Time Series Analysis of California Hospitality Industry Employees

Oğuzhan AYDIN
Department of Statistics
Middle East Technical University
Ankara, Turkey
aydin.oguzhan_01@metu.edu.tr

***Abstract*— This paper presents the prediction of hospitality employees in California with different forecasting models such as ARIMA, TBATS, PROPHET, ETS, and Neural Network. R-Studio is used mainly in the analysis of this dataset. Before the analysis and forecasting part, some pre-analysis operations such as data cleaning, outlier, and anomaly checking, and so on have been applied. Then, unit root and stationary checking with some time series tests are conducted. Different types of forecast models are fitted and their performances on the test and train data are compared. Finally, best forecasting performance is estimated based on statistical model evaluation criteria.**

Keywords— Forecast, Stationary, Time Series Analysis, Diagnostic Checking, Hospitality Employees, ARIMA, ETS, Prophet, TBATS, NNETAR

I. INTRODUCTION

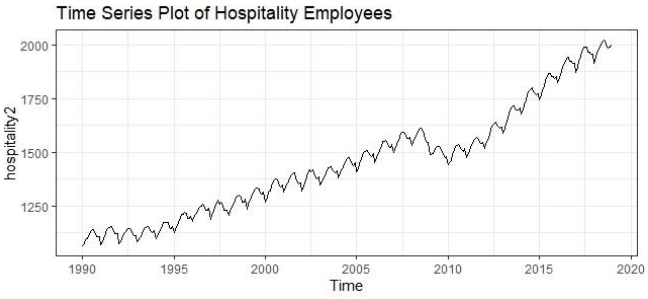
The number of employees in thousands of persons from January 1990 to December 2018 has been recorded as monthly averages. Hospitality employees of California are used by the California State Presidency and U.S. National Travel and Tourism Office to analyze the state's tourist attractions. Depending on the number of employees, the quality of hotels and accommodations can also be evaluated.

The main objective of this study is to analyze the employees in California and how more people will be employed in this field. For this purpose, the number of employees in the California hospitality industry between 1990 and 2018 is analyzed and explored.

While analyzing the time series, the future has been also forecasted by using ARIMA, TBATS, Neural Network, ETS, and Prophet models. After modeling, using statistical metrics such as mean absolute error and root mean squared error, which is used to measure accuracy, the performance of the models is compared. During the study, R Studio version 4.2.2 is used for all analyses.

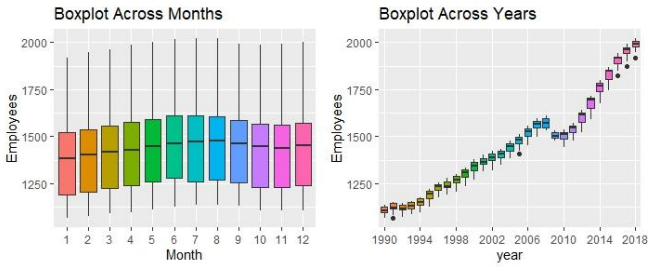
II. DATA DESCRIPTION AND PREPROCESSING

The data set is taken from https://www.kaggle.com, a subsidiary of Google LLC. The data set contains 348 observations recorded from 1990 to 2018 monthly.



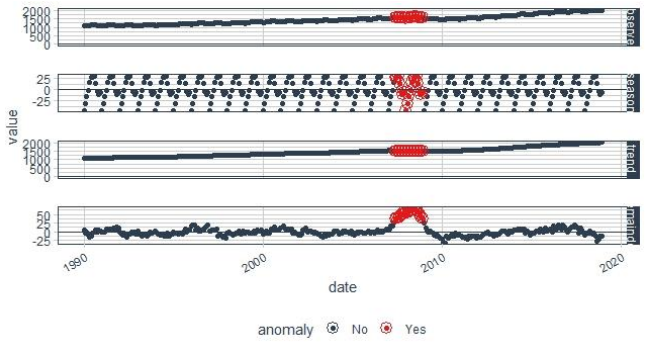
***Graph 1**: Time Series Plot of Hospitality Employees*

This graph represents the hospitality employees in California. Focusing on the plot, it is clear that the series is not stationary in the mean. There is an increasing trend. Since the line is not straight, the series might have a stochastic trend. Also, it seems that there is seasonality.



***Graph 2**: Boxplot across months and years*

From the first boxplot, since the median value for each month is not equal, it appears that we have a seasonal component each year. Also, we see that there are no outliers present monthly. From the second boxplot, we can see that there is an increasing trend. Also, there are some outliers on a yearly basis.

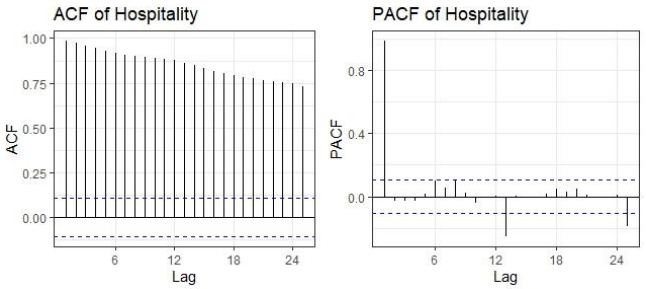


***Graph 3**: Anomaly Detection Plot*

Plots show that we have anomalies, so we need to get rid of them. So, thanks to the tsclean() function under the forecast library, the anomalies are removed and replaced by interpolated values.

After getting rid of the anomalies, the data set is divided into two train and test data sets. The last 12 observations are kept test data set, and the rest of them are train data set.

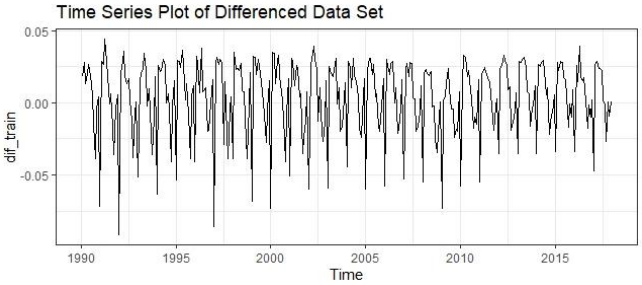
Covariance-stationarity is frequently needed in order to start the modeling process when working with time series. To stabilize the variance, we generate a lambda value for our data set. Since our lambda value is close to 0 (0.097), we could apply log transformation and box-cox transformation. I have applied box-cox transformation for the lambda value on the train data set to stabilize the variance. After achieving stationarity in variance, ACF/PACF plots and related tests are checked.



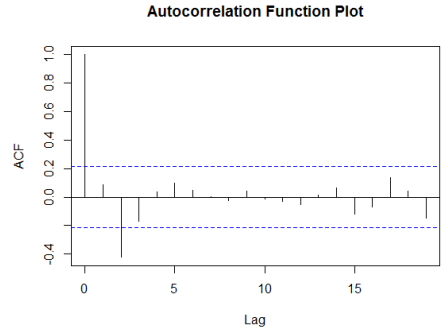
***Graph 4**: ACF and PACF Plots of Hospitality*

Focusing on the ACF plot, we can easily see that there is slow linear decay in the ACF plot. This slow linear decay indicates that the series is not stationary. Since we decide on the nonstationarity, there is no need to interpret the PACF plot. Now, check the related tests.

The following tests are applied to check the stationarity of the series: ADF Test, KPSS Test, and HEGY Test. All of these three tests suggest that the series is not stationary. Also, KPSS Test suggests that there exists a stochastic process and HEGY Test suggests that there exists both regular unit root and seasonal unit root. In the light of this information, one regular difference is taken.



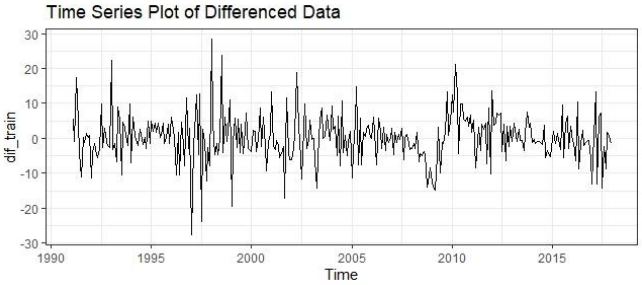
***Graph 5**: Time Series Plot of Differenced Data Set*



***Graph 6**: ACF and PACF Plots of Differenced Data Set*

After the one regular difference is taken, the same tests are applied to the differenced data set. This time, KPSS Test and ADF Test suggest the series is stationary. However, HEGY test still suggests that there exists both regular and seasonal unit root. Checking the ACF and PACF plots, it was clear that there is a slow linear decay in the seasonal lags. Therefore, one seasonal difference is taken to the differenced data set.

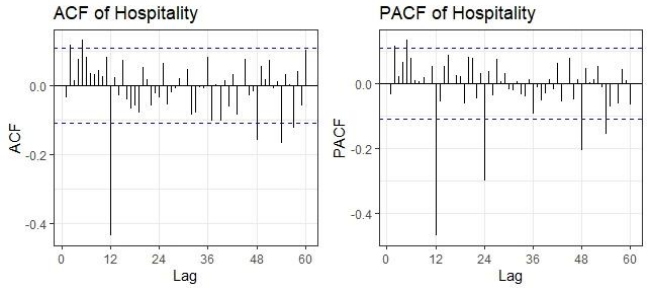
After the seasonal difference is taken, both ACF/PACF plot and related tests suggest that there is no seasonal unit root and regular unit root. The series is stationary. Also, the Canova-Hansen test suggests that the series is purely deterministic and stationary.



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

III. MODEL SUGGESTION

After obtaining the stationary series, it is time to suggest model by looking at the ACF and PACF plots. Also, we may obtain a plot using auto.arima function in R as an additional method.



***Graph 8**: ACF and PACF Plots of Stationary Data Set*

Focusing on the ACF plot, it cuts off after lag 2, and lag 5 in the regular part and cuts off after lag 12 in the seasonal part. Also, it shows exponential decay in the seasonal lags. Focusing on PACF plot, it cuts off after lag 2 and lag 5 in the regular part and cuts off after lag 24 in the seasonal part. Below are some suggested model for these plots:

- SARIMA(2,1,2)(2,1,1)$_{12}$
- SARIMA(2,1,2)(2,1,0)$_{12}$
- SARIMA(2,1,2)(0,1,1)$_{12}$
- SARIMA(5,1,2)(0,1,1)$_{12}$
- SARIMA(0,1,5)(2,1,1)$_{12}$
- SARIMA(5,1,3)(2,1,1)$_{12}$ **Suggested by auto.arima function*

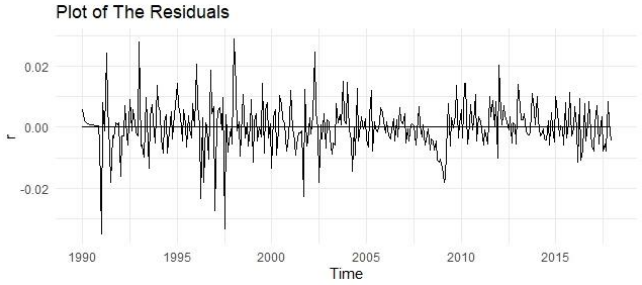
IV. MODELLING AND DIAGNOSTIC CHECKING

After determining the models, we fit the data to the model. Their significance and performance are compared with respect to some criteria under the 95% confidence level. As a result, out of these 6 models, SARIMA(2,1,2)(2,1,0)$_{12}$ and SARIMA(2,1,2)(0,1,1)$_{12}$ are found significant. After checking AIC and BIC values, SARIMA(2,1,2)(0,1,1)$_{12}$ is selected as best models since it has the lowest AIC and BIC values.

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

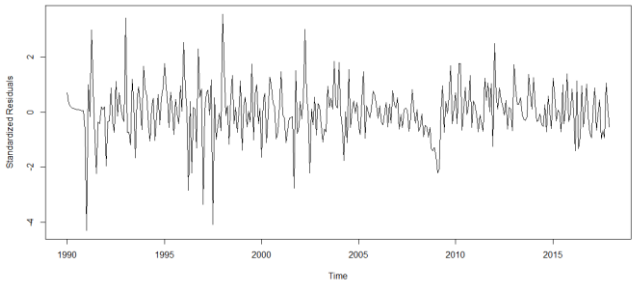
SARIMA(2,1,2)(0,1,1)$_{12}$	*AR2*	*MA2*	*SMA1*
Coefficient	-0.5714	0.6639	-0.7414
s.e	0.2272	0.1874	0.0376
AIC=-2190.04 AICc=-2189.77 BIC=-2167.37			

The goodness of fit of the model and the validity of assumption should be checked after identifying and estimating the model. There are 3 assumptions that should be satisfied: Normality, linearity and equality of variance.



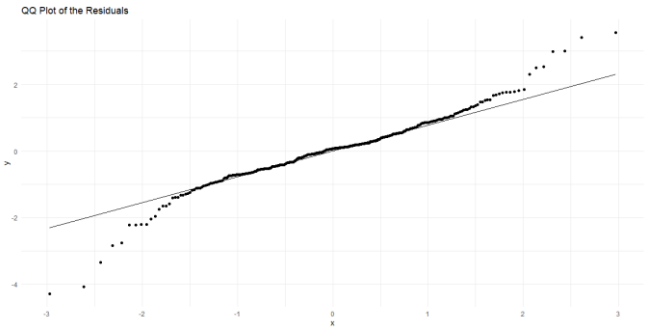
***Graph 9**: Plot of the residuals*

Residuals are scattered around zero and it can be interpreted as zero mean. Variability seems high.



***Graph 10**: Plot of the standard residuals*

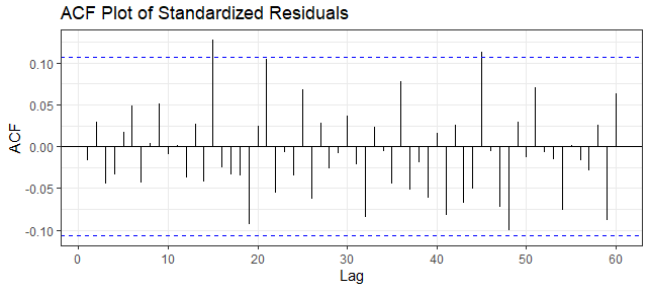
I. **Normality:** Residuals should follow the normal distribution. Normality can be checked by both visually such as Q-Q Plot, and formally such as Shapiro-Wilk and Jarque-Bera Tests.



***Graph 11**: Q-Q Plot of the standard residuals*

Q-Q Plot exhibits S-shaped. Seems symmetric however there exists outliers. Residuals may not follow the normal distribution. To be sure about non-normality, we should apply formal test. P-value for both Shapiro-Wilk and Jarque-Bera test is smaller than 0.05. Meaning residuals are not normally distributed. transformation on residuals should be applied.

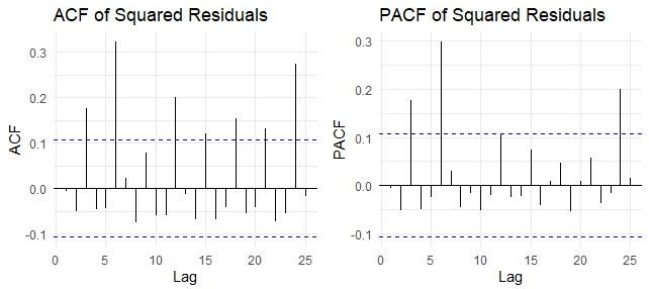
II. **Linearity:** There should be no serial autocorrelation among the residuals. To detect the linearity, there are two different ways, visual way by ACF/PACF plots of residuals, and formal way by Box-Pierce test, Breusch Godfrey test and Ljung-Box test.



***Graph 12**: ACF Plot of the standard residuals*

There are some significant spikes that are out of WN band. It might be the indication of autocorrelation. Formal tests should be applied to be sure that the residuals are normally distributed. P-values for Box-Pierce test(≈0.77), Breusch Godfrey test(≈0.83) and Ljung-Box(≈0.74) test are higher than critical value 0.05. Therefore, it can be concluded that there is no autocorrelation among the residuals. Residuals are uncorrelated.

III. **Equality of Variance:** The variance should be equal for all sample. To detect the equality of variance, there are two different ways; visual way by ACF/PACF plots of squared residuals, and formal way by studentized Breusch-Pagan test and ARCH LM test.



***Graph 13**: ACF and PACF Plots of the squared residuals*

There are some significant spikes that are out of WN band. It might be the indication of heteroscedasticity. Formal tests should be applied to be sure that if the heteroscedasticity exists. P-values for studentized Breusch-Pagan test and ARCH LM test is smaller than the critical value 0.05. It can be concluded that there is a heteroscedasticity problem. There is no constant variance over time and ARCH(lag) effects are present. It should be modelled.

The high values in the lower and upper extremes destroy the normality due to high variation. Most probably normality test on residuals will fail.

- Normality Assumption has failed.
- Serial Correlation Assumption is successful. There is no serial correlation among residuals.
- Homoscedasticity Assumption has failed. There exists heteroscedasticity.

Another type of model is exponential smoothing models, ETS. In the modelling part, train data is used and ets function under the forecast package also suggests the model structure. The model structure is suggested as additive error, additive trend, and additive seasonality. The best exponential smoothing model for the series is given below.

***Table 2:** Summary of ETS Model*

ETS(A,A,A)		
Smoothing Parameters		
Alpha	Beta	Gamma
0.8293	0.1195	0.1043
Model Selection Criteria		
AIC	AICc	BIC
3122.797	3124.721	3187.688

After fitting the model, the residuals of the ETS model is checked if normality assumption is satisfied by Shapiro-Wilk Test. P-value of the test is smaller than the 0.05. Meaning that the residuals do not follow the normal distribution.

As a third model, TBATS model is fitted to the series. TBATS model for the series is given below.

***Table 3:** Summary of TBATS Model*

TBATS	
Parameters	
Alpha	1.729
Beta	0.303
Damping Parameter	1
Gamma 1	-0.0004
Gamma 2	-0.0014
AIC	3164.171

After fitting the model, the residuals of the TBATS model are checked if normality assumption is satisfied by Shapiro-Wilk and Jarque Bera tests. P-value of both tests is smaller than the 0.05. Meaning that the residuals do not follow the normal distribution.

After TBATS model, Neural Network model is fitted to the series. In the Neural Network, past observations are considered as input variables. Also, by tunning the parameters, lowest AIC value was obtained at size equals 30 and repeat number equals 10. The model details are given below:

Table 4: *Summary of NNETAR Model*

NNAR(1,1,30)	
Size = 30	Repeat = 10
σ^2 = 0.00026	

After fitting the model, the residuals of the Neural Network model are checked if normality assumption is satisfied by Shapiro-Wilk test. P-value of the test is smaller than the 0.05. Meaning that the residuals do not follow the normal distribution.

Lastly, Prophet model has been fitted. To find the best prophet model, different models are created with different parameters changing change point and seasonality of prior scale and change point rate. Finally, parameters for the best model are decided on with default value since it has the lowest MAPE and RMSE value. The model details are given below.

After fitting the Prophet model, since the R function does not provide the residuals for Prophet model, it is calculated by subtracting response from the fitted value. As a result, it is obtained that the residuals do not follow the normal distribution since the P-value of the Shapiro-Wilk test is smaller than the 0.05.

After fitting all the models, forecast values from each method has been obtained and accuracy has been calculated. Test accuracy explains how closely fitted and observed values are related to one another. The train and test accuracy for each model is given below:

Table 5: *The train accuracy of models*

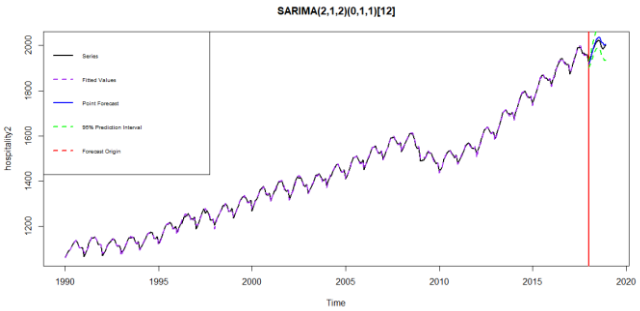
	ARIMA	*ETS*	*TBATS*	*NNETAR*	*Prophet*
ME	0.0002	0.169	0.023	0.059	**-0.0002**
RMSE	**0.008**	5.408	5.599	11.181	6.833
MAE	**0.006**	4.126	4.285	8.241	5.194
MPE	0.002	0.016	0.003	-0.002	**-0.003**
MAPE	**0.054**	0.296	0.309	0.585	0.361
ACF1	**-0.016**	0.004	-0.001	0.525	0.683

Table 6: *The forecasting performance of models*

	ARIMA	*ETS*	*TBATS*	*NNETAR*	*Prophet*
ME	**-8.692**	-0.308	2.975	-3.738	-1.557
RMSE	12.740	**6.306**	7.357	8.945	31.431
MAE	9.937	**5.035**	6.849	7.199	24.074
MPE	**-0.434**	-0.017	0.148	-0.193	-0.103
MAPE	0.498	**0.254**	0.345	0.364	1.221
ACF1	0.760	**0.362**	0.396	0.413	0.632

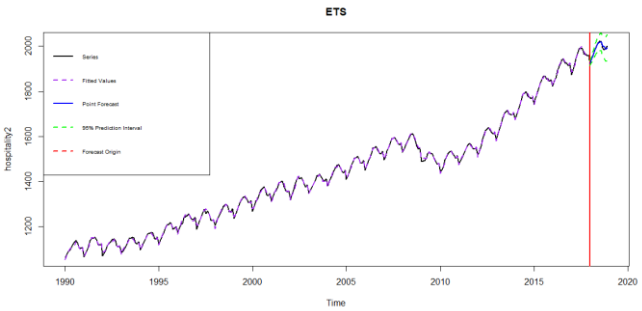
The train accuracy of models shows that ARIMA forecasting has the best fitting while NNETAR has the worst fitting with respect to the RMSE and MAPE values. The test accuracy of the model table shows that ETS has the best forecasting performance while Prophet has the worst. ETS model is followed by TBATS, NNETAR, ARIMA forecasting and Prophet, respectively.

Five forecast graphs are in below. Each graph was created by each forecast model. These graphs show the forecasting performance for each model. For all graphs, black line represents the original data set, purple dashed line represents the fitted value, blue line represents the point forecast, green dashed lines represent 95% upper and lower prediction intervals and vertical red line represents the forecast origin which is year of 2018.



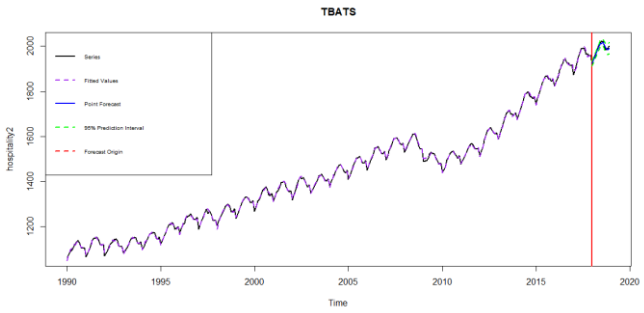
Graph 14*: Forecast Plot of SARIMA*

Graph 14 represents the forecast obtained by SARIMA modelling. Fitted values are perfectly match with the original series. Prediction interval for the forecast value includes the original value but prediction interval is large compared to the ETS and TBATS. Forecast is plausible.



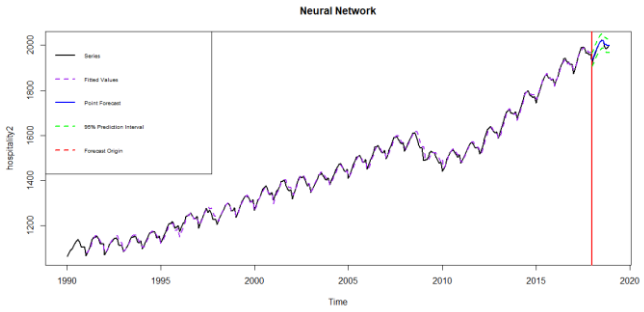
Graph 15*: Forecast Plot of ETS*

Graph 15 represents the forecast obtained by ETS modelling. Fitted values are close to the original series. Prediction interval for the forecast value is close to the original series and contains the original value. Point forecast is also nearly match with the original series. Forecast is plausible.



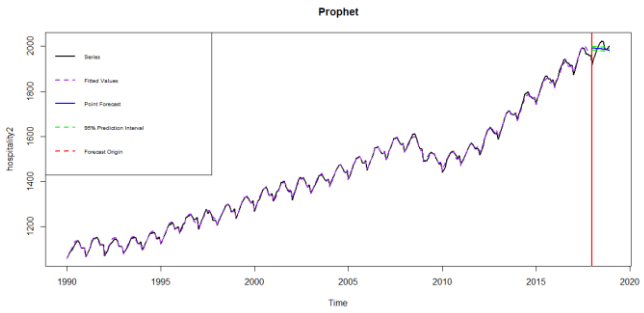
***Graph 16**: Forecast Plot of TBATS*

Graph 16 represents the forecast obtained by TBATS modelling. Fitted values are close to the original series. Prediction interval for the forecast value is small, close to the original series and contains the original series. Forecast is plausible.



***Graph 17**: Forecast Plot of Neural Network*

Graph 17 represents the forecast obtained by Neural Network modelling. Although there are some differences in certain points, it can be said that fitted values are close to the original series. Prediction interval for the forecast value is close to the original series and includes the original series. Forecast might be plausible.



***Graph 18**: Forecast Plot of Prophet*

Graph 18 represents the forecast obtained by Prophet modelling. Fitted values are close to the original series although there are little differences in certain points. Prediction interval for the forecast value is not good and does not contain the original values. Forecast is not plausible.

V. DISCUSSION AND CONCLUSION

In this study, the first step was to check the time series plot of the whole series. In the plot, it was obtained that there exists non-stationarity, increasing trend and seasonality in the series. Then, via both monthly and yearly boxplots, outliers and seasonality were checked. There were some outliers on the yearly basis. After anomaly detection, data cleaning and dividing, stationarity of the series was checked. First, box-cox transformation applied on the train set for variance stabilizing. Then, via the ACF and PACF plots, and formal tests such as KPSS and ADF tests, it was obtained that the series is not stationary. Since there was a stochastic trend, differencing on the series was applied on the series. After one regular difference, even though KPSS and ADF tests indicate that the series is stationary, ACF/PACF plots and HEGY test indicates that there exists a slow linear decay in the seasonal lags. HEGY test also suggested that there is a seasonal unit root. To overcome this problem, one seasonal difference was taken on the differenced series. After the seasonal difference, all the formal tests which are KPSS test, ADF test, HEGY test, Canova Hansen test and ACF/PACF plots suggested that the series is stationary. Then, looking at the ACF/PACF plots and using auto.arima function in R, different SARIMA models were obtained. After obtaining the SARIMA models, significance and performance of the models were compared based on model evaluation criteria such as BIC and AIC.

After fitting the best model, diagnostic checks were applied to the model. However, model has failed in two assumptions out of the three. Normality and homoscedasticity assumptions were failed. For nonnormality, transformation should be applied on the residuals. For heteroscedasticity, ARCH(lag) effects are present and it should be modelled. However, it was confirmed that the errors are uncorrelated. In addition to the best ARIMA models, four different forecasting methods, which are ETS, NNETAR, TBATS, and Prophet, were considered and forecast from them were produced. These five models were compared based on RMSE and MAPE values and it was obtained that ARIMA forecasting has the best fitting while ETS has the best forecasting performance for the future compared to other models. ETS has been followed by TBATS, NNETAR, ARIMA forecasting and Prophet, respectively. The biggest challenge in this study was the non-normality and heteroscedasticity problem of residuals. Because of this problem, the analysis might not show the accurate results. Despite the problems encountered, the best model could be obtained. At the end of the study, it is comprehended that how to analyze time series data sets, how to interpret ACF and PACF plots, how to detect and overcome non-stationarity, how to check the diagnostics and overcome the problem, and how to find forecast values of the series.

VI. REFERENCES

Santello, G. (2022, August 26). Hospitality employees - time series dataset. Kaggle. Retrieved January 22, 2023, from https://www.kaggle.com/datasets/gabrielsantello/hospitality-employees-time-series-dataset