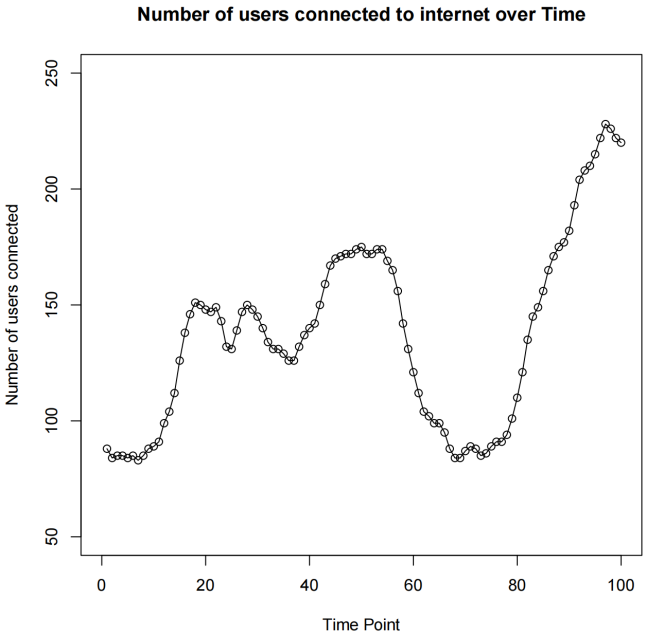

Fig. 1. Original Data Plot.

In addition, we can observe a slight trend of increase in the data, indicating that the original data is not a weakly stationary time series whose mean should be constant throughout time. Thus, the original data is demanded to take differencing later to fit ARIMA models. Moreover, we cannot figure out an apparent seasonal fluctuating in the data plot.

After analyzing the data plot, we can calculate the original data's ACF and PACF plots (Lag < 100), shown in Fig. 2. The ACF data is consistent with our previous analysis; that is, the ACF curve has no dies down. Although the PACF image cuts off at lag 2, we have observed that the data is unstable, so it is not helpful for the derivation of the ARIMA model. In addition, the image of PACF when lag is minimal is significantly higher than that of others, which also shows that the original image is unstable.

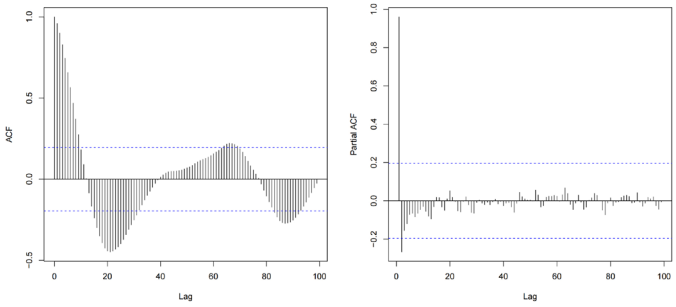

Fig. 2. ACF and PACF of Original Data.

Since the original data is unstable and has trending components, we need to perform a difference transform if we want to use the ARIMA model to fit it. To get the appropriate order, we examine the data after first-order and second-order differencing in a series of tests, including the Augmented Dickey-Fuller Test (ADF), the Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS), and the Philipps-Perron Test (PP). The results are shown in Table I. The 1% critical values for test statistics of KPSS and ADF are 0.739 and -2.6.

TABLE I
P-Value of Different Test

Data	KPSS	ADF	PP
First-Order	0.218	-4.071	0.072
Second-Order	0.052	-9.983	0.01

Observing the results in Table I, we found that both first-order and second-order differentiation can make the data into stationary time series. Specifically, the data obtained after the second-order differential is more consistent with the hypothesis, but we should select smaller parameters when setting the ARIMA model to reduce the complexity.

Therefore, we will discuss both the first-order and second-order differencing, respectively.

II. First-Order Differencing Analysis

We can implement the first-order differencing in order to the original data to make the data stationary, removing the trending component. The plot of first-order differencing data is shown in Fig. 3.

We first perform ACF and PACF analysis. The ACF and PACF plot after first-order differencing are shown in Fig. 4.

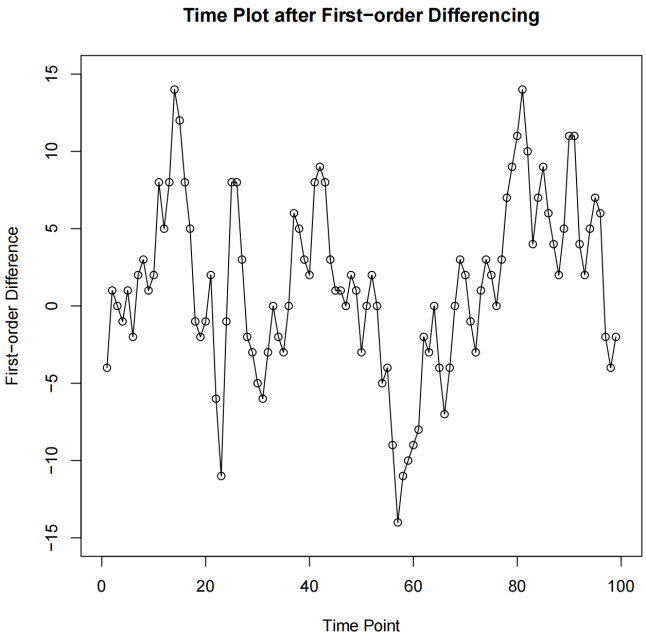

Fig. 3. First-Order Differencing Data Plot.

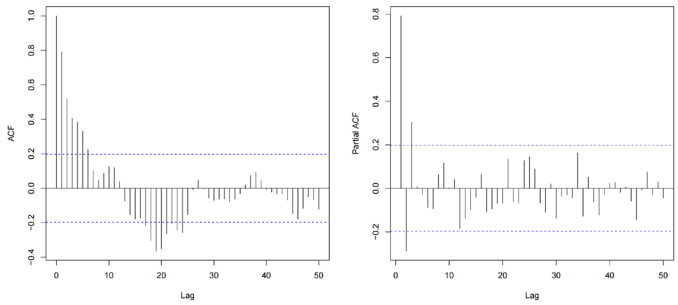

Fig. 4. ACF and PACF of First-Order Differencing Data.

For the ACF plot, we can find that ACF shows a wave pattern and does not cuts off until lag 24; while PACF cuts quickly at lag 3. This pattern of ACF and PACF indicates that we should implement an AR model instead of an ARMA model, which means ARIMA(3,1,0) may be an appropriate model.

On the other hand, the *ar.yw()* function, the Yule-Walker test to estimate the AR coefficient, also gives the parameter of AR as 3. As a result, the ARIMA(3,1,0) is utilized to fit the original time series data.

We can use the *tsdiag()* function and *checkresiduals()* function to implement the diagnostic check smoothly. Furthermore, we can plot the Normal Q-Q Plot of residuals. Fig. 5 shows the results. (Due to the same sub-plot in the two different test functions, we rearrange the plot to delete the overlapping part and add the Q-Q plot in it.)

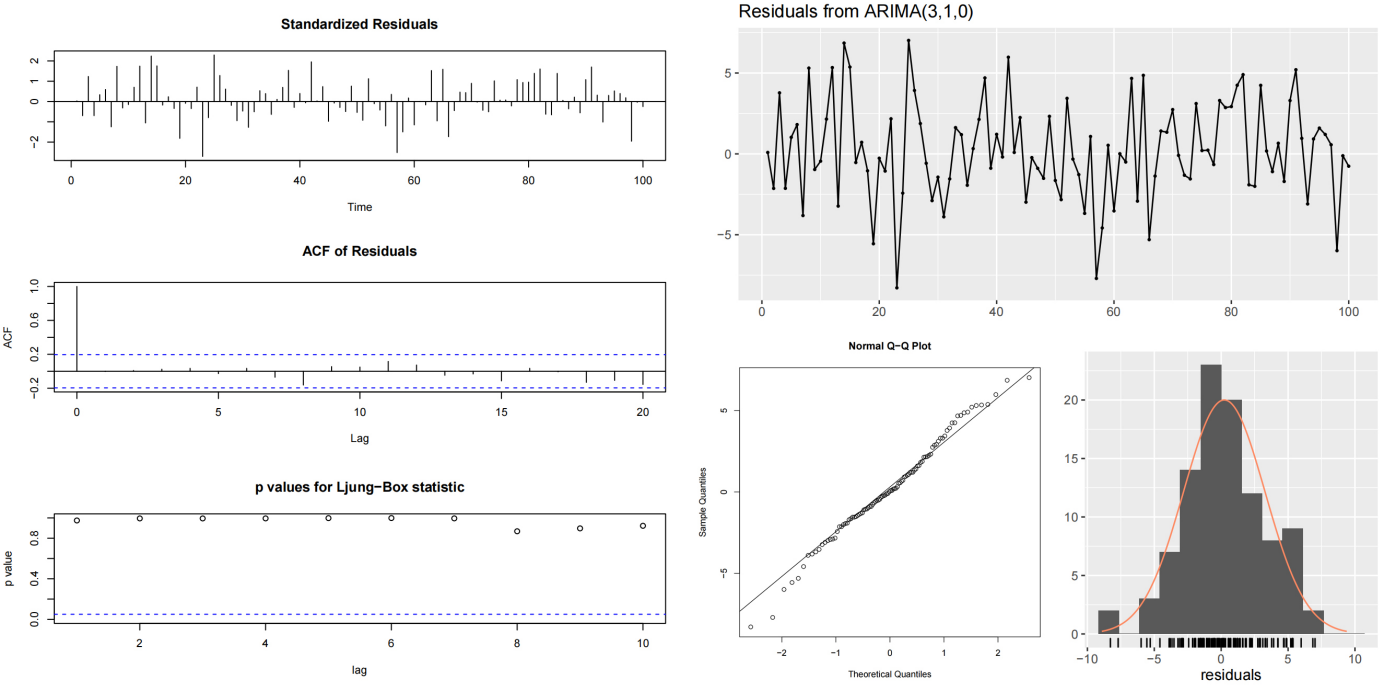

Fig. 5. Diagnostic Check of ARIMA(3,1,0).

It can be observed that ARIMA(3,1,0) provides adequate fit for the time series data. More specific:

1. The residuals (standardized residuals) look random enough.
2. The histogram of residuals and Q-Q plot of residuals show that residuals are enough to be regarded as white noise.
3. The ACF of the residuals cuts off after lag 0.
4. The p-value for the Ljung-Box statistic is all high enough. We can get the result of 0.9749, which is high beyond 0.05, through the Box.test() function.
5. The AIC and BIC of ARIMA(3,1,0) are 511.994 and 522.3745, respectively.

Having concluded that ARIMA(3,1,0) provides an adequate fit, we implement the forecasting and validation with the model. We can use the *forecast()* function to create the forecasting data easily, which is shown in Fig. 6.

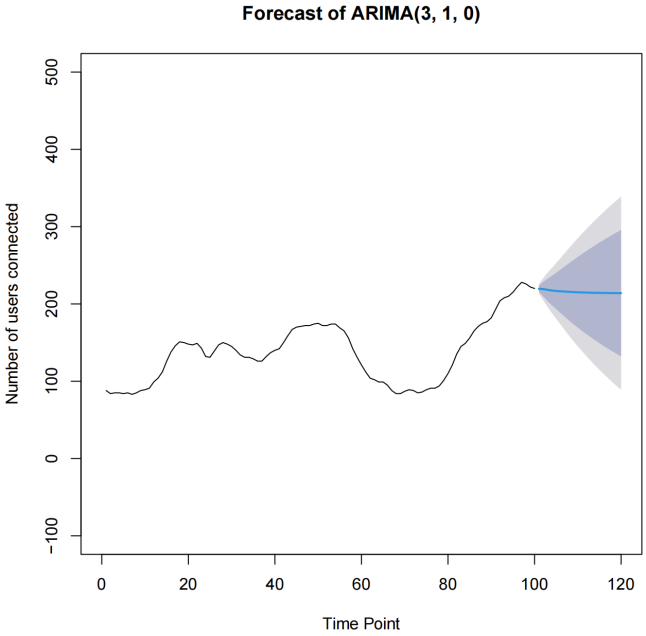

Fig. 6. Forecast of following 20 Points with ARIMA(3,1,0).

Since we have no accurate data about the following time point, we cannot evaluate the accuracy of the prediction about the time of the following model. Therefore, we need to use the existing data to divide it into a training set and a validation set to verify the existing data.

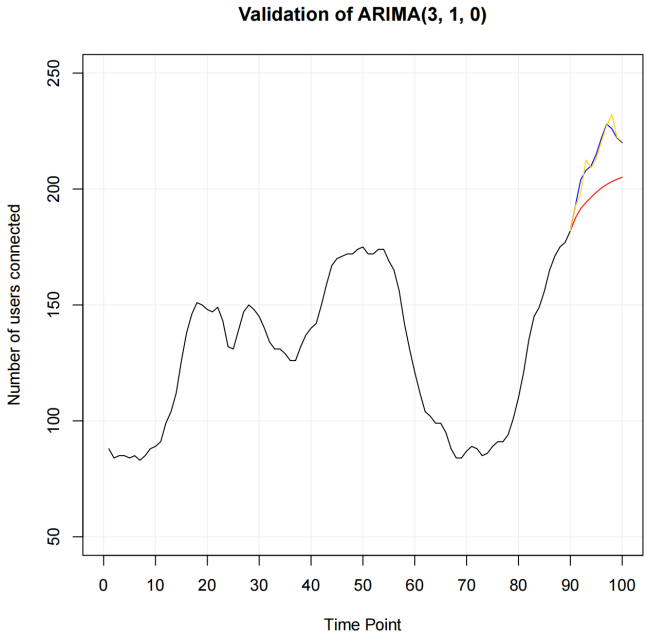

Fig. 7. Validation of 90:10 Points with ARIMA(3,1,0).

We divide the original data into the training set and the validation set with the proportion of 9:1. ARIMA(3,1,0) is used to fit the data on the training set, and the predicted data are compared with the real data of the validation set. We have two prediction methods: 10-ahead validation, using only the previous 90 data points; the other is one-step validation, utilizing ground truth data step by step. The results are shown in Fig. 7. The blue line in the figure is real data, the red line is 10-ahead validation, and the gold line shows the one-step validation. We can find that the outcome of one-step verification is better than the 10-ahead prediction, almost fitting the real data. A sample explanation is that we continue to use GT data in one-step prediction.

If the model is only explained from the 10-ahead validation's result, it is apparently not fit enough. Because all predicted values are lower than the real value, that is, if we add a small constant e to the predicted result, the result must be better than the existing model.

In addition, we can use the *accuracy()* function to verify the verification results quantitatively. This discussion will be carried out later in the following section.

III. Second-Order Differencing Analysis

In Section I, we cannot determine the differencing order for the time being. Therefore, similar to the first-order differencing, we can further carry out the second-order differencing and analyze the results. The plot of second-order differencing data is shown in Fig. 8.

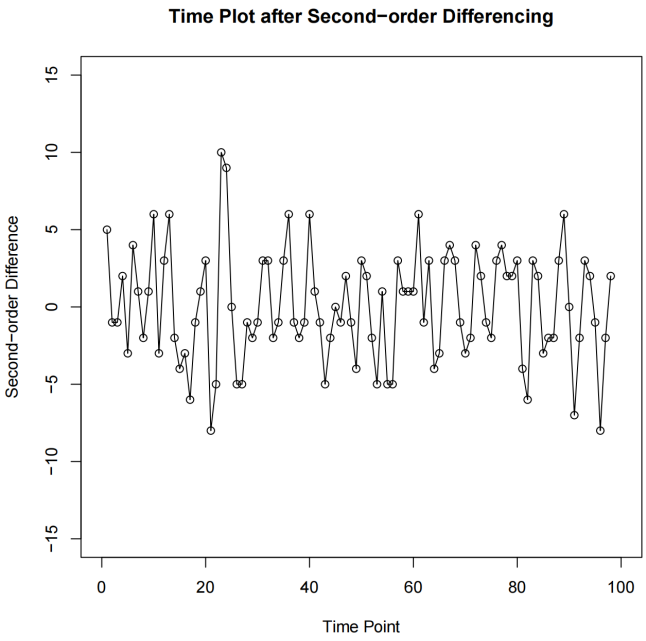

Fig. 8. Second-Order Differencing Data Plot.

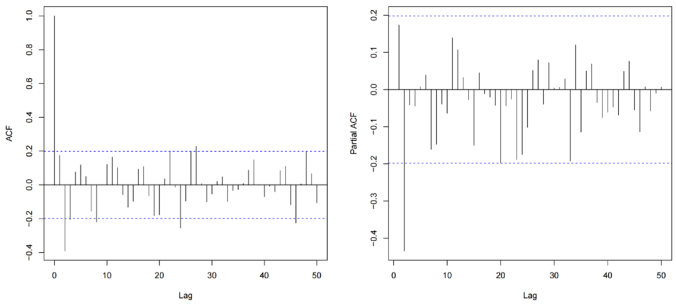

Fig. 9. ACF and PACF of Second-Order Differencing Data.

Fig. 8 shows that the data after the second-order differencing is more random than the first-order differencing, in line with the previous speculation. We continue to observe the ACF plot and PACF plot of the second-order, shown in Fig. 9.

The properties of the ACF plot and PACF plot of second-order differencing are roughly similar to those of first-order differencing. The ACF plot does not cut off until lag 27, while the PACF plot cuts off quickly at lag 2. Therefore, similar to the first-order differencing model, we choose the second-order AR model ARIMA(2,2,0) as the possible model, which is also suggested by the Yule-Walker function.

As a result, the ARIMA(2,2,0) is utilized to fit the original time series data. Same as the previous model, we can use the *tsdiag()* function and *checkresiduals()* function to implement the diagnostic check easily. The results are shown in Fig. 10.

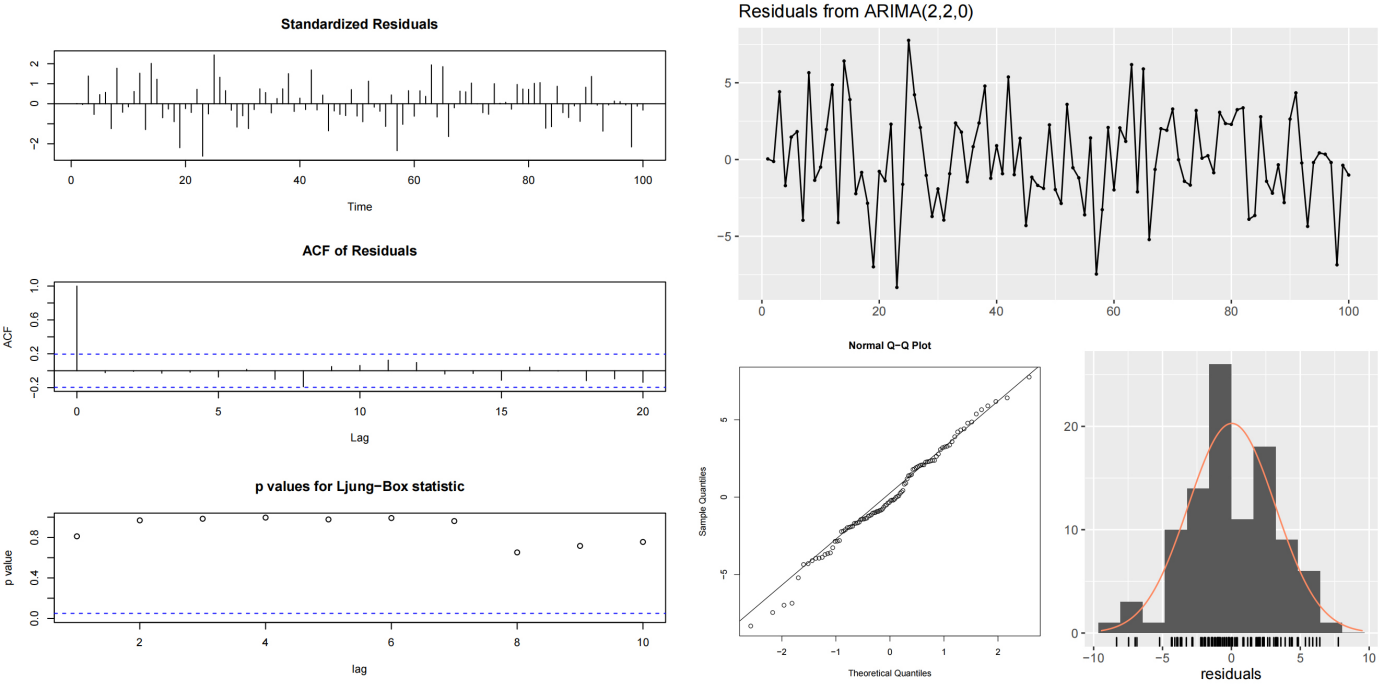

Fig. 10. Diagnostic Check of ARIMA(2,2,0).

The diagnostic check result analysis process is similar to the previous analysis and will not be repeated here. We can get the conclusion that the model of ARIMA(2,2,0) provides adequate fit for the time series data. Then, we do the same process of forecasting and validation as the previous model. The results are shown in Fig. 11.

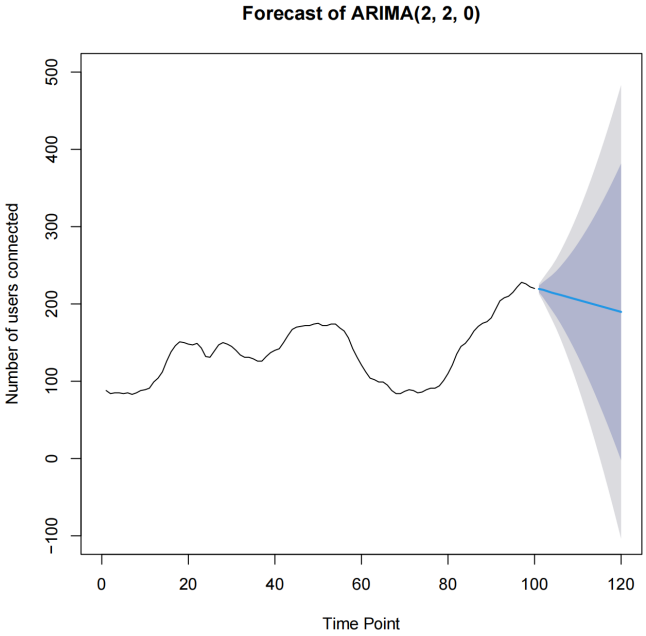

Fig. 11. Forecast of following 20 Points with ARIMA(2,2,0).

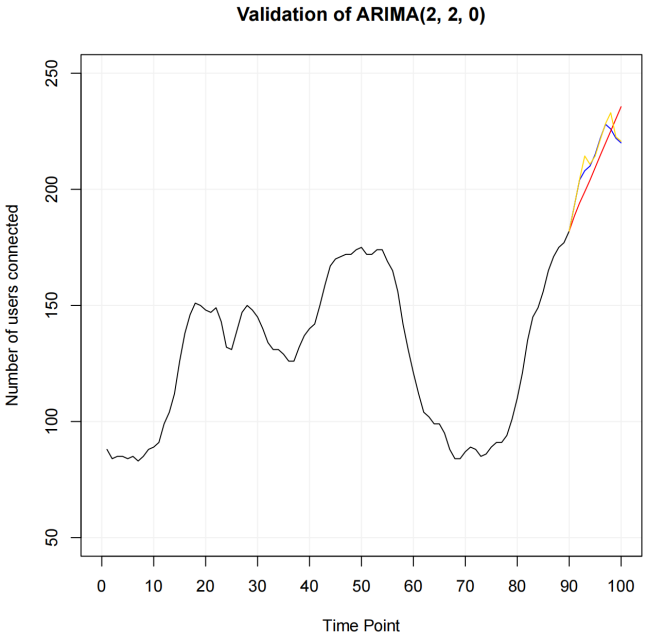

Fig. 12. Validation of 90:10 Points with ARIMA(2,2,0).

We can find that the upper confidence and lower confidence edge of ARIMA(2,2,0) in forecasting have a significantly more extensive range compared with the ARIMA(3,1,0).

Intuitively, this may indicate that the prediction performance of ARIMA(2,2,0) is not exceptional. This conclusion will be confirmed in the subsequent analysis of MASE. We also test the validation performance of ARIMA(2,2,0). The figure about the validation is shown below in Fig. 12.

Interestingly, although the result deviation of one-step validation seems more significant, which coincides with the discussion above, the result of 10-ahead validation is much better than ARIMA(3,1,0), which is worthy of our attention.

IV. Auto ARIMA

In the previous discussion, we determined the parameters of the ARIMA model by analyzing the basic attributes of data and observing the ACF plot and the PACF plot. However, in R's function library, *auto.arima()* is a function to help us automatically determine ARIMA parameters according to the AIC, BIC, and AICC values. Therefore, we will use *auto.arima()* in this section to automatically determine the optimal parameters, which are (1,1,1).

As a result, the ARIMA(1,1,1) is utilized to fit the original time series data. Same as the previous model, we can use the *tsdiag()* function and *checkresiduals()* function to implement the diagnostic check easily. The results are shown in Fig. 13.

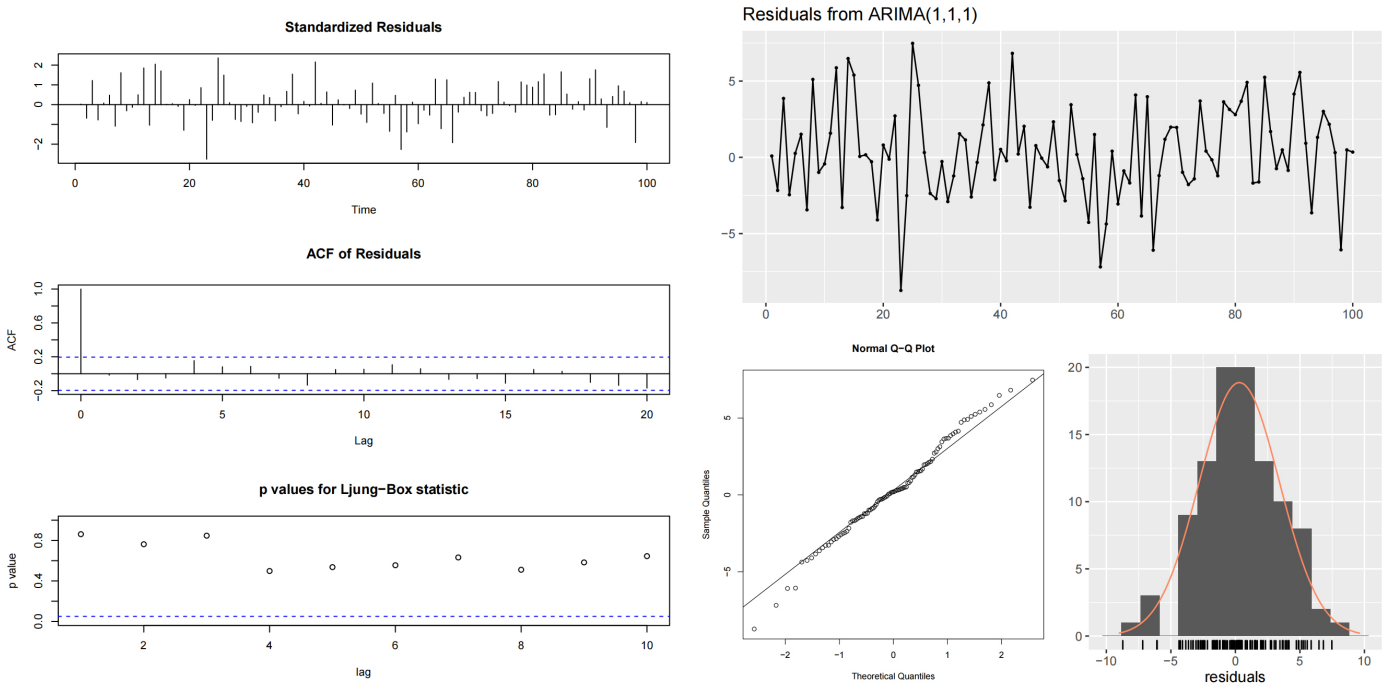

Fig. 13. Diagnostic Check of ARIMA(1,1,1).

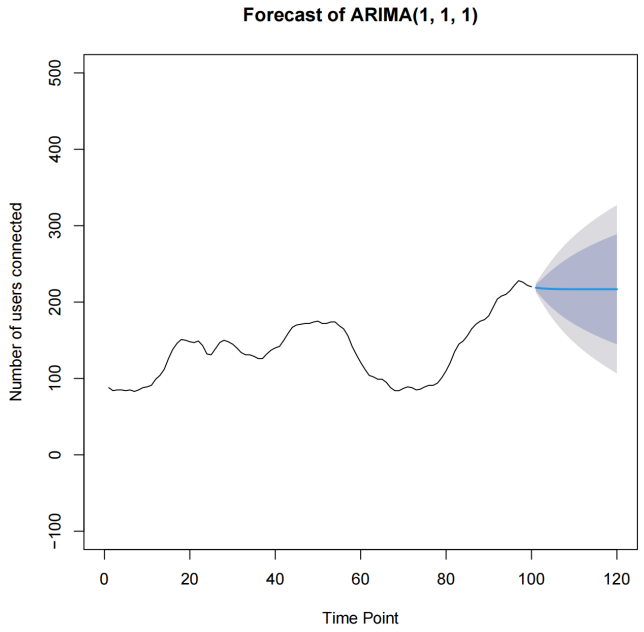

Fig. 14. Forecast of following 20 Points with ARIMA(1,1,1).

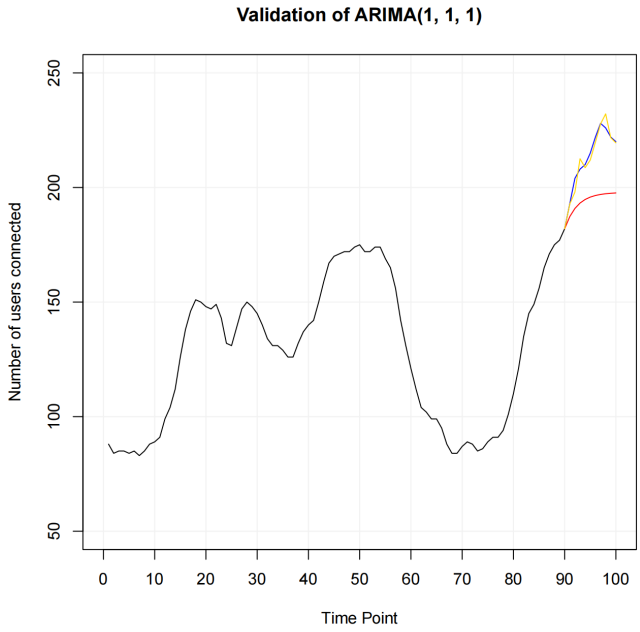

Fig. 15. Validation of 90:10 Points with ARIMA(1,1,1).

The analysis process of the diagnostic check result is similar to the previous analysis and will not be repeated here. It is worth noting that the p-value obtained by the Ljung-Box test is lower than that of the previous two models, but it is still far beyond the confidence interval of 0.05.

We can get the conclusion that the model of ARIMA(1,1,1) provide adequate fit for the time series data. Then, we do the same process about forecasting and validation as the previous model. The results are shown in Fig. 14.

The result of forecasting is very similar to ARIMA(3,1,0). Next, carry out validation, and the results are shown in FIG. 15. It can be found that the deviation between the results of 10-ahead validation and the real data is more significant than ARIMA(3,1,0).

V. Brute-Force to Finding Parameters

In the previous discussion, we discussed three possible models. The parameters of these models are not only from the analysis of data characteristics but also from the automatic search of functions. After analyzing these three models, we found that they effectively fit the existing data. However, we did not find a "best" model parameter. Therefore, in this section, we will use the brute force method to traverse and search the model parameters and record the AIC and BIC of the model.

TABLE II
AIC Value of Different Parameters

Diff=1		q						
		0	1	2	3	4	5	6
p	0	N.A.	549.8	519.8	520.2	519.3	518.8	518.2
	1	529.2	529.2	529.2	529.2	529.2	529.2	529.2
	2	522.1	516.2	517.3	515.7	513.2	518.0	520.0
	3	**511.9**	513.9	515.6	514.4	514.7	516.4	514.1
	4	513.9	515.9	516.1	519.0	515.3	518.1	517.2
	5	515.8	517.6	513.5	521.6	**511.1**	**512.7**	514.7
	6	517.4	519.6	518.5	519.4	521.3	N.A.	N.A.

Diff=2		q						
		0	1	2	3	4	5	6
p	0	N.A.	523.9	517.2	512.3	513.8	515.7	517.3
	1	530.4	523.5	513.1	513.8	515.7	513.8	515.1
	2	**511.4**	513.2	515.1	515.7	513.8	514.6	516.9
	3	513.2	**510.7**	512.6	514.4	513.2	513.5	515.7
	4	515.1	517.1	514.5	515.0	517.8	519.9	517.0
	5	517.1	518.7	516.3	518.3	514.1	**509.8**	511.6
	6	518.7	520.0	518.5	516.4	516.0	N.A.	N.A.

The three most minor (best) values in each table are marked in bold.

TABLE III
BIC Value of Different Parameters

Diff=1		q						
		0	1	2	3	4	5	6
p	0	N.A.	554.9	527.6	530.6	532.3	534.4	536.4
	1	534.4	**522.0**	**526.6**	527.5	530.6	534.4	538.7
	2	529.9	526.6	530.3	531.3	531.4	538.8	543.4
	3	**522.3**	526.9	531.1	532.5	535.5	539.7	540.0
	4	526.9	531.5	534.3	539.8	538.7	544.0	545.8
	5	531.4	535.8	534.3	544.9	537.0	541.3	545.9
	6	535.6	540.3	541.8	545.4	549.9	N.A.	N.A.

Diff=2		q						
		0	1	2	3	4	5	6
p	0	N.A.	529.0	524.9	**522.6**	526.7	531.2	535.3
	1	535.6	531.3	**523.5**	526.7	531.3	531.9	535.8
	2	**519.2**	523.5	528.0	531.2	531.9	535.3	540.2
	3	523.6	523.6	528.1	532.5	533.9	536.7	541.6
	4	528.0	532.6	532.6	535.7	541.1	545.7	545.5
	5	532.6	536.8	537.0	541.6	539.9	538.2	542.6
	6	536.8	540.7	541.7	542.3	544.4	N.A.	N.A.

The three most minor (best) values in each table are marked in bold.

Because we want to avoid the high complexity of the model, we limit the model parameters to less than six while only searching for the first-order and second-order differencing. The search results are listed in Table II and Table III.

Through the table, we can observe that the parameter values with small AIC and BIC values are approximately consistent with or identical to the parameter values we analyzed before. For example, the three model parameters we discussed earlier are all bold. This indicates that our previous analysis is accurate. In addition, we found a new set of parameters, (5,5) parameter pairs. Since the complexity of the model significantly constrains BIC, this set of parameters is not bold in the BIC table and not be suggested in auto.arima(). However, if only AIC is considered, this set of parameters shows outstanding properties.

Therefore, we discuss the pair of parameters (5,2,5). The ARIMA(5,2,5) is utilized to fit the original time series data. Same as the previous model, we can use the *tsdiag()* function and *checkresiduals()* function to implement the diagnostic check easily. The results are shown in Fig. 16. We can see from the figure that the diagnostic diagram of ARIMA (5,2,5) can almost be considered to be better than the previous three models.

We can get the conclusion that the model of ARIMA(5,2,5) provide adequate fit for the time series data. Then, we do the same process about forecasting and validation as the previous model. The results are shown in Fig. 17.

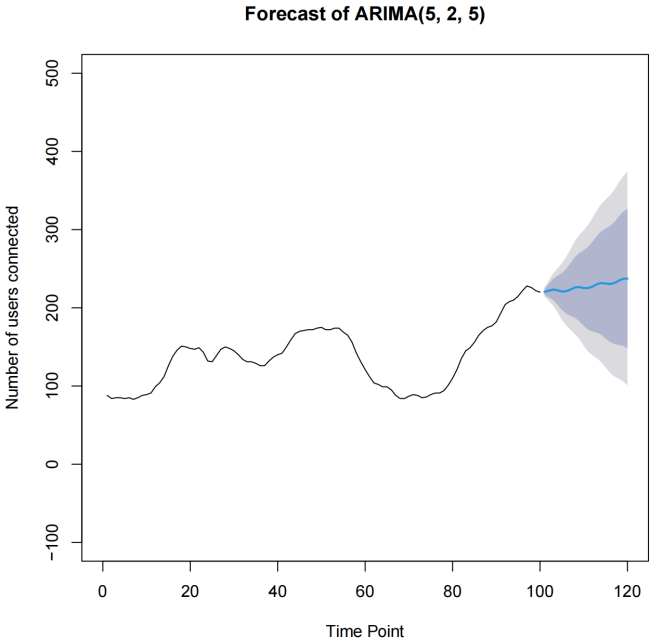

Fig. 17. Forecast of following 20 Points with ARIMA(5,2,5).

In the forecasting plot, we found that different from the ARIMA model with simple parameters before, there are fluctuations in the predicted image of ARIMA (5,2,5). On the one hand, this is in line with our intuition in real life, but it does not mean that it is more reasonable. We need to continue observing the validation result, shown in Fig. 18.

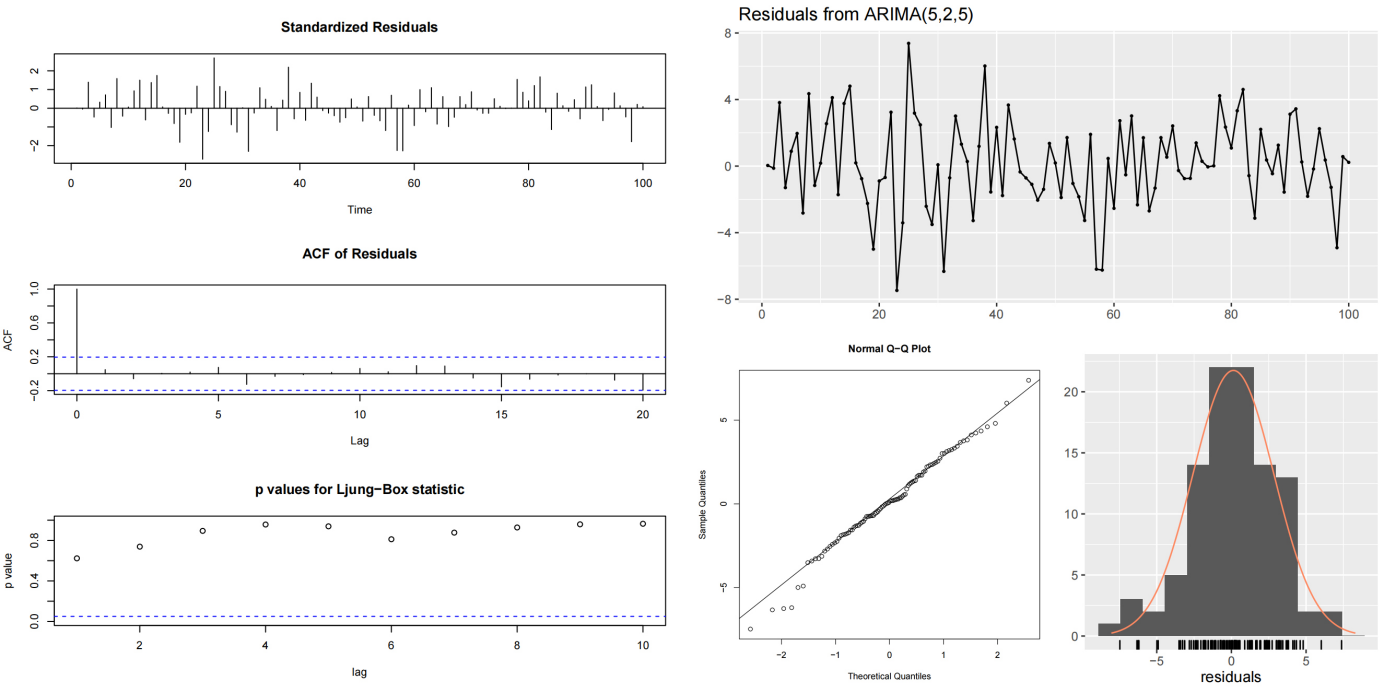

Fig. 16. Diagnostic Check of ARIMA(5,2,5).

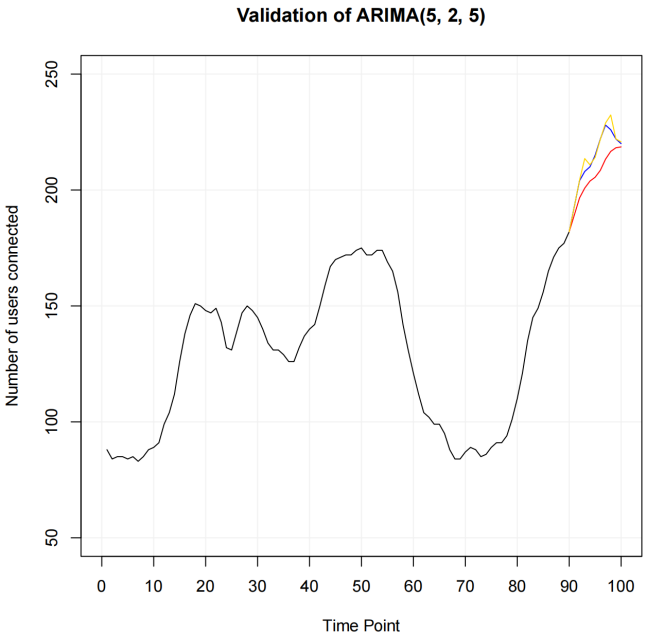

Fig. 18. Validation of 90:10 Points with ARIMA(5,2,5).

Firstly, we find that the 10-ahead validation of the two second-order differencing models discussed in this paper, (2,2,0) and (5,2,5), fits the real data much better than the first-order differencing model, (3,1,0) and (1,1,1). However, we cannot quantitatively analyze the fitting degree of the model, so we will conduct MASE analysis on the four models in the next section.

VI. Accuracy Analysis

In the previous discussion, we made a qualitative analysis of the model's prediction ability, but there is a lack of quantitative data to compare it. Therefore, we use the *accuracy()* function to analyze the above four models in this section. In order to simplify the analysis procedure, we select RMSE, MAE, and MASE. The results are shown in Table IV.

TABLE IV
Accuracy

Set	Model	RMSE	MAE	MASE
Training	(3,1,0)	3.044	2.367	0.523
	(1,1,1)	3.113	2.405	0.531
	(2,2,0)	*3.150*	*2.511*	*0.555*
	(5,2,5)	**2.713**	**2.098**	**0.463**
Test	(3,1,0)	12.071	10.086	2.229
	(1,1,1)	**11.351**	**9.265**	**2.047**
	(2,2,0)	*14.750*	*13.118*	*2.899*
	(5,2,5)	12.528	9.415	2.080

The best result in each criterion of data is marked in bold and the worst result is marked in italics.

It can be seen that the conclusions represented by various criteria are very consistent. Among them, ARIMA(2,2,0) has the worst prediction accuracy in both training set and test set, which shows that this model may not be ideal. This conclusion is consistent with the too extensive confidence interval in the prediction results we analyzed in Section III.

ARIMA(5,2,5), the best model obtained by searching parameters, performs best in the training set, while ARIMA(1,1,1), the model given by automatic parameter function, performs best in the test set. These two models have their own advantages and disadvantages. ARIMA(1,1,1) has lower parameter complexity, but the residual analysis of the ARIMA(5,2,5) model appears to be more likely as Gaussian white noise. In practical application, how to choose them needs to be analyzed in detail, combined with the complexity requirements of the model.