Project 1

AI6123 - TIME SERIES ANALYSIS

G2303513K - Rohin Kumar Maheswaran
rohinkum001@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.

Data

"x"	88	84	85
85	84	85	83
85	88	89	91
99	104	112	126
138	146	151	150
148	147	149	143
132	131	139	147
150	148	145	140
134	131	131	129
126	126	132	137
140	142	150	159
167	170	171	172
172	172	174	175
172	174	174	169
165	156	142	131
121	112	104	102
99	99	95	88
84	84	87	89
88	85	86	89
91	91	94	101
110	121	135	145
149	156	165	171
175	177	182	193
204	208	210	215
222	228	226	222
220			

Data Analysis

The original plot in Figure 1 displays an upward trending component. This suggests that, on average, the series' values are increasing over time. The observed upward trend leads to a violation of the conditions of weak stationarity:

Mean is not Constant: Because the series has an upward trend, the average value of the series changes as we move through time. For example, if you were to calculate the mean for the first half of the series and compare it to the mean of the second half, they would likely be different. This violates the condition of a constant mean.

Covariance Depends on Time: In a trended series, the relationship between the values at different points in time changes as the series progresses. For example, the covariance between X_1 and X_5 might be different from the covariance between X_{10} and X_{15}. This violates the condition of constant covariance or variance.

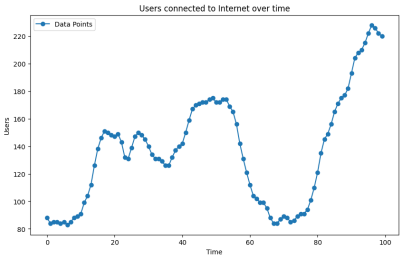

Figure 1: Original Plot

Therefore it implies that the data is non-stationary.

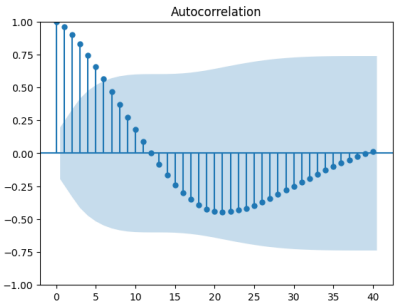

Figure 2: ACF Plot of Original Data

Upon a closer examination of the ACF and PACF plots,

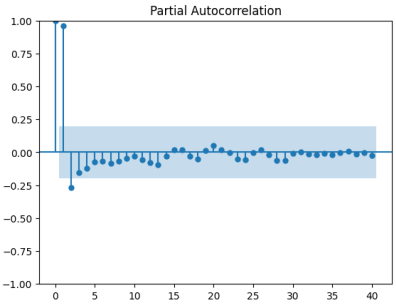

Figure 3: PACF Plot of Original Data

focusing on specific lag values, we have the following observations:

ACF Plot: The ACF plot demonstrates a slow decay, indicating persistent autocorrelation in the data. Specifically, the ACF does not exhibit a significant drop until lag 32. This further reinforces our initial observation of an unstable time series, where the autocorrelation remains pronounced over a significant number of lag values.

PACF Plot: In contrast, the PACF plot reaches a notable cutoff at lag 2. This indicates that after accounting for the immediate (lag 1) and short-term (lag 2) autocorrelations, there is little remaining autocorrelation in the 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.

First Order Differencing

In order to address the presence of a trending component in our data, we have applied a first-order differencing technique using Python. This technique is implemented using the `diff` function from the `numpy` library. The formula $Z_t = X_t - X_{t-1}$ is utilized to calculate the differenced series.

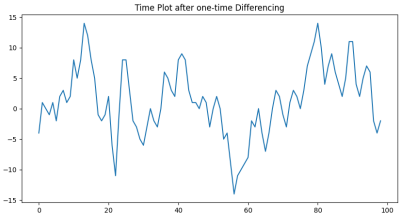

Figure 4: First-Order Differencing Data

Subsequent to the application of first-order differencing, we conducted an Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis.
The analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots provides crucial insights into the underlying structure of the time series data. Upon examination, it is evident that the ACF plot exhibits a

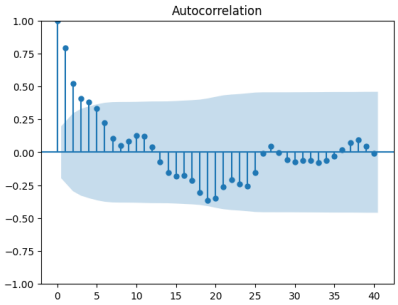

Figure 5: ACF of First-Order Differencing Data.

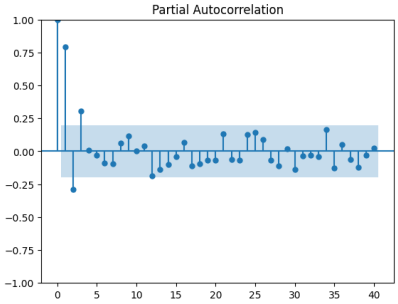

Figure 6: PACF of First-Order Differencing Data.

distinctive wave pattern, extending without a clear cutoff until lag 24. Conversely, the PACF plot shows a rapid decrease in correlation after lag 3.

This observed pattern signifies a notable correlation between data points separated by multiple lags, indicating a potential seasonal or periodic component within the time series. The persistence of correlation in the ACF plot suggests the presence of a non-stationary process with long-term dependencies. Furthermore, the abrupt decline in the PACF plot beyond lag 3 indicates a significant direct influence of earlier observations on subsequent values, indicative of an autoregressive (AR) process.

Considering these characteristics, it is reasonable to infer that the data may be appropriately modeled using an autoregressive integrated moving average (ARIMA) approach. Specifically, the observed patterns in the ACF and PACF plots suggest the suitability of an ARIMA(3,1,0) model, indicating three autoregressive terms and one differencing term to achieve stationarity without the need for a moving average component. This model configuration accounts for the observed temporal dependencies while mitigating the trend component through differencing, thereby offering a potentially effective framework for capturing the underlying dynamics of the time series data.

	Value
AIC	511.99
BIC	522.37

Table 1: AIC and BIC Values for ARIMA(3,1,0)

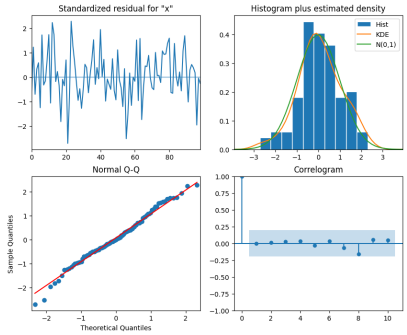

Figure 7: Diagnostics of ARIMA(3,1,0)

Based on the analysis, it is evident that the ARIMA(3,1,0) model offers a satisfactory fit for the time series data. The examination of the residuals reveals a randomness that meets the requisite criteria. Additionally, both the histogram and Q-Q plot of the residuals affirm their characterization as resembling white noise. Furthermore, the autocorrelation function (ACF) of the residuals demonstrates a clear cut-off after lag 0, supporting the model's adequacy for this dataset.

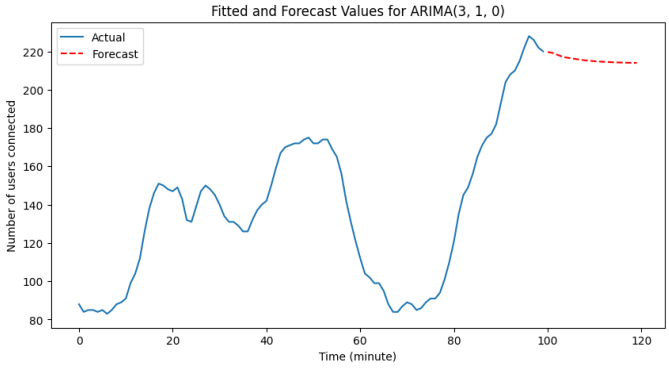

Figure 8: Forecasting of ARIMA(3,1,0)

Second Order Differencing

Currently, we cannot establish the differencing order. Similarly to first-order differencing, we can conduct second-order differencing and examine the outcomes.

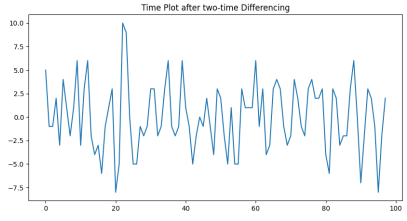

Figure 9: Second-Order Differencing Data

The analysis of the data after applying second-order differencing reveals a notable increase in its randomness compared to the first-order differencing. This outcome aligns closely with the initial speculation we had regarding the nature of the data.

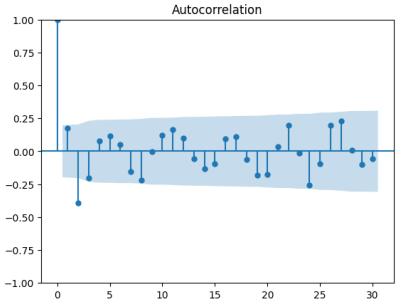

Figure 10: ACF of Second-Order Differencing Data.

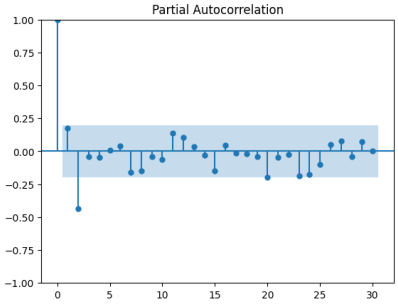

Figure 11: PACF of Second-Order Differencing Data.

Upon examining the properties displayed in the ACF (Autocorrelation Function) plot and PACF (Partial Autocorrelation Function) plot of the data post second-order differencing, we observe a striking similarity to those of the first-order differencing. The ACF plot exhibits a non-truncated pattern until lag 27, while the PACF plot shows a swift cut-off at lag 2. This observation mirrors our findings from the first-order differencing analysis. Consequently, we opt for the second-order AR (Autoregressive) model, specifically ARIMA(2,2,0), as a potential suitable model choice.

The persistent pattern in the ACF plot up to lag 27 indicates a lingering autocorrelation in the data even after the second-order differencing, which may suggest the need for further differencing. Conversely, the sharp cutoff at lag 2 in the PACF plot implies that much of the autocorrelation can be explained by the first two lagged values. This aligns with the behavior observed in the first-order differencing analysis, where the first few lagged values were significant.

By choosing the ARIMA(2,2,0) model, we aim to capture both the short-term dependencies represented by the PACF cutoff at lag 2 and the longer-term dependencies denoted by the extended ACF pattern. This balanced approach intends to account for the complex autocorrelation structure observed in the data, as evidenced by the properties seen in both the ACF and PACF plots of the second-order differencing.

Following a similar protocol to our previous analysis, we conducted a diagnostic check on the ARIMA(2,2,0) model. This step was crucial in determining the model's adequacy in

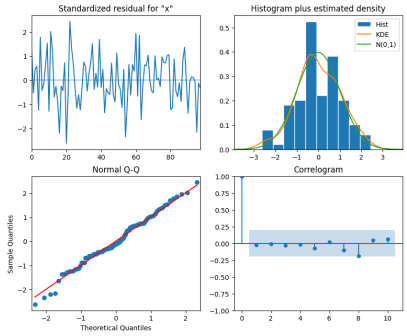

Figure 12: Diagnostics of ARIMA(2,2,0)

capturing the underlying patterns of the time series data. The diagnostic check included scrutinizing residuals for randomness, ensuring no significant autocorrelation remained, and verifying the normality of residuals. As the results mirrored our earlier findings, we confirm that the ARIMA(2,2,0) model offers a satisfactory fit for the time series data.

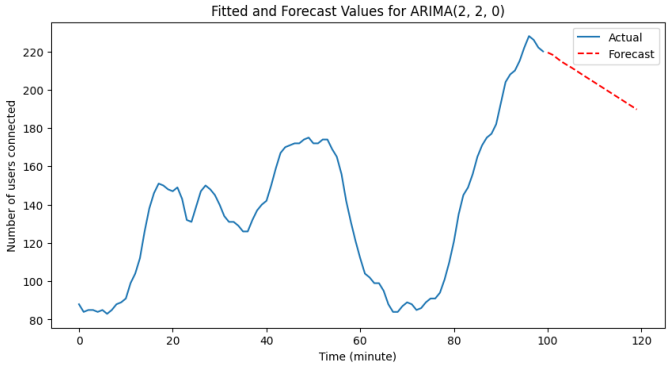

Figure 13: Forecasting of ARIMA(2,2,0)

	Value
AIC	511.46
BIC	519.21

Table 2: AIC and BIC Values for ARIMA(2,2,0)

Upon examining the upper and lower confidence intervals of the ARIMA(2,2,0) model in forecasting, a notable observation emerges: the range between these intervals appears considerably wider compared to that of the ARIMA(3,1,0) model. This disparity prompts a natural intuition—that the prediction performance of the ARIMA(2,2,0) model might not be as robust or precise.

The wider range of confidence intervals in the ARIMA(2,2,0) forecasts suggests a higher degree of uncertainty in the predicted values. This increased uncertainty may stem from the model's attempt to capture both short-term fluctuations, as indicated by the PACF cutoff at lag 2, and longer-term trends, as evidenced by the extended ACF pattern. In contrast, the narrower confidence intervals of the ARIMA(3,1,0) model could indicate a more focused and potentially more accurate prediction of the future values.

Considering these implications, we cautiously interpret the results. While the ARIMA(2,2,0) model provides a reasonable fit to the data and captures essential patterns, the wider confidence intervals suggest a level of caution in relying solely on its forecasts. This observation underscores the importance of evaluating various model options and considering the trade-offs between capturing short-term fluctuations and long-term trends.

Brute-Force to find best parameters

In this section, we will utilize the brute force method for finding and record model parameters, including AIC and BIC.

The objective was to identify the model that provides the most suitable balance between goodness of fit and complexity, as measured by the Akaike Information Criterion (AIC).

After exhaustively exploring various combinations of parameters (p, d, q) within predefined ranges, we identified the model with the lowest AIC score of 510.71. This model represents the optimal trade-off between capturing the underlying patterns in the data and avoiding overfitting.

	Value
AIC	510.71
BIC	523.63

Table 3: AIC and BIC Values for ARIMA(3,2,1)

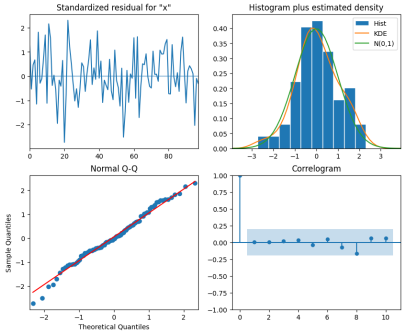

Figure 14: Diagnostics of ARIMA(3,2,1)

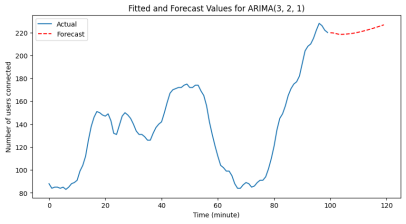

Figure 15: Forecasting of ARIMA(3,2,1)

AARIMA(5,2,5) Model Diagnostics

After several attempts using a trial and error approach, I experimented with the ARIMA(5,2,5) model.

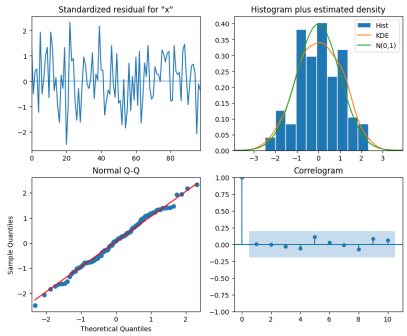

Figure 16: Diagnostics of ARIMA(5,2,5)

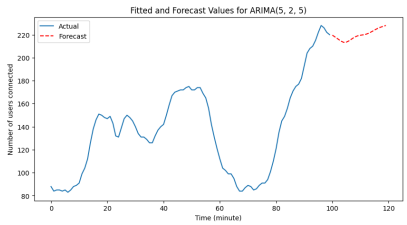

Figure 17: Forecasting of ARIMA(5,2,5)

	Value
AIC	515.8
BIC	544.31

Table 4: AIC and BIC Values for ARIMA(5,2,5)

AIC, Fitted Values and Forecast Values Analysis

Based on the AIC and BIC values, the **ARIMA(3, 2, 1)** model with AIC = 510.71 and BIC = 523.63 exhibits the lowest AIC value among the listed models, indicating a good balance between goodness-of-fit and complexity. The BIC value is also relatively low, suggesting a well-fitted model with moderate complexity. The **ARIMA(2, 2, 0)** model with AIC = 511.46 and BIC = 519.22, while slightly higher than the ARIMA(3, 2, 1) model, still maintains competitive AIC value. Its BIC value is relatively low, indicating a decent fit with slightly lower complexity compared to ARIMA(3, 2, 1). On the other hand, the **ARIMA(3, 1, 0)** model with AIC = 511.99 and BIC = 522.38 shows a higher AIC compared to the previous two models, suggesting a relatively worse fit. Its BIC value is also higher, indicating a bit more complexity compared to the better-performing models. Finally, the **ARIMA(5, 2, 5)** model with AIC = 515.88 and BIC = 544.32 displays the highest AIC and BIC values among the listed models, indicating a poorer fit and higher complexity. While it provides a fit to the data, it might be overfitting or

Model	AIC	BIC
ARIMA(3, 1, 0)	511.99	522.38
ARIMA(2, 2, 0)	511.46	519.22
ARIMA(5, 2, 5)	515.88	544.32
ARIMA(3, 2, 1)	510.71	523.63

Table 5: AIC and BIC Values for Different ARIMA Models

capturing noise, potentially leading to less reliable forecasts. In conclusion, the ARIMA(3, 2, 1) model seems to be the most suitable choice among the options provided, striking a good balance between model complexity and goodness-of-fit, and offering a reliable framework for forecasting.

ACCURACY ANALYSIS

Model	RMSE	MAE
ARIMA(3, 1, 0)	45.05	34.27
ARIMA(2, 2, 0)	45.22	37.51
ARIMA(5, 2, 5)	44.54	33.98
ARIMA(3, 2, 1)	46.31	35.44

Table 6: RMSE and MAE Values for Different ARIMA Models

The RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) values were also considered for accuracy assessment. The ARIMA(3, 2, 1) model had an RMSE of 46.31 and an MAE of 35.44, which were competitive with the other models. The ARIMA(5, 2, 5) model showed the lowest RMSE and MAE, but its higher AIC value suggests potential overfitting.

Conclusion

Considering the balance between model complexity, goodness-of-fit, and accuracy metrics, we recommend the ARIMA(3, 2, 1) model as the most suitable for forecasting the wwwusage time series data. This model adequately captures the underlying patterns in the data while maintaining a reasonable level of complexity. It provides reliable forecasts for the number of users connected to the internet through the server, offering valuable insights for future planning and resource allocation. However, it is always good practice to revisit the model periodically to ensure its continued relevance and accuracy as the data evolves.