Business forecasting and predictive modeling
Seminar Report

Pop David Alexandru

January 6, 2025

Contents

Chapter 1

Introduction

In this piece of work, we will go through the forecasting, analyzing, and interpreting the results of the different forecasting approaches. More concretely, we will start with a small idea of the dataset used and how it is structured, along with the meaning of each column. After we have a rough idea about the dataset, some very notorious forecasting approaches are presented, such as finding trends, seasonality, cycles, running statistical tests with the purpose of finding stationary. In the end, future predictions will be done in order to see how does these models can predict the future.

This paper is structured into four chapters: Introduction, Related Concepts, Practical Work, and Conclusions. In the second chapter, the theory of the newly discovered forecasting method will be presented. In the third chapter, all the steps involved in the final solution will be presented along with explanations and plots. In the final chapter, we will conclude this work with objective and subjective thoughts regarding the results.

Chapter 2

Related Concepts

2.1 Prophet

For the additional forecasting method Prophet had been chosen due to its relevance on the market and predicting relevant results. This model's goal is to be simpler than ARIMA models, which sometimes are hard to parameterize and don't always handle seasonality well. Prophet works in three main steps: First is the preprocessing step in which the model is cleaning and preparing the data. The tool automatically handles missing values, anomalies and outliers. The modeling then takes into account the trend and seasonality to estimate the components associated with each time observation. Each season can be daily, weekly, monthly, annual or customized according to context. After this estimation, Prophet forecasts for the desired future period [1].

Chapter 3

Practical work

The dataset used in this work is called *Electric Production*. It contains time series which spans from January 1985 to January 2018 related to electricity production in the United States. It consists of two columns:

1. DATE: This column represents the date in which the observation was made in the following format: month/day/year (eg. 10/01/2016).

2. IPG2211A2N: Numeric values of the Industrial Production Index for electric power generation, transmission and distribution. The values represent the output of electricity production in a month. The values in this dataset represent monthly changes in electricity production, with higher values indicating greater production levels.

3.1 Exploratory data analysis

The first step is to plot the current data to check if we can see changes over time such as trends, seasonality or cycles or unusual observations. As we can see in the Figure 3.1, the graph has an ascending trend over the period of time. The features which can be seen in the plot shall be included into the forecasting methods. As we can see in the in the original time series, in the Figure 3.2, the red line indicates an ascending trajectory over time. We can observe a clear seasonality pattern, with regular oscillations. The trend line smooths out the seasonal and irregular components, evidentiating the increasing over time. The detrended line represents the residuals after

removing the trend and seasonal components, suggesting that there is no pattern only white noise.

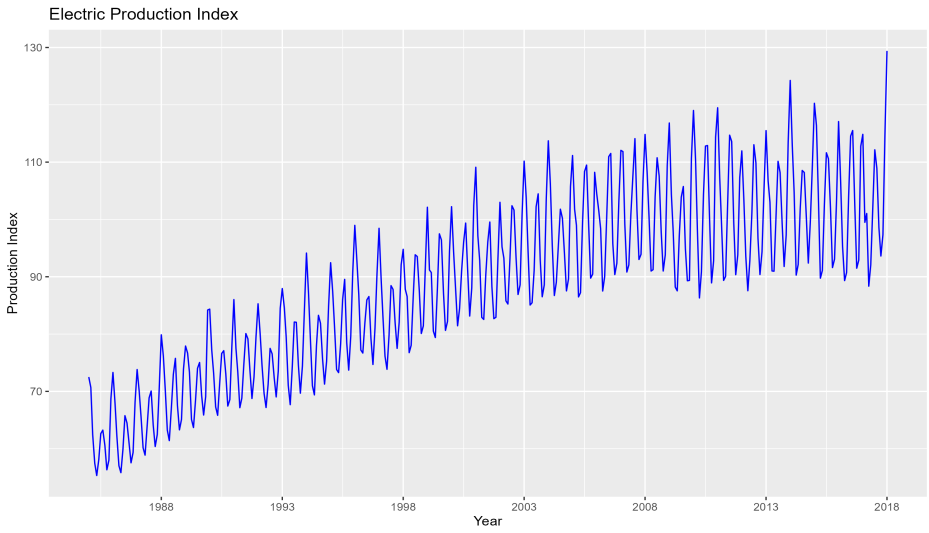

Figure 3.1: Electric Production Index

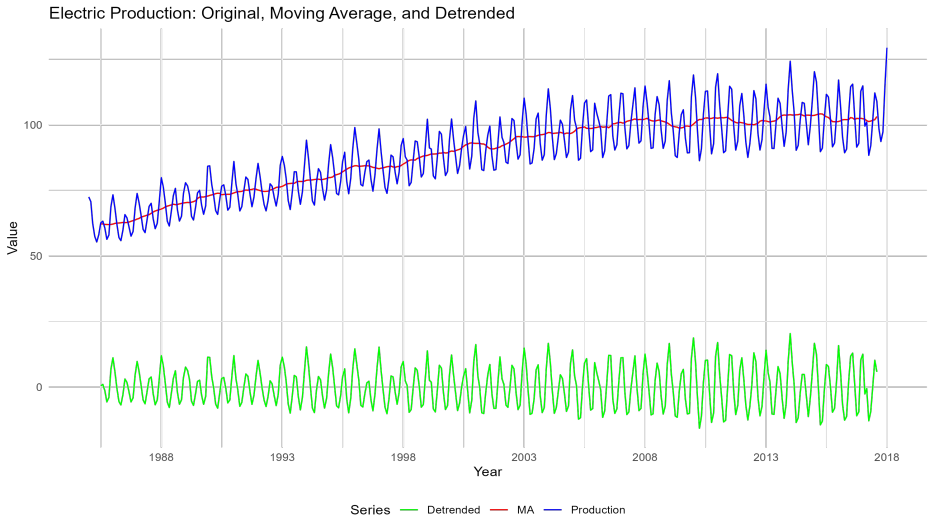

Figure 3.2: Electric Production Index detrended

A time series is non-stationary if any of the statistical properties (mean,

variance or covariance) change over time. This can be observed in the plots, that, trend has an increase in the time series. To check if a time series is stationary or non-stationary a statistical test was used, Augmented Dickey-Fuller, shorten ADF, test. The p-value obtained is greater than the significance level of 0.05 and the value of the ADF test is higher than any of the critical values. Hence, the time series is non-stationary. This can also be observed from the KPSS test.

KPSS Stat	**KPSS p-value**
6.31	0.01

Table 3.1: KPSS test results

Dickey-Fuller	**Lag order**	**p-value**
-1.4408	12	0.8134

Table 3.2: Augmented Dickey-Fuller test results

Because data is not-stationary, we need to transform data to stationary data. To do that, we will apply difference transformations such as logarithmic transformation with box-cox. After that we will detrend the time series, so we will remove the trend from the time series. As we can see a comparision between the box-cox plot and the initial plot in Figure 3.3.

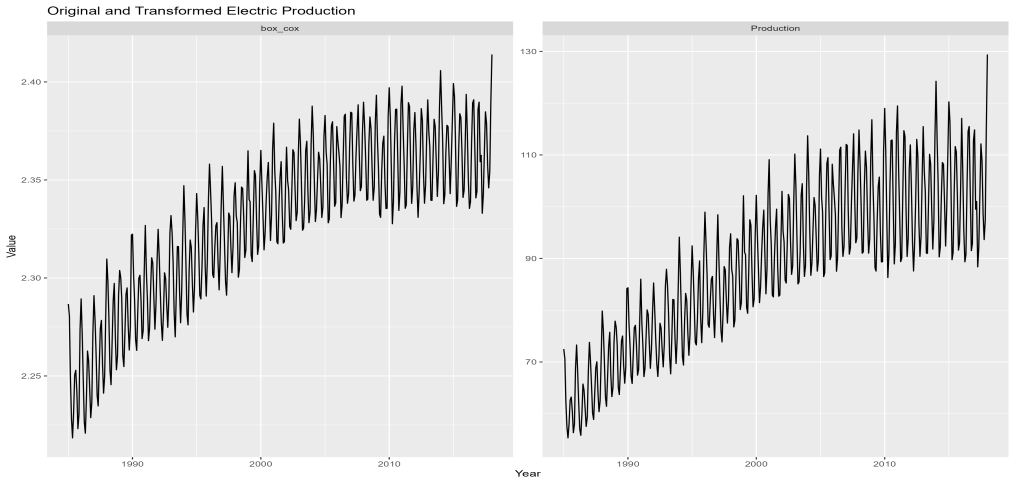

Figure 3.3: Box-Cox transformation

After removing the moving average, we obtain a new plot represented in Figure 3.4.

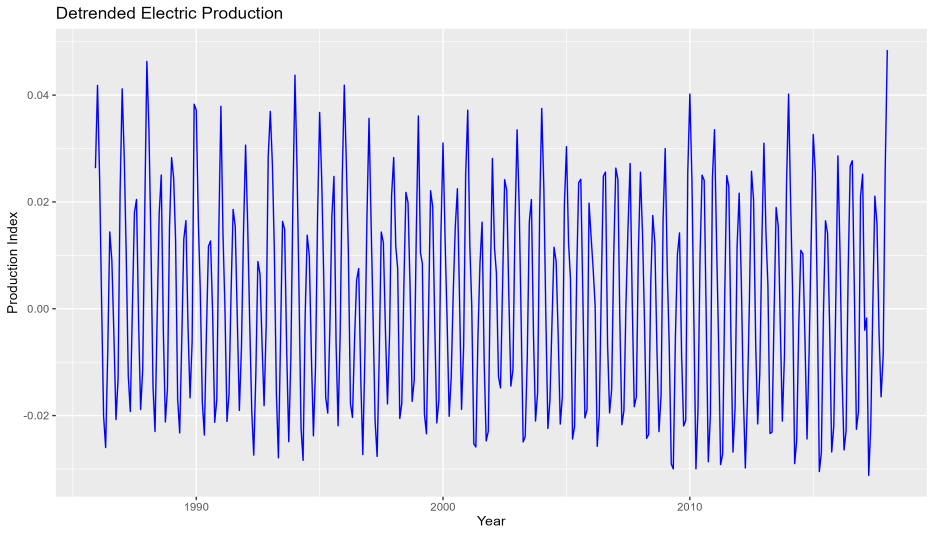

Figure 3.4: Detrended plot

For the new time series we run again the ADF test and we get the following results:

Dickey-Fuller	**Lag order**	**p-value**
-6.0996	12	0.01

Table 3.3: Augmented Dickey-Fuller test results

Because the p-value is less than the significance level of 0.05, we reject null hypothesis and the time series is stationary. We will decompose the time series into three components: trend, seasonal, reminder:

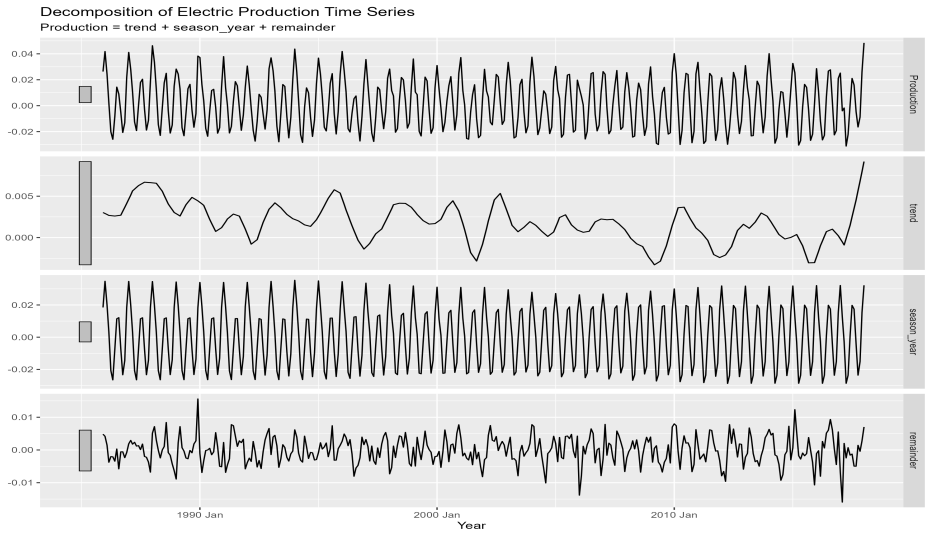

Figure 3.5: Decomposition

We will use autocorrelation to summarize the strength of relationship. We can observe in the Figure 3.6 that might exist a potential seasonal pattern for lag 6 and 12.

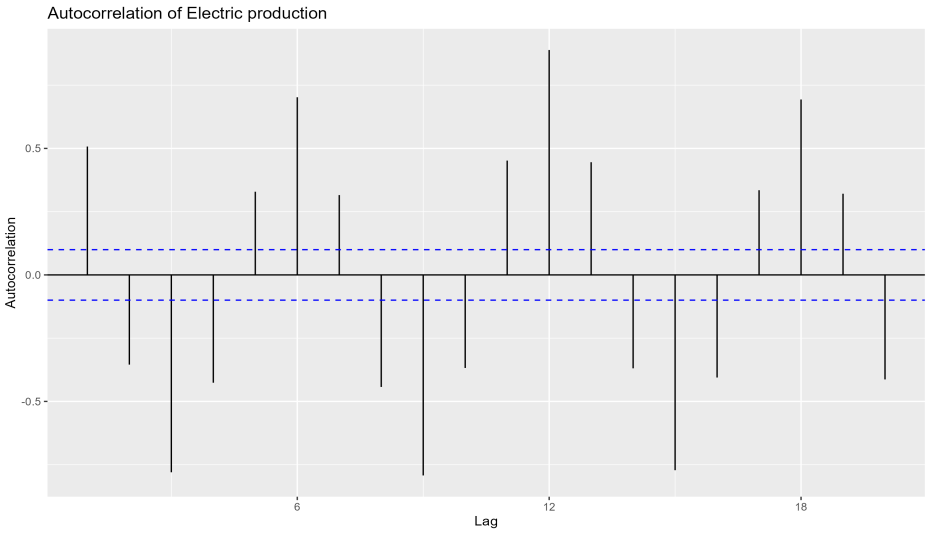

Figure 3.6: Autocorrelation

3.2 Forecasting models

For this work, twelve forecasting methods had been used to compare the results and their accuracies, more concrete:

1. Mean
2. Naive
3. SNaive
4. Drift
5. ETS_ANN
6. ETS_AAN
7. ETS_Auto
8. ARIMA_Auto
9. ARIMA_201
10. NNETAR
11. TSLM
12. Prophet

For the Arima model, the values for p, d, q parameters are: 2, 0, 1. The data has been splitted into two, the training data, 80% percentage, and the testing data. The output of each forecast can be observed in the Figure 3.7.

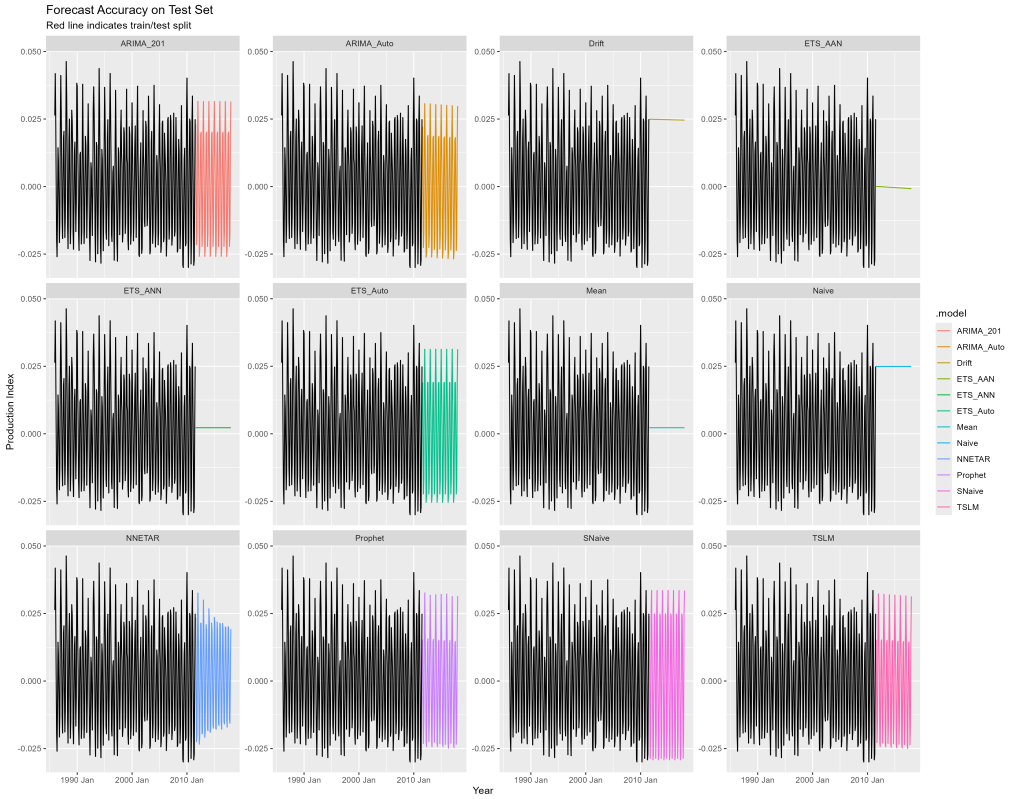

Figure 3.7: Accuracy comparison

To measure the accuracy, root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) methods have been used. For the above methods, we got the following accuracies:

Model	**RMSE**	**MAE**	**MAPE**
ETS_Auto	0.00620	0.00473	64.4
Arima_Auto	0.00621	0.00474	63.5
Arima_201	0.00624	0.00475	68.0
TSLM	0.00678	0.00527	76.5
Prophet	0.00684	0.00530	76.3
SNaive	0.00748	0.00586	86.5
NNETAR	0.00879	0.00690	81.5
ETS_AAN	0.0198	0.0168	99.7
ETS_ANN	0.0198	0.0169	109
Mean	0.0198	0.0169	109
Drift	0.0312	0.0258	392
Naive	0.0313	0.0259	394

Table 3.4: Accuracy table

Chapter 4

Conclusions

As we observe in the accuracies table, to get the best model we are looking for the model with lowest RMSE, MAE and MAPE. The best accuracy was obtained by the ETS_Auto model, and the lowest accuracy was obtained by the Naive model. In the table 4.1 we can observe the most effient models.

Model	**RMSE**	**MAE**	**MAPE**
ETS_Auto	0.00620	0.00473	64.4
ARIMA_Auto	0.00621	0.00474	63.5
ARIMA_201	0.00624	0.00475	68.0
TSLM	0.00678	0.00527	76.5
Prophet	0.00684	0.00530	76.3
SNaive	0.00748	0.00586	86.5
NNETAR	0.00879	0.00690	81.5
ETS_AAN	0.01980	0.01680	99.7
ETS_ANN	0.01980	0.01690	109.0
Mean	0.01980	0.01690	109.0
Drift	0.03120	0.02580	392.0
Naive	0.03130	0.02590	394.0

Table 4.1: Sorted Models by Accuracy

Bibliography

[1] S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

List of Figures