Time Series Analysis of ESB Connections

Ebru Şakar
Department of Statistics
Middle East Technical University
Ankara, Turkey
e214634@meu.edu.tr

Abstract— **In this paper, SARIMA, ETS, TBATS, PROPHET and NN approaches are provided to forecast quarterly electricity connections processes by using R-Studio. After splitting data into training and test sets, several data preprocessing techniques, such as anomaly detection and stationarity checking, are conducted. MAPE and MASE are used for comparing forecast accuracy on both train and test sets to find the best fit model.**

Keywords—Time series analysis, ESB Connections, SARIMA, ETS

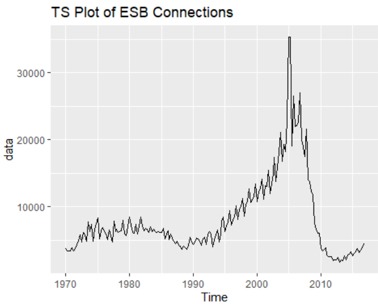
I. INTRODUCTION

ESB Networks is a company which provides energy services in Ireland. They also maintain the flow of electricity to 2.3 million customer premises that contain domestic, commercial and industrial customers.

This study is aimed not only to determine a good model that can be used to predict new electricity connections but also to understand timely patterns in the data. Hence, quarterly new ESB Connections dataset is analyzed with different types of data forecasting methods such as SARIMA, ETS, TBATS, Prophet and Neural Network. Then, to find the best fit model according to their forecast accuracy, the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE) values are compared by using R-Studio.

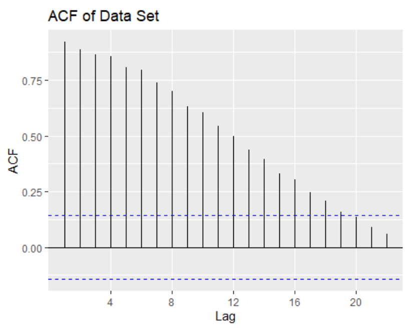
II. DATA DESCRIPTION AND PREPROCESSING

The data set is taken from https://www.gov.ie/en/publication/6dc45-esb-connections/. This dataset of new electricity connections which are recorded every month by ESB Networks measures the number of homes connected to the electricity grid and thus becoming available for use. The data set includes the total number of Local Authority Housing, Voluntary & Co-operative Housing and Private Houses, and contains 188 observations recorded from 1970 to 2016 quarterly.



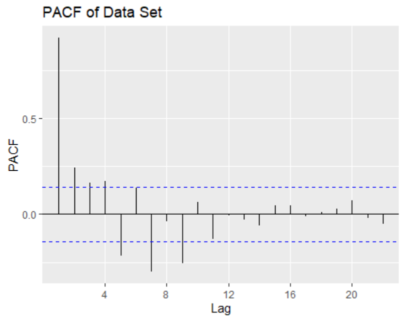
Graph 1*: Time Series Plot of Data Set*

As seen in Graph 1, the series has some ups and downs over time. Besides, it shows increasing and decreasing patterns at different periods. It seems not stationary. However, further research should be conducted to test these hypotheses.



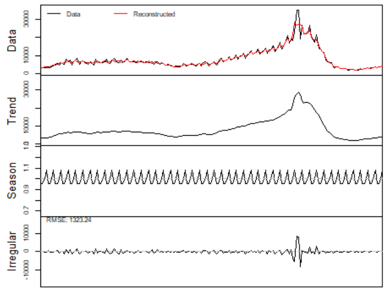
Graph 2*: ACF Plot of Data Set*

As can be seen in the ACF plot, there is a linear decay which indicates that the process is not stationary.



Graph 3*: PACF Plot of Data Set*

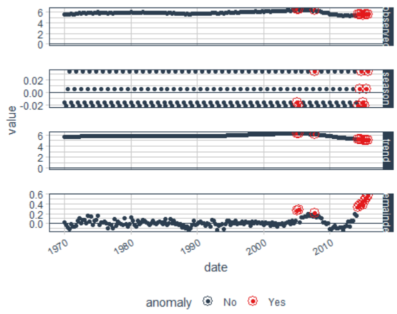
As it is decided that the process is non-stationary according to ACF and time series plot, there is no need to interpret the PACF plot in this case.



Graph 4*: Decomposition Plot*

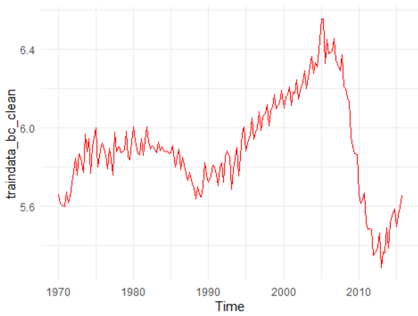
The trend is not linear, there are clear seasonality and some irregularity as well.

First of all, the data set is divided into a train set and test set which includes the last 4 observations of the data. After generating a lambda value for the data, it can be seen that the transformation is needed since the lambda value is different than 1 (lambda = -0.098). For that reason, Box-Cox transformation is applied to stabilize the variance. Then, anomaly detection is applied to the train data.



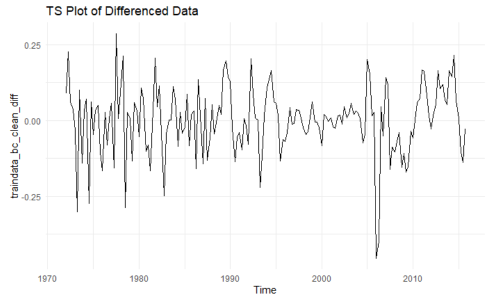
Graph 5*: Anomaly Detection Plot*

Thus, it is shown that the series has anomalies. This problem is solved by replacing outliers with estimated values.



Graph 6*: Time Series Plot After Anomaly Detection*

The function does not detect the outlier in the data. Afterwards, the Hegy Test shows that one regular difference is required to pass the stationarity tests. According to the Canova-Hansen test, one seasonal difference will be enough to remove trends in the series. After taking one regular and one seasonal differencing, there is no seasonality problem in ACF plot. Besides, it seems stationary by looking at the result of the HEGY Test.

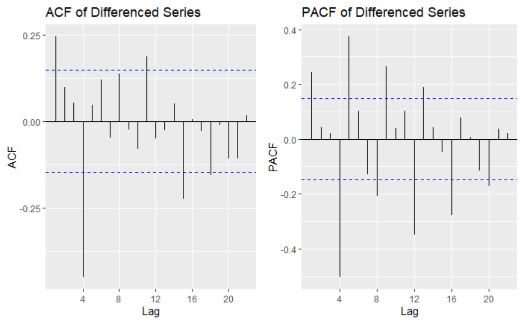


***Graph** 7: Time Series Plot of Differenced Data Set*

After taking one regular and one seasonal difference, it can be said that the series seems stationary around mean 0.

III. MODEL SUGGESTION

The ACF and PACF of the differenced data are examined to determine possible models.



***Graph 8**: ACF and PACF Plot of Differenced Data Set*

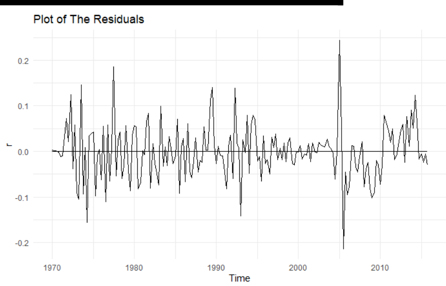
It is seen that ACF and PACF cut off after lag 1 for nonseasonal processes. Therefore by looking both ACF and PACF plot, the suggested models are SARIMA(0,1,1)(1,1,1)[4], SARIMA(0,1,1)(2,1,1)[4], and SARIMA(0,1,3)(0,1,1)[4] for this data set. All models are significant.

IV. MODELLING AND DIAGNOSTIC CHECKING

After identifying some tentative models, the AIC, BIC, log-likelihood values are compared to find the most appropriate one. The model having minimum AIC and BIC value is SARIMA(1,1,3)(0,1,1)[4]. And afterwards, diagnostic checks of the model are conducted.

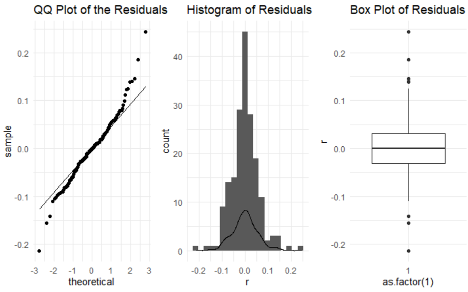
***Table 1:** Summary of Model*

[1]	Model					
[2]	Coefficients	ar1	ma1	ma2	ma3	sma1
[3]		0.8577	-1.2414	0.2672	0.1856	-0.7882
[4]	s.e.	0.0753	0.1041	0.1216	0.0753	0.0589
[5]	sigma^2 estimated as 0.00377: log likelihood=246.43					
[6]	AIC=-480.86 AICc=-480.37 BIC=-461.74					



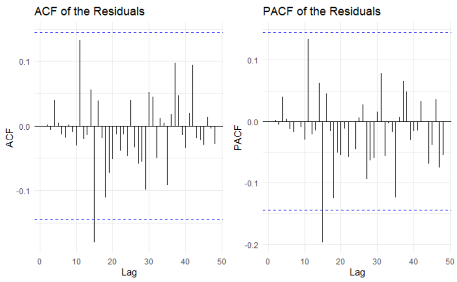
***Graph 9**: Plot of Residuals*

They are scattered around zero and it can be interpreted as zero mean. Normality assumptions are checked with the visual inspection tool Q-Q plot, and the Shapiro-Wilk test.



Graph 10*: QQ plot of the standard residuals*

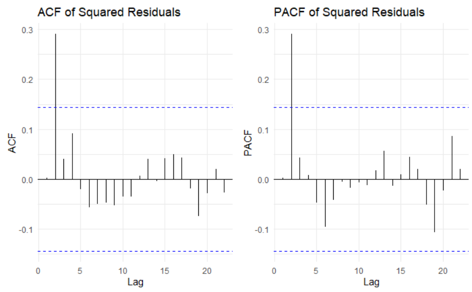
QQ Plot shows that the residuals of the model seem to have an S shape. Besides, it has outliers. Similarly, the histogram of the residuals shows that they might have a symmetric distribution but it has outliers. Addition to this, the box plot shows that the data may have a symmetric shape, but there are lots of outliers which may violate normality assumption among the residuals. To be sure about normality, Shapiro-Wilk should be applied. Unfortunately, Shapiro-Wilk shows that errors do not distribute normally($p<0.05$). Since the lambda value is almost equal to 1, it can be said that the transformation is not needed. Hence, the Box-Cox transformation is not applied. Thus, we have residuals with non-normal distribution.



Graph 11*: ACF and PACF of the Residuals*

Both plots show that some spikes are out of the white noise bands. However, they are not significant. Besides, according to the results of the Ljung-Box Test, the residuals of the model are uncorrelated ($p>0.05$). Consequently, our residuals are uncorrelated because formal tests are more reliable than plots.

The last assumption checked is the heteroscedasticity of the residuals. To test this assumption, we can look at the ACF and PACF plot of squared residuals.



Graph 12*: ACF and PACF of the Squared Residuals*

Both plots show that some spikes are out of the white noise bands that is an indication of heteroscedasticity problem. Besides, according to the results of ARCH Engle's Test, there is a heteroscedasticity problem ($p<0.05$), and it confirms the presence of arch effects. As a result, error variance is not constant and it should be modelled using ARCH and GARCH methods.

After fitting the SARIMA model, to find the best exponential smoothing model, Holt's Winter Additive and Multiplicative Methods are applied. Then, the performances of both methods are compared with respect to MASE and MAPE values. Last of all, the best ETS model for the process is given below.

Table 2: *Summary of ETS Model*

Forecast method: ETS(M,Md,M)
Smoothing parameters:
alpha = 0.3892
beta = 0.1248
Gamma = 0.063
phi = 0.9371
Initial states:
l = 3211.2203

b = 1.0043
s = 1.005 1.025 0.9483 1.0217
sigma: 0.1595
AIC AICc BIC
3539.839 3541.111 3571.989

In the end, it seems that the best forecasting method ETS(MMM) according to MASE and MAPE criteria. After fitting the model, the residuals of the ETS model are checked by the Shapiro-Wilk test, and it is seen that they do not follow normal distribution ($p<0.05$).

The study then continues with the evaluation of accurate PROPHET models. It is played with the parameters so that better prophet models can be obtained. These parameters are seasonality mode, yearly seasonality, weekly seasonality, daily seasonality. After that, their performances are compared with respect to MASE and MAPE criteria. Last of all, the residuals of the best-fitted PROPHET model are checked by the Shapiro-Wilk test which proves that they do not follow normal distribution ($p<0.05$).

Table 3: *Summary of TBATS Model*

TBATS (0, {0,0}, 0.897, {<4,1>})
Call: tbats(y = train)
Parameters
Lambda: 0
Alpha: 0.2678583
Beta: 0.1430728
Damping Parameter: 0.89651
Gamma-1 Values: 7.713294e-05
Gamma-2 Values: 0.0001838556
Sigma: 0.1538165
AIC: 3532.566

After fitting the model, the residuals of the TBATS model are checked by the Shapiro-Wilk test which proves that they follow normal distribution ($p>0.05$).

Afterwards, the study continues with the evaluation of accurate NNETAR models. It is played with the parameters so that better nnetar models can be obtained. These parameters are number of non-seasonal lags used as inputs(p), number of seasonal lags used as inputs(P), number of nodes in the hidden layer(size) and number of networks to fit with different random starting weights(repeats). Then, their performances are compared with respect to MASE and MAPE criteria. Last of all, the best NNETAR model for the process is given below.

Table 4: *Summary of NNETAR Model*

Forecast method: NNAR(15,1,7)[4]
Model Information:
Average of 50 networks, each of which is
a 15-7-1 network with 120 weights
options were - linear output units

After fitting the model, the residuals of the NNETAR model are checked by the Shapiro-Wilk test which shows that they do not follow normal distribution ($p<0.05$). The train accuracy of all the models is evaluated and given below.

Table 5: *The train accuracy of models*

	ME	RMSE	MAE	MAPE	MASE	ACF1
SARIMA	-2.41e-04	5.97e-02	0.0438	0.7435	0.570	-9.4e-05
ETS	-173.63	2352.31	1074.79	11.63	0.66	0.27
TBATS	-43.162	2131.80	1090.55	11.90	0.674	0.154
NNETAR	1.1096	394.00	283.75	4.870	0.175	0.121
PROPHET	-9.262	4269.32	2894.20	48.724	1.790	0.868

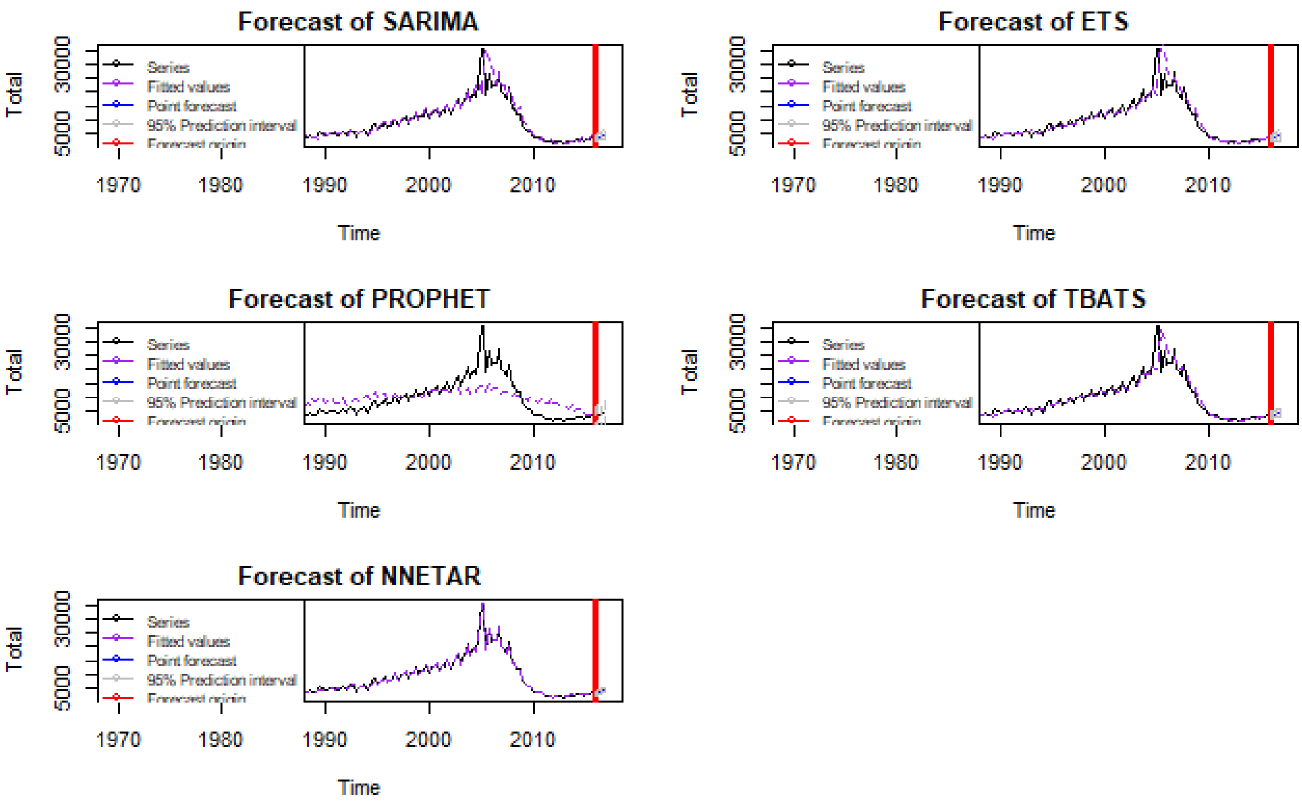
The test accuracy of all the models is evaluated and given below.

Table 6: *The forecasting performance of models*

	ME	RMSE	MAE	MAPE	MASE	ACF1
SARIMA	7.1293	64.411	54.232	1.459	0.033	-0.385
ETS	**41.77**	**81.19**	**69.42**	**1.773**	**0.043**	**0.241**
TBATS	-46.74	217.10	162.03	4.553	0.100	-0.097
NNETAR	-67.581	95.083	78.61	2.223	0.048	-0.627

PROPHET	-477.91	712.27	657.2	17.31	0.406	-0.623

Both tables show that the SARIMA model outperforms the other methods in comparison of both train and test set based on all measures. However, since the residual of the SARIMA model does not follow the normal distribution, its forecast result is not reliable. Hence, it would not be right to choose it as the best model. For that reason, it seems that ETS model, which has the second-best forecasting performance of test set based on MAPE and MASE criteria, is the best fitting model for the process, although its train accuracy is not that good.



***Graph 13**: Forecast Plots*

According to the forecast plots above, the NNETAR method has a narrower prediction interval compared to others. However, as its residuals do not follow the normal distribution, its prediction interval is not reliable. Also, its accuracy results do not support this one. Unfortunately, although there is almost a perfect match between fitted and observed values, its forecasting performance is not good enough based on MAPE and MASE values. Similarly, the residuals of ETS do not follow the normal distribution, its prediction interval is not reliable either. As a result, since the best way for the final decision is the performance comparison of the models via accuracy measures, ETS model is the best-fitted model for the process.

V. DISCUSSION AND CONCLUSION

In this study, firstly, the existence of a trend, seasonality, outliers are determined by time series plot. Secondly, the last 4 observations are referred to as test data sets. Then, it is seen that Box-Cox transformation is needed to stabilize the variance. After the transformation, the series is cleaned from anomalies. On the other hand, ACF and PACF plots show some stationarity problems. Similarly, the Hegy Test and Canova-Hansen test shows that one regular and one seasonal differencing is required to pass the stationarity tests. Therefore, the same unit root tests are applied again after the differencing, and the results state that the problem is solved. Then, some tentative SARIMA models are identified by looking at ACF and PACF plots. The model having minimum AIC and BIC values is selected as the best one.

After diagnostic checking, it is confirmed that there are some problems such as non-normal residuals and heteroscedasticity. However, the residuals are uncorrelated according to results of the Ljung-Box Test. In addition to the best SARIMA model, four different forecasting methods are applied and compared with respect to MAPE and MASE values. Finally, it may be concluded that ETS has the best performance in predicting future values compared to other models. Consequently, this study helped me to learn how to interpret ACF and PACF plots, how to solve unit root problems, how to find and compare the forecast accuracy of different methods. Shortly, I have learnt how to analyse time-series data. Based on what I have learned from this course, I tried to obtain the best model, but it is seen that further research should be conducted to solve the heteroskedasticity problem of SARIMA model.

VI. REFERENCES

[1] ESB Connections. (2020, December 5). ESB Connections by sector quarterly. Gov.Ie. https://www.gov.ie/en/publication/6dc45-esb-connections/