Project report
SF2943 Time Series Analysis
Group 7

Shengyang Zhuang
szhuang@kth.se
20000802-T172

Ashot Hovsepyan
ashoth@kth.se
990103-7458

Kevin Johansson
kevinjo@kth.se
920411-4798

Mohammed Rafid
mohra@kth.se
960316-5631

Vlado Grujcic
grujcic@kth.se
980401-0156

May 22, 2022

Part A. Analysis of a non-financial time series

1 Introduction

This report aims to fit a stationary time series model to a data set of quarterly Australian electricity production ranging from 1956 to 2010. The production volume is given in kWh. The analysis will be conducted in RStudio, which is a software used for statistical analysis. The source of the data comes from key2stats data set.

2 Classical decomposition model

2.1 Preliminary analysis

- Step 1: Plotting the data and making a judgment

The time series represents the quarterly Australian electricity production in kWh from 1956 to 2010. (Fig.1) shows the data.

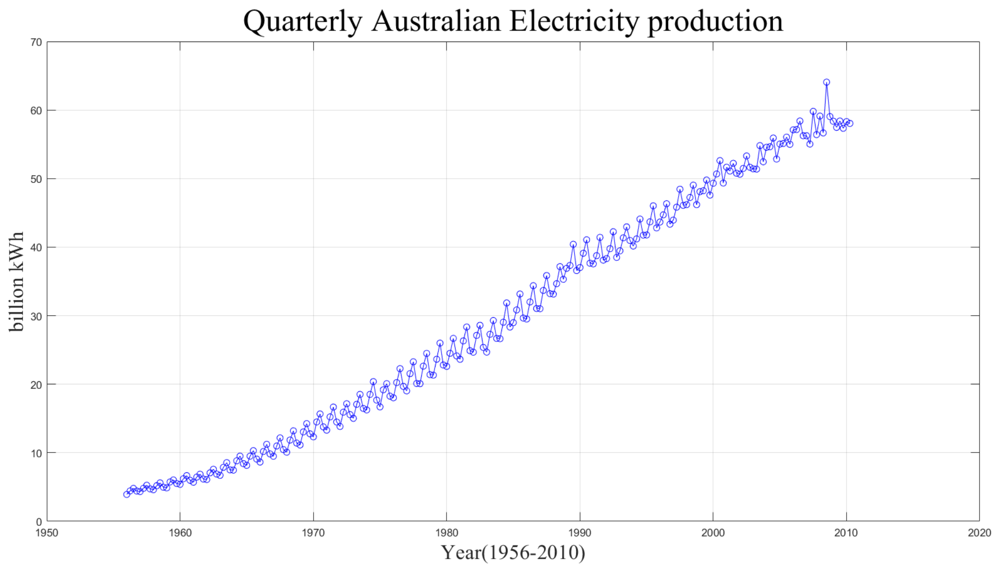

Figure 1: Australian quarterly electricity production from 1956 to 2010.

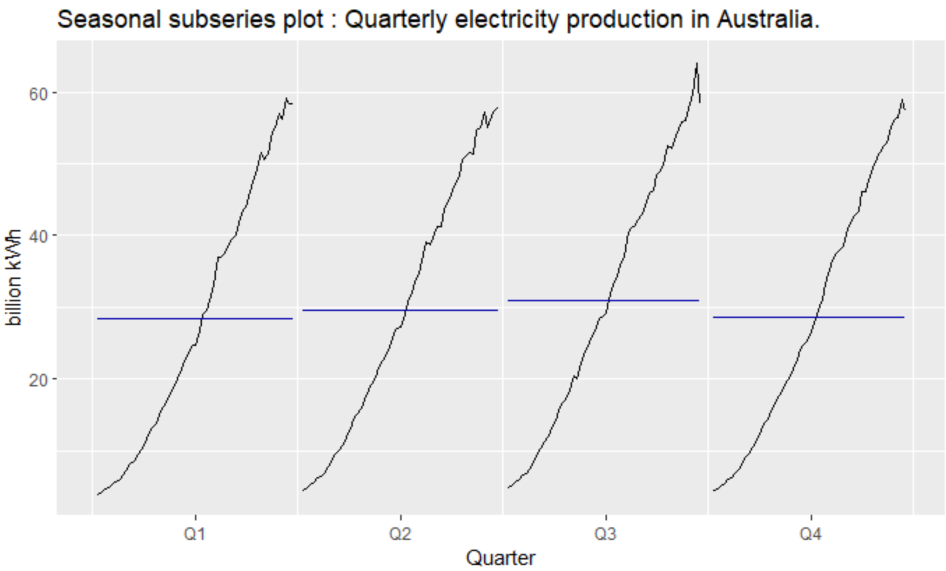

Figure 2: Seasonal subseries plot of quarterly electricity production in Australia.

From Fig.1, we can see that there is an apparent increasing trend. As we collect the data for each quarter in separate mini time plots (Fig.2), we can see that there is a strong seasonality. In addition, there are neither any abrupt changes nor any biased data. Based on these facts, we choose to model the data by the classical decomposition model:

$$X_t = m_t + s_t + Y_t$$

where m_t is the trend component, s_t is the seasonal component ($s_{t+d} = s_t$, $\sum_{j=1}^{d} s_j = 0$), and Y_t is the residual stationary time series ($EY_t = 0$).

2.2 Trend and seasonality removal

- Step 2: Removing the trend and seasonal components

The method of differencing is used to eliminate the trend and seasonal components of the time series. With seasonality of period $d = 4$, we obtain

$$\nabla_4 X_t = X_t - X_{t-4} = \left(1 - B^4\right) X_t = m_t - m_{t-4} + Y_t - Y_{t-4}$$

Thus, the seasonality is removed, as shown in Fig.3.

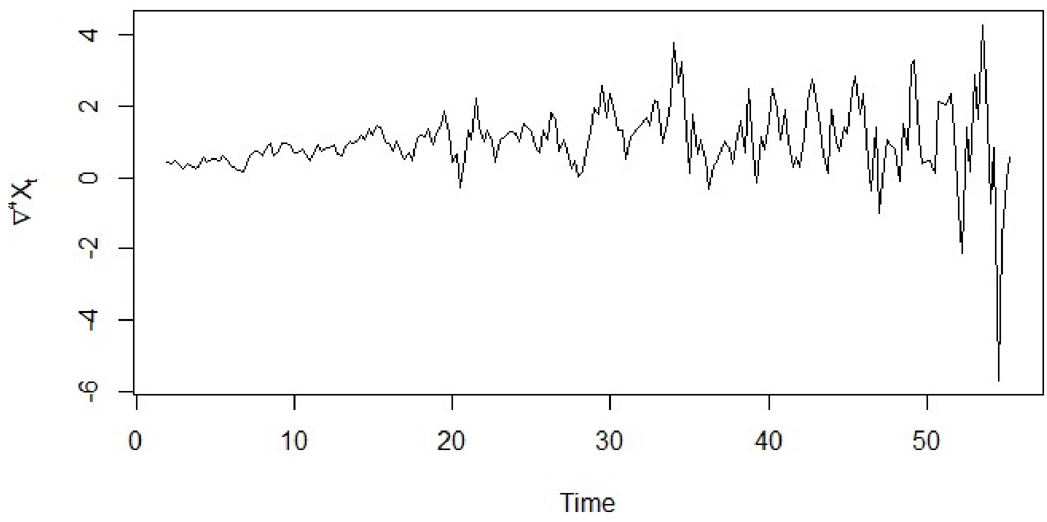

Figure 3: Time series of the lag-4 difference of the quarterly electricity production in Australia.

Close inspection of Fig.1 would suggest that there is a linear increase over time, so we'll set `differences = 1` here to remove the trend, as shown in Fig.4.

$$\nabla\nabla_4 X_t = \nabla(m_t - m_{t-4}) + \nabla(Y_t - Y_{t-4})$$

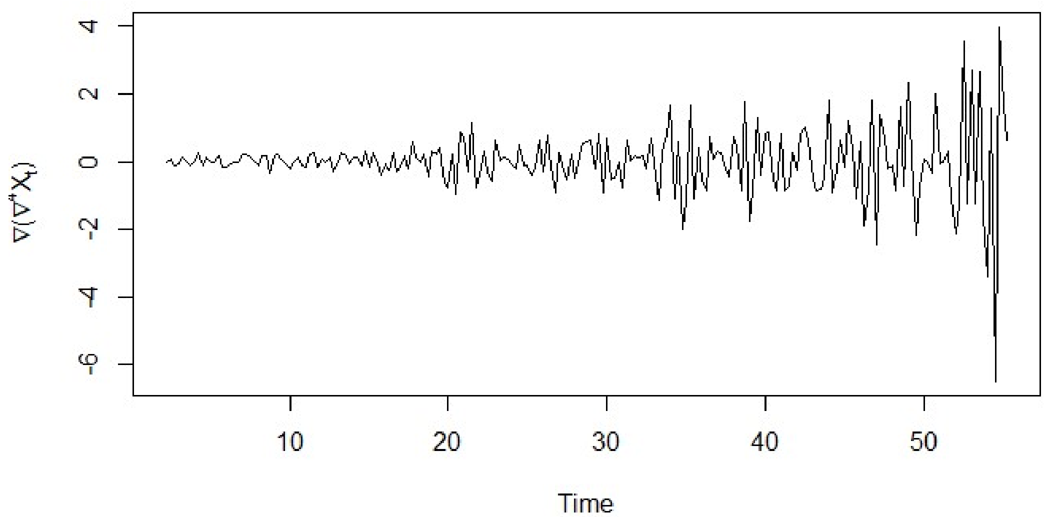

Figure 4: Time series of the first-difference of the lag-4 difference quarterly electricity production in Australia.

To verify that the removal of trend and seasonality is successful, we use the Augmented Dickey-Fuller test (ADF test) to check the the stationarity of the time series. Using the null hypothesis that the time series is not stationary, we can reject it if the test yields a p-value smaller than 0.05. The test is conducted in RStudio by the function `adf.test()` in the library `tseries`. The result below shows stationarity of $\hat{X}_t$, since the p-value is less than 0.05.

```
Dickey-Fuller = -12.097, Lag order = 5, p-value = 0.01
alternative hypothesis:  stationary
```

2.3 Model selection

- Step 3: Deciding a stationary model and using it for inference

Before fitting the stationary time series $\hat{X}_t$ with a standard model, we first calculate both ACF and PACF, since these are important tools when trying to identify the appropriate order of p and q. Our aim is to find an appropriate ARIMA model based on the ACF and PACF shown in Fig.5.

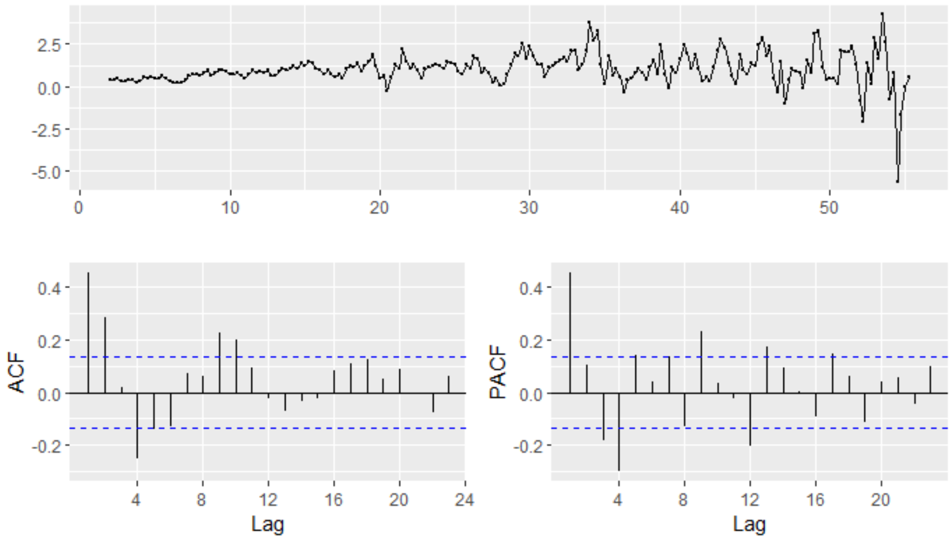

Figure 5: Correlogram of the ACF and PACF of transformed data.

Since the seasonality is already proven, it is suitable to fit the time series to a seasonal ARIMA model.

3 SARIMA model

3.1 Fitting model parameters manually

Our aim is to find an appropriate ARIMA(p,d,q) model based on the ACF and PACF shown in Fig.5. The significant spike at lag 1 in the ACF suggests a non-seasonal MA(1) component, and the significant spike at lag 4 in the ACF suggests a seasonal MA(1) component. Consequently, we begin with an ARIMA(0,1,1)(0,1,1)[4] model (Fig.6).

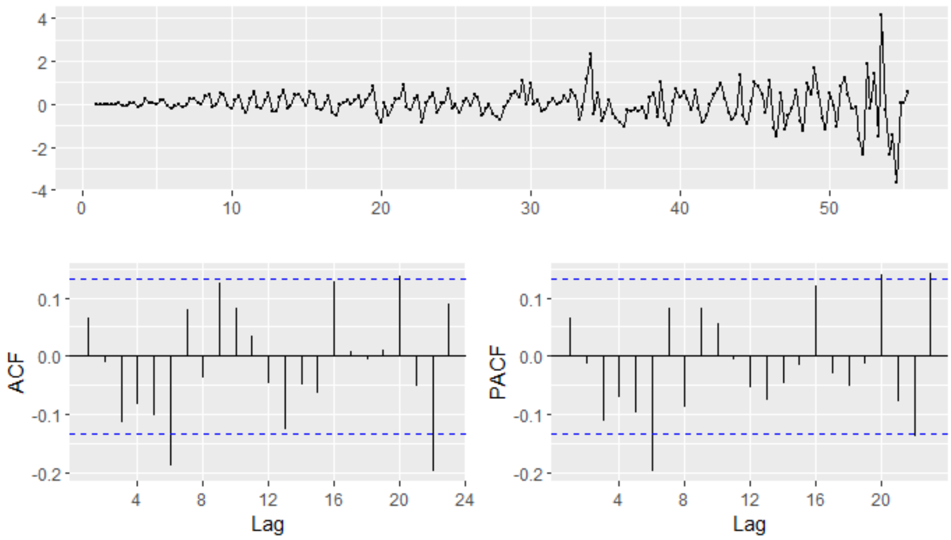

Figure 6: Residuals from the fitted ARIMA(0,1,1)(0,1,1)[4] model for the quarterly electricity production in Australia data.

From Fig.6, both the ACF and PACF show significant spikes at lag 6, indicating that some additional non-seasonal terms need to be included in the model. Since the estimation of p, q and P, Q by ACF and PACF are based on empirical data, other models with nearby parameter values still need to be considered. The best model structure is selected based on the minimum AICc value, which is shown in Tab.1.

Table 1: Parameters

Model Structure	**AICc**	**BIC**	**Model Structure**	**AICc**	**BIC**
ARIMA(0,1,1)(0,1,1)[4]	483.48	493.45	ARIMA(2,1,1)(0,1,1)[4]	478.77	495.29
ARIMA(0,1,2)(0,1,1)[4]	482.08	495.34	ARIMA(2,1,2)(0,1,1)[4]	480.67	500.43
ARIMA(0,1,3)(0,1,1)[4]	479.36	495.88	ARIMA(2,1,3)(0,1,1)[4]	/	/
ARIMA(1,1,1)(0,1,1)[4]	479.74	492.43	ARIMA(3,1,1)(0,1,1)[4]	480.5	500.26
ARIMA(1,1,2)(0,1,1)[4]	479.65	496.17	ARIMA(3,1,2)(0,1,1)[4]	481.63	504.61
ARIMA(1,1,3)(0,1,1)[4]	480.35	500.11	ARIMA(3,1,3)(0,1,1)[4]	472.63	498.82

Consequently, we choose ARIMA(3,1,3)(0,1,1)[4], whose residuals are plotted in Fig.7.

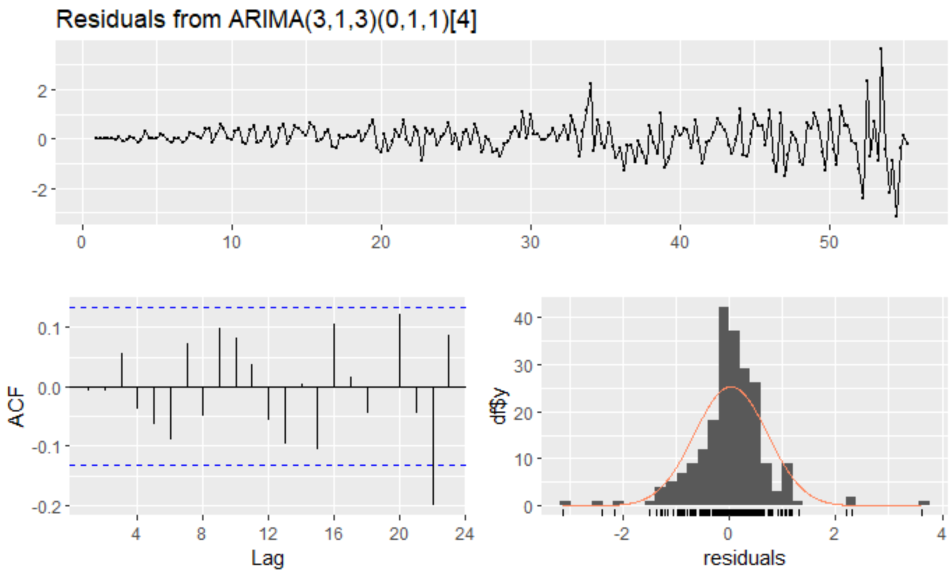

Figure 7: Residuals from the fitted ARIMA(3,1,3)(0,1,1)[4] model.

```
Ljung-Box test
data:  Residuals from ARIMA(3,1,3)(0,1,1)[4]
Q* = 9.2494, df = 3, p-value = 0.02615
Model df:  7.  Total lags used:  10
```

The residuals from this model are shown in Fig.7. The histogram of the residuals seems to be normally distributed with a mean around zero. There is still one significant spike in the ACF, and the model fails the Ljung-Box test. The model can still be used for forecasting, but the prediction intervals may not be accurate due to the correlated residuals.

3.2 Fitting model parameters with `auto.arima()`

Another approach to estimate the parameters of the seasonal ARIMA model is to use the function `auto.arima()` in RStudio. The result is shown in Tab.2 and Fig.8. However, the model still fails the Ljung-Box test. Sometimes it is just not possible to find a model that passes all of the tests.

Table 2: Coefficients

`ARIMA(1,0,1)(1,1,2)[4] with drift`						
Parameters	**ar1**	**ma1**	**sar1**	**sma1**	**sma2**	**drift**
Estimate	0.9158	−0.4411	0.8928	−1.6921	0.7737	0.2299
s.e.	0.0467	0.0942	0.1420	0.1425	0.0954	0.0456
`sigma^2 = 0.5082:  log likelihood = -230.58`						
`AIC=475.16 AICc=475.71 BIC=498.72`						

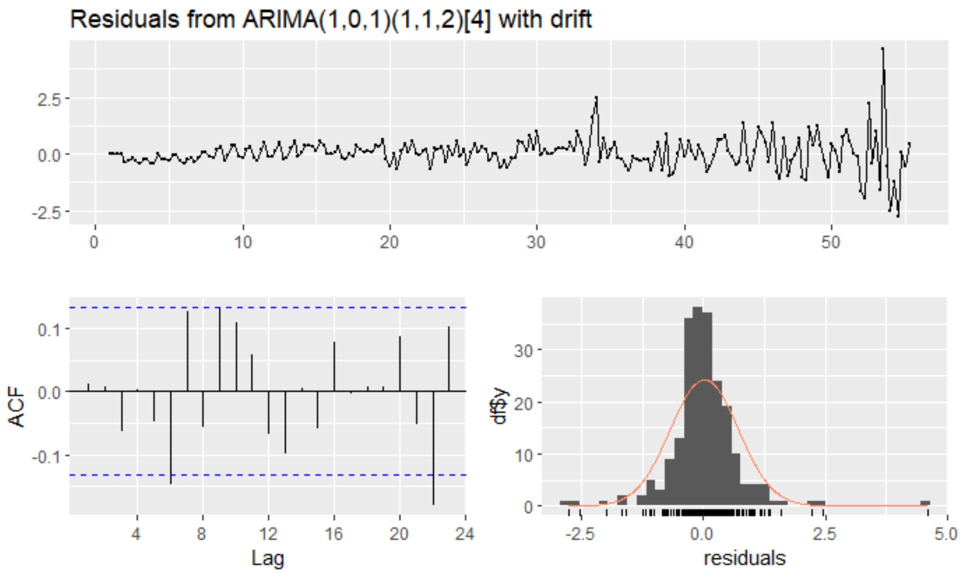

Figure 8: Residuals from the fitted ARIMA(1,0,1)(1,1,2)[4] model for the quarterly electricity production in Australia data.

3.3 Forecasting SARIMA process

None of the models considered here pass all of the residual tests. In practice, we would normally use the best model we could find, even if it did not pass all of the tests. Compared with the results shown in the previous section, the model produced by the manual approach is better since the AICc value is lower. Therefore, we forecast this SARIMA process from ARIMA(3,1,3)(0,1,1)[4], which is shown in Fig.9.

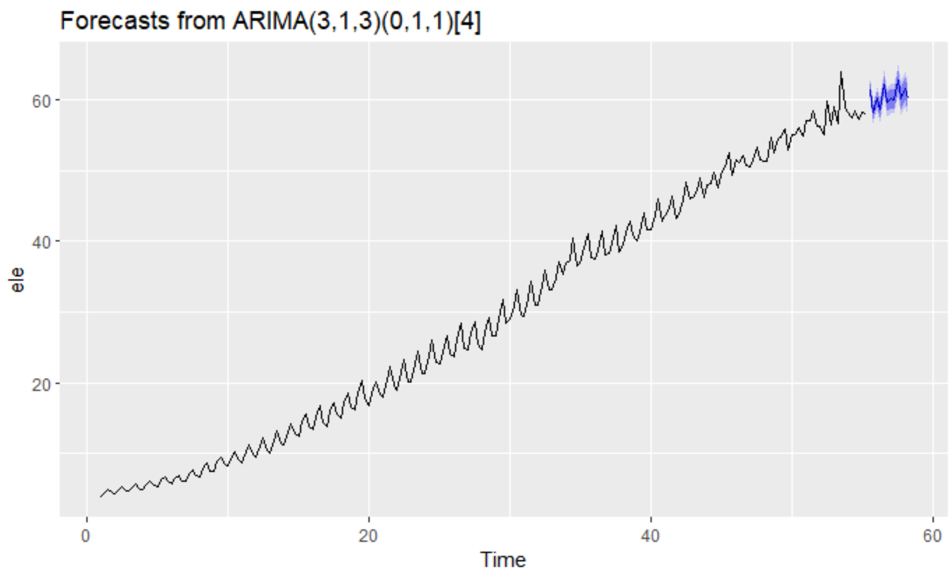

Figure 9: Forecasts from the ARIMA(3,1,3)(0,1,1)[4] model applied to the quuarterly Australian electricity production data.

The forecast follows the historical pattern and seems to be relatively stable and the prediction intervals allow for the data to trend upwards and downwards during the forecast period.

4 Conclusion

4.1 Results

In this report, the time series was modelled by the classical decomposition model. Trend and seasonality was detected and removed from the data set using the differencing method. A seasonal ARIMA(3,1,3)(0,1,1)[4] model was fitted after evaluating and finding the smallest AICc value for different parameter values. After analysing the residuals, the model was deemed satisfactory. Lastly, a forecast was made, and it provided a relatively reliable prediction.

4.2 Difficulties and further discussion

During the modeling procedure, we first conducted the model fitting mannually according to ACF and PACF. However, even the model structure with the smallest AICc value failed the Ljung-Box test, with a few spikes in the ACF. This suggests some correlated residuals and the corresponding forecast may not be accurate enough. Next we tried using the automatic ARIMA algorithm. However, the residuals still failed the test. To further investigate and make a more accurate model, we may need to consider the ARCH or GARCH models.

Alternatively, we could also use more advanced forecasting methods, such as Neural network models, which is an artificial-intelligence based method. The procedure is based on training and is more efficient and accurate.

References

[1] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on <current date>.

[2] Holmes, E. E., M. D. Scheuerell, and E. J. Ward. Applied time series analysis for fisheries and environmental data. Edition 2021.

[3] P.J. Brockwell and R.A. Davis: Introduction to Time Series and Forecasting, Springer.

Peer review: the report of Group 12

The report is well structured and the analysis of the chosen data is clear and thorough where relevant methods from the course are included. A good understanding of the data plot was shown when the multiplicative time series was chosen instead of the classical additive time series, which further allowed them to log-transform the data. Methods such as the Augmented Dickey-Fuller test and the unit root test were implemented to confirm that the time series was stationary. Further, they properly used the results from ACF and Partial ACF to provide a suitable assumption for the choice of model and verified it with the help of the Akaike Information Criterion and Bayesian Information Criterion.

To provide more information to the model, methods for the diagnostics of the residuals such as the Ljung-Box test and a QQ-norm test could be used. In addition, the authors seem to trust `auto.arima()` function without any conditions. As we all know, this automatic algorithm in R is only based on finding the smallest AICc value. However, this is not a guarantee that the selected model structure is the fittest one. It is possible that this model still could not pass all the residual tests. Therefore, residual tests must be conducted to prove this point. Lastly, the results from the ADF test can be presented in a table for more clarity.

Part B. Discussion of research paper 8

1 Summary

This research paper, titled Short-term Electricity Load Forecasting with Time Series Analysis, mainly discusses the ARIMA and SARIMA model, with the applications in forecasting short-term electricity load demand. Statistical forecasting in this field has become more and more crucial than ever for utility companies due to increased deregulation of the energy market. From this perspective, time series analysis is a significant tool for them to support their decisions about the operation, planning, and maintenance. Based on the facts above, this paper is of high research value.

To begin with, the preliminary analysis proposed by this paper is rigorous since it considers four different types of load forecasting. This makes sense due to the complexity in changes of electricity demand in terms of different time dimensions. The authors also use different models to generate these different types of forecasts, which shows accuracy in analysis. It is worth affirming that the authors also mention the uncertainty in its time series, for example, the deployment of the Smart Grid. In order to forecast the time series with the greatest accuracy, the paper also compares different forecasting methods, including ARIMA, SARIMA, ANN, MLR and other advanced methods. This makes the forecasting results more convincing to readers. Additionally, unlike unwarranted predictions, this paper forecasts the existing data based on the chosen model structure. Through the error between forecasting results and actual data, readers can have an intuitive understanding of the strengths and weaknesses of the chosen model.

2 Model assumptions

In terms of time series modelling, the paper uses a slightly different classical decomposition model introduced in the course content, which has an additional cyclical component $c(t)$.

Aside from basic model assumptions, the paper first discusses three basic concepts, which are stationary ARMA models, non-stationary ARIMA models and non-stationary SARIMA models. The description of this part is a little bit redundant since the equations are all basic ones and there is no need to display them again which can be easily obtained from textbooks and references. Moreover, Eq.(2) to Eq.(4) are only general rules and there is no doubt that it could be expressed in other forms. Although basic concepts described in this section seems to be redundant, the *Accuracy Measurements and Errors Calculation* part is the icing on the cake for the analysis, which is a strong tool to assess performances and to make comparisons between similar models.

3 Model analysis

In the section of *Case study*, this paper constructed a model for time series forecasting with the purpose of showing the process of the analysis in detail. The descriptions in this section largely determined how restrictive the obtained results were. To begin with, the author simply used the arithmetic averaging method to handle the missing and extra observations. Although this might seem to be the fittest method to deal with the data pre-processing, more evidence and reasons on the feasibility must be presented. Otherwise the analysis won't be convincing since it is possible that there might be some biased points happening in the interval, especially for this kind of data with large uncertainty.

In the part of *Model Fitting and Parameter Estimation*, the author improved the classical decomposition model with an additional cyclical component. This improvement is rigorous and corresponds to reality. However, the rigour of the following part is questionable. The reason of the existence of trend component makes sense but the identification of directly using a linear least square regression to estimate lacks a convincing proof and needs further demonstrations. Even though the result in Figure 3 shows good, the abrupt and direct use of the LLSR model still seems confusing. Again in terms of the seasonality and the cyclicality adjustment, LLSR model was used to refine the trend component. This procedure is rigorous and improves the accuracy of the analysis.

After removing the trend, seasonal and cyclical components, the author unjustifiably confirm this as stationary data. This needs to be improved instead of only using a sentence "*Provided that the deterministic components have been successfully isolated, a time series plot should now exhibit the characteristics of a stationary time series*".

Finally, when it comes to model structure fitting, although the author mentions *Accuracy Measurements and Errors Calculation* in previous sections, we do not find any further explanation after model selection. Data visualization needs to be done to illustrate which model structure is the best one.

3 Robustness and potential use in practice

In this paper, the author discusses different models parameters and different modelling methods. In order to improve the robustness of the analysis, the author mentioned that they conducted additional experiments based on Artificial Neural Network (ANN), Multiple Linear Regression (MLR), etc. However, they do not provide detailed proof of why ARIMA and SARIMA model is the fittest one. This might affect the rigor of the analysis. In the end, the author mentions that ARIMA and SARIMA have limitations when it comes to producing accurate forecasts. This makes sense since we also encountered the same problem in part A. However, we do not see the authors provide certain solutions to that. According to what we mention above, we believe the analysis lacks robustness since ACF and PACF still show correlated residuals. And most importantly, the paper does not deal with uncertainty due to rapid change in technology, etc.

In conclusion, this model is very useful to forecast electricity loads for various reasons. Load forecasting is a technique used by power companies to predict the future load demands. It is important to ensure that there would be enough generation to cope with the increasing demand of electricity. As presented in this paper, modeling electricity consumption by using a time series approach has shown to be successful. However, it is not certain that the model presented would generate reliable forecasts today due to the change of influencing factors. The rapid innovation of technology and the increasing efficiency of power generators alongside the increasing demand presents a new set of influences to be considered.