Assignment 2

Rushabh Khara

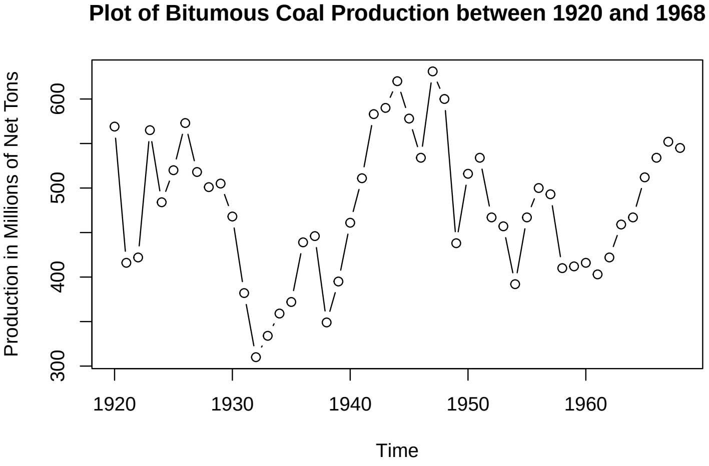
Introduction

This report is intended to provided an in-depth analysis of the Bituminous coal data between 1920 and 1968. The analysis starts with visualisation of the time series data, variance examination, transformation and further going on to decipher trends, seasonal effects, and finally the irregular effects. Intention is to provide a simple yet powerful model that captures the data appropriately. Decisions made during the analysis have been backed up with reasoning and clear explanations.

Analysis

Variance Analysis

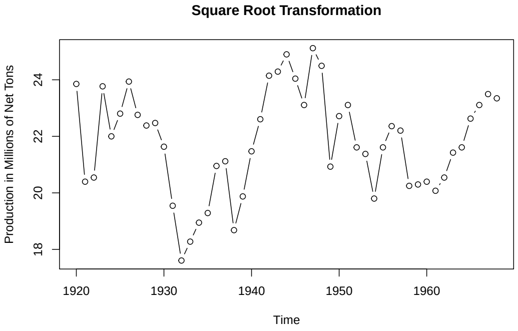
Throughout the time period from 1920 to 1968, the data points show a fairly uniform spread around a central tendency. There isn't a noticeable pattern where the spread of points gets wider or narrower as time progresses. If we were to draw imaginary horizontal lines capturing the bulk of the data for each period, the distance between these lines would remain relatively consistent throughout the timeline.

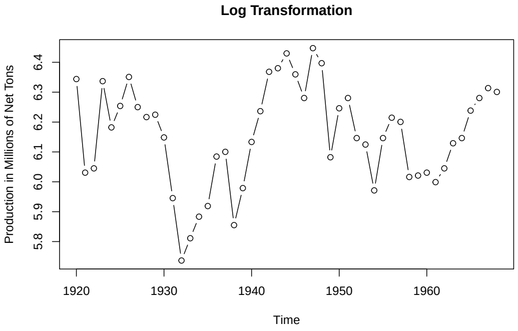


Transformation

The data doesn't seem to have heteroscedasticity problem. However, it would be beneficial to check for transformations to see if the data can have more constant variance.

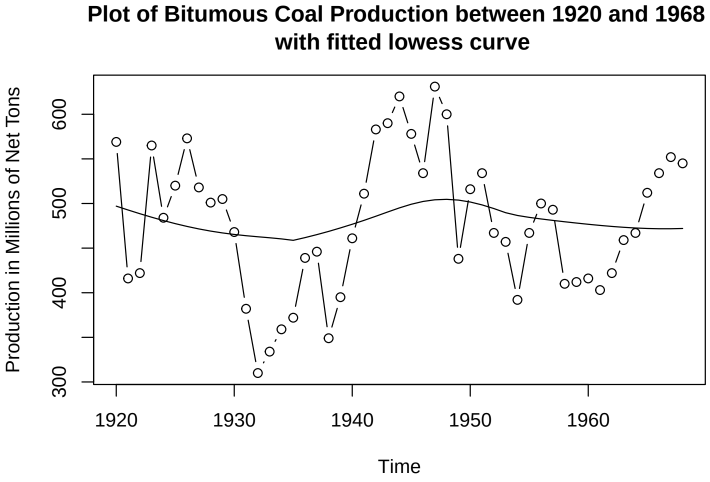
The square root and log transformation plots below exhibit a nearly identical spread of data. Aside from a shift in the y-axis scale, the differences are hardly noticeable. Given this, it's advisable not to apply any transformations. Implementing them would only add complexity to the model without yielding significant benefits. Hence, we proceed with the original scale.





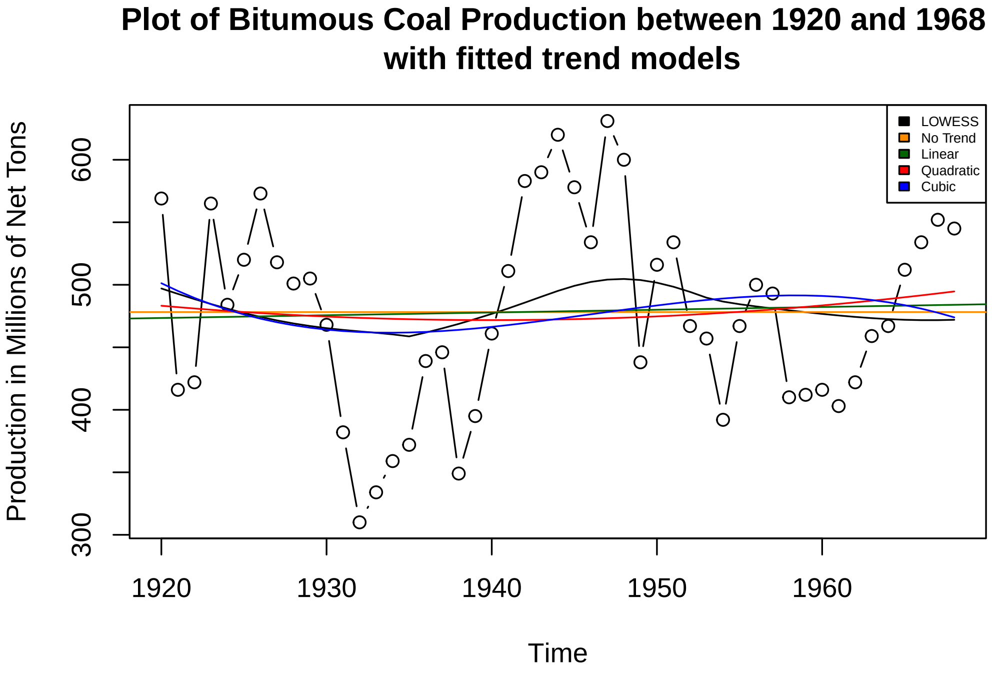
Trend

Fitting the lowess curve, we observe a rather linear behaviour with a slight cyclic trend in the time series. There is no point assuming their is a trend without checking by fitting models as the lowess curve might appear non-linear or linear based on shape and size of the plot.



To determine the most suitable trend for the data, we'll evaluate four models: no trend, linear, quadratic, and cubic. It is important that we center the time before fitting the cubic model to avoid computational errors.

Looking at the models in plot below, it is observed that no trend, linear, and quadratic models have highly similar pattern with slight difference. The quadratic model is limited in its ability to capture trends, as it can curve only once. The cubic model identifies the initial drop but lags significantly in capturing subsequent rises. Due to this lag, the cubic model isn't well-suited for this data. Given the similarities in fit between the no trend and linear models, we prioritize simplicity. Therefore, the no trend model is chosen.

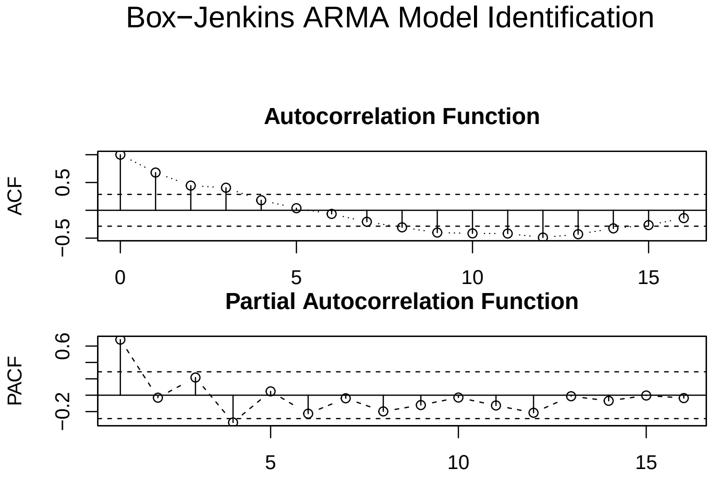


Seasonal Effect

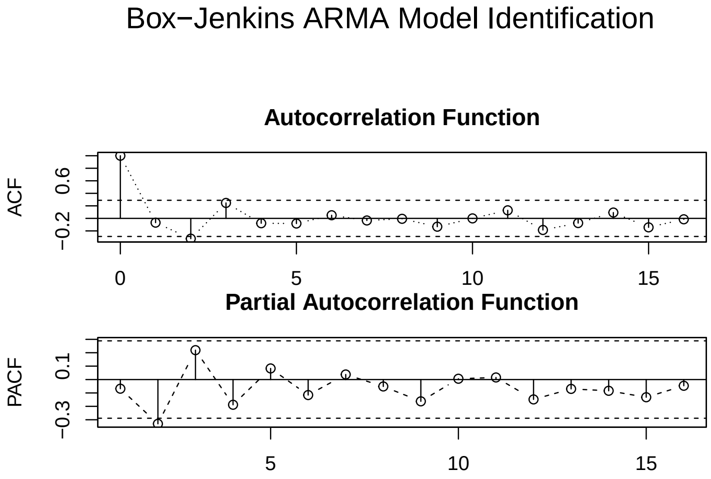
Upon inspecting the plotted time series data, discerning a definitive seasonal effect proves to be complex. The observed wave-like fluctuations hint at potential seasonal trends, yet the limited time span of the data prevents a confident identification of consistent patterns across multiple cycles. Given the dataset's frequency of 1, advanced decomposition methods like STL are inapplicable, limiting our analysis. Furthermore, without comprehensive context or extended data, it's challenging to differentiate genuine seasonal effects from anomalies or external factors. Consequently, while there are indications of potential seasonality, making conclusive determinations from the present data would be premature. Hence, we choose to ignore seasonal effects.

Irregular Effect

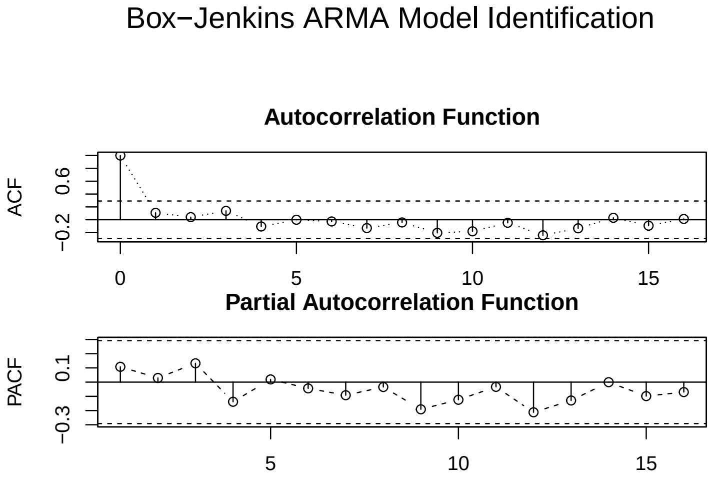
We'll employ the Box-Jenkins ARMA Model Identification to discern the irregular effects and select the optimal autoregressive model. Observing the ACF plot, a slow decay is evident, signaling potential stationarity issues in the data. The PACF reveals significant spikes at lags 1 and 4. However, before fitting an AR model based on the PACF, we'll consider differencing the data. Afterward, we'll re-examine these plots to determine if the stationarity concern has been addressed.



The ACF and PACF plots for the differenced data are presented below. The stationarity issue appears to have been addressed with only one significant spike at lag 0 and spike that barely touches the threshold at lag 2. Notably, the PACF plot shows just one significant spike at lag 2. This represents a notable improvement over the non-differenced data, where an AR(4) model would have been required to be fitted. Now, we can simply fit an simpler AR(2) model.



After implementing the AR(2) model, the PACF plot below shows no significant spikes. The ACF plot displays just one spike at lag 0, which is anticipated. Both ACF and PACF indicate no lingering dependence. It seems the AR(2) model has effectively captured the majority of the dependence.



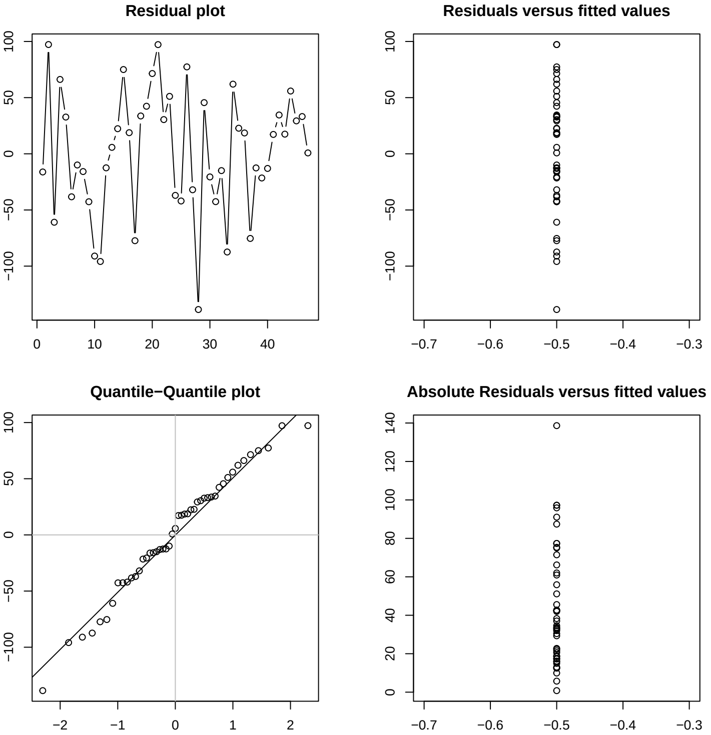
Diagnostic Plots

Residual Plot: A consistent spread is observable, indicating homoscedastic behavior.

Residual vs Fitted Plot: The adoption of a ‘no trend’ model results in all fitted values being the mean of the data, explaining the straight line formation.

Quantile-Quantile (QQ) Plot: The data points closely adhere to the normal line, underscoring the satisfaction of the normality assumption.

Absolute Residual Plot: Due to the lack of variance in the x-axis fitted values, the lowess curve isn’t discernible. This uniformity can be attributed to our ‘no trend’ approach where all values converge to the mean.



Final Model

The final model can be expressed as follows:

- Let $Coal_t - Coal_{t-1}$ represent the *differenced data.*
- Let X_i denote the *Irregular Effect.*

Given these definitions:

$$Coal_t - Coal_{t-1} = \text{mean}(Coal_t - Coal_{t-1}) + X_i$$

$$X_i = -0.136 \times X_{i-1} - 0.296 \times X_{i-2}$$

Conclusion

The Bituminous coal data from 1920 to 1968 underwent a comprehensive analysis, revealing no strong seasonal or trend patterns within this time frame. Initial observations pointed to potential variance fluctuations, but detailed assessments dispelled these concerns. While transformations like the square root and log were considered, they offered no added advantage, prompting the retention of the original data scale. To tackle the non-stationarity observed in the initial irregular component analysis, the data was differenced, leading to the adoption of an AR(2) model as opposed to an AR(4) model that would have been adopted in a non differenced data. This model adeptly captured the series’ nuances, a claim further solidified by diagnostic plots which aligned well with our modeling assumptions. In conclusion, during these years, the coal production remained relatively stable with deviations primarily attributed to irregular factors.