1 Part I: Imputation

1.1 Data

In this scenario, we have a monthly resolution time series of the amount of beer produced in Australia from January 1956 to August 1972 and from March 1975 to March 1992. Additionally, we have Australian steel production, gas consumption, and electricity consumption data from January 1956 to March 1992. Additionally, we have monthly mean high-temperature data from November 1943 to March 1992 and car production data from July 1961 to March 1992 in Australia.

1.2 Methodology

First, we plot the time series to examine its general behavior and characteristics.

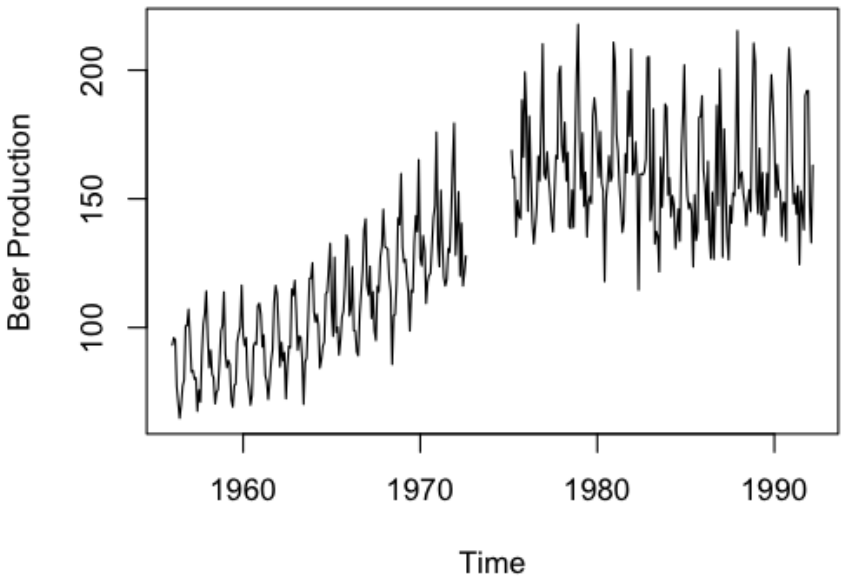

Figure 1: Australian Beer Production Time Series

We observe that from the late 1950s to the early 1960s, the beer data appears stationary. From mid mid-1960s to the early 1970s, there is a positive trend. From mid mid-1970s onward, we observe almost no trend, but some seasonality.

Additionally, we see that the variance of the time series appears to be increasing. The variance in the 1960s is less than the variance in the 1980s.

Kalman smoothing is a technique that can be used to impute missing values and smooth time series data. It takes into account both past and future observations to provide the best linear unbiased estimates. This characteristic makes it suitable to impute missing data in the middle of the time series. To impute the missing data, we have two approaches.

The first approach is to fit an ARIMAX model to the time series before the missing values and forecast the missing values. As data are observed, we apply Kalman smoothing to update the state estimates by taking into account future values. ARIMAX takes advantage of the other time series data in Australia. These other time series may have correlations with the beer production time series since they all represent the activity of Australians. For instance, higher temperatures may lead to higher consumption of beer, thus increasing beer production. Thus, starting with ARIMAX can incorporate any impact that other time series have on the beer production time series. The result of this approach is shown below.

Note that we use the latest start date among all series to fit the ARIMAX model. The time series for cars has the latest start date, which is July 1961.

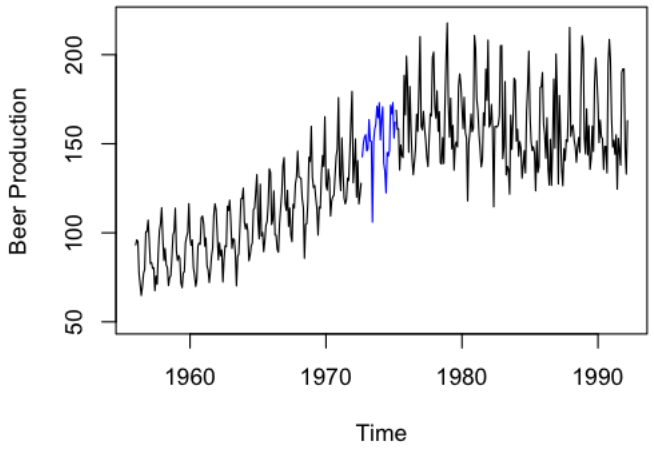

Figure 2: Australian Beer Production Time Series Imputed by ARIMAX and Kalman Smoothing

The second approach is to use Kalman smoothing in combination with StructTS or SARIMA. Since we observe trends and seasonality in the observed log-transformed time series, a simple Kalman smoothing that assumes linearity may not be suitable. Hence, we consider integrating Kalman smoothing with StructTS or SARIMA.

StructTS and SARIMA models provide a structured approach to decompose time series into trend, seasonality, and residuals, which can then be modeled in a state-space framework suitable for Kalman smoothing. Furthermore, since different periods of the beer data exhibit different trends, StructTS is more suitable than SARIMA. StructTS can dynamically adjust to changes in the trend and seasonality, which is advantageous for long-time series with structural breaks or shifts. SARIMA is more suitable when the trends are consistent. By integrating these StructTS, the smoothed estimates from Kalman can benefit from the component-based modeling of StructTS, leading to more accurate and robust forecasting and missing data imputation.

However, there is increasing variance in the time series, and StructTS does not handle non-constant variance well. Thus, we consider applying log transformation to the time series in the second approach. As seen in the figure below, log transformation has removed the non-constant variance.

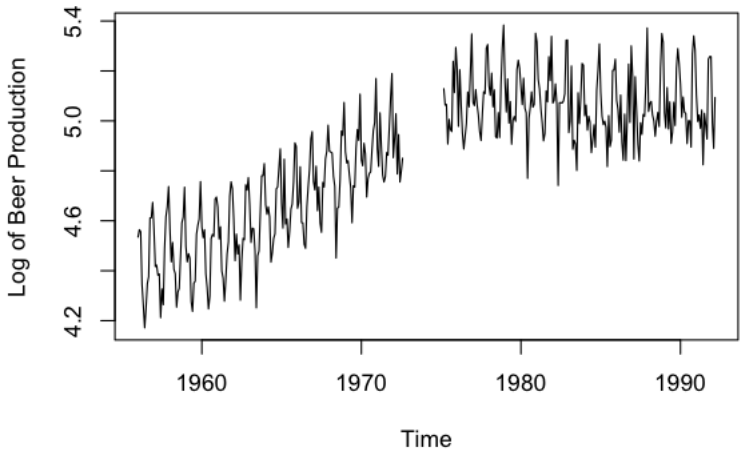

Figure 3: Log of Australian Beer Production Time Series

The imputed time series using the second approach is shown in Figure 22.

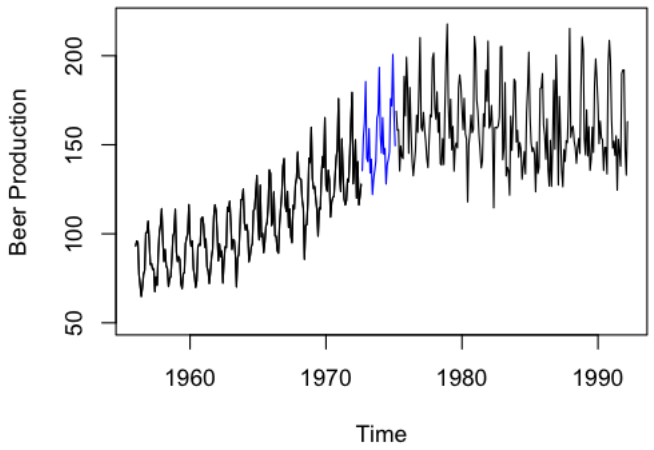

Figure 4: Australian Beer Production Time Series Imputed by StructTS and Kalman Smoothing

By comparing Figure 20 and Figure 22, we observe that the ARIMAX method imputes slightly lower values than the StructTS method. The ARIMAX method tends to follow the pattern shown after the 1980s, whereas the StructTS method follows the pattern shown in the mid-1960s to 1970s. However, we observe that the maximum value imputed by the ARIMAX method is lower than most of the peaks observed afterward. Moreover, in the observed time series, we see a short cycle approximately every year. The imputed portion derived by ARIMAX does not display these short cycles as clearly. Thus, the StructTS method aligns with the observed pattern better.

1.3 Diagnostics

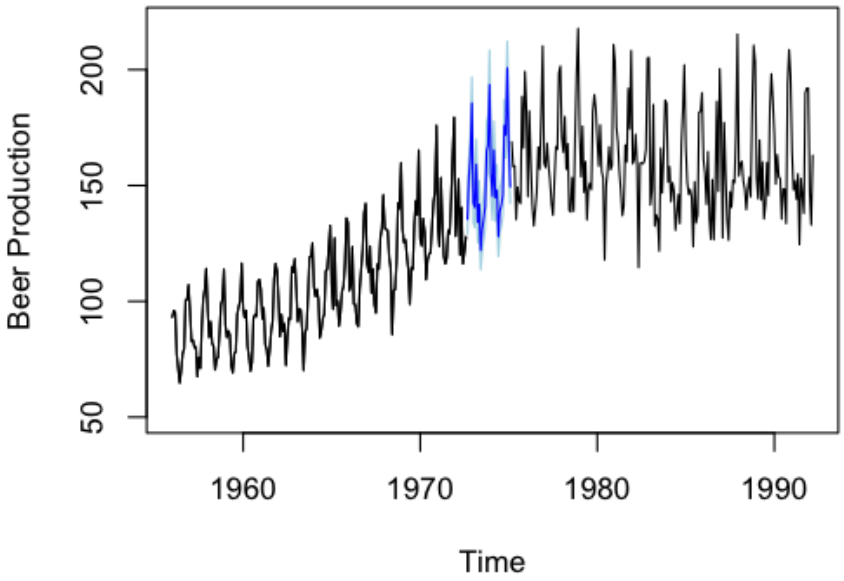

Figure 5: Australian Beer Production Time Series Imputed with Prediction Interval

The imputed result is shown in dark blue and the 95% prediction interval is shown in light blue. The imputed portion highlighted in blue matches the general pattern from 1970 to 1980. We see that the trend and the variance are slowly increasing. The seasonal pattern follows the same period as the other portions.

2 Part II: Forecasting

2.1 Data

In this scenario, we use the imputed beer data from Scenario 3 in combination with other data regarding Australia. We have Australian steel production, gas consumption, and electricity consumption data from January 1956 to March 1992. Additionally, we have monthly mean high-temperature data from November 1943 to March 1992 and car production data from July 1961 to March 1992 in Australia.

2.2 Model Selection

Forecasting beer production requires considering various economic and environmental factors that can significantly influence production volumes. Factors such as steel production, gas consumption, electricity consumption, temperatures, and car production can all influence beer production. For example, steel is a critical component in the manufacture of brewing equipment and storage containers like kegs and cans. Likewise, The brewing process uses natural gas extensively, especially for heating brew kettles. Variations in gas consumption or its cost directly affect production costs and operational efficiency, which can, in turn, influence beer production volumes. To forecast beer production in the next 24 months, not only do we have to consider beer production in the past, but we also have to consider these other influencing factor's data in the past.

Both ARIMAX and the VAR (Vector AutoRegression) model take into account the influence of other time series on beer production. However, since we do not have future data for the other time series, we cannot use ARIMAX to forecast beer production's future value. The VAR model does not require any future data and is suited for this analysis due to its ability to handle multiple interdependent time series. This model captures the linear interdependencies among the variables, which allows for an exploration of how beer production is influenced by the others. It focuses on the autoregressive (AR) terms, which describe how past values of all variables in the system influence current values.

To find the best order for the VAR(p) model, we iteratively check all $p \in [0, 20]$ and identify the best order that gives the lowest AIC score. The resulting model is VAR(13), resulting in an AIC of 53.59.

2.3 Model Diagnostics

To evaluate the fit of the model, we can check the autocorrelation of the residuals using the ACF plot.

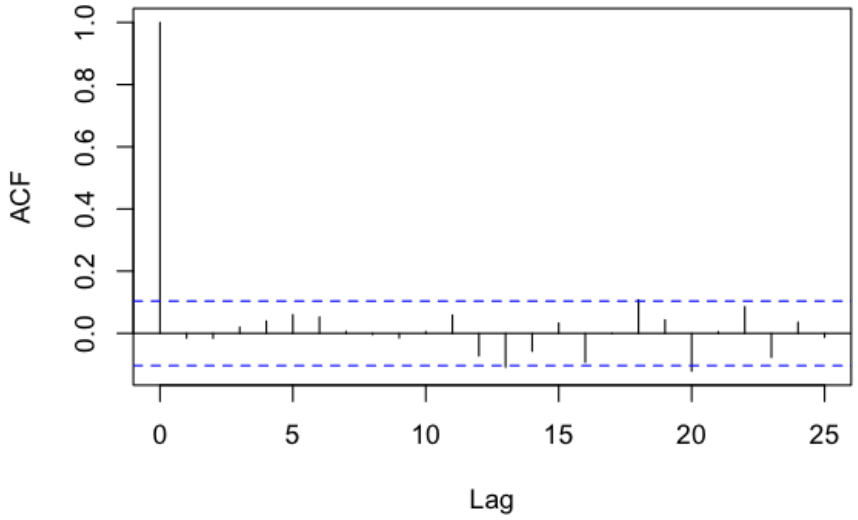

Figure 6: 2 Years Forecast of Australian Beer Production

As seen in the ACF plot, there is no significant autocorrelation at any lag except for lag=0. This indicates that the residuals behave like white noise and the trends in the data are effectively captured by the VAR(13) model.

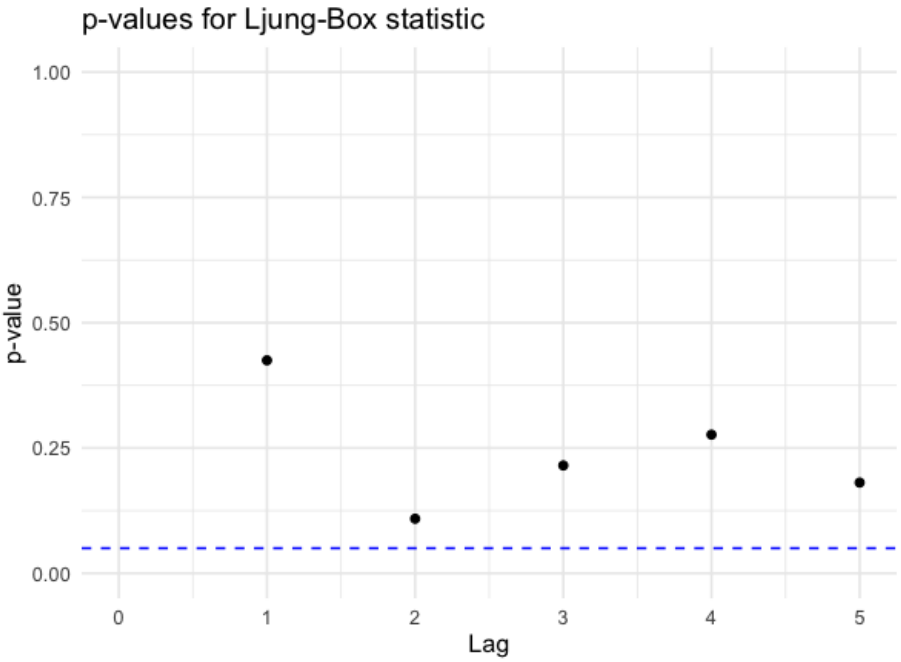

Figure 7: 2 Years Forecast of Australian Beer Production

The Ljung-Box test for the first few lags has p-values greater than 0.05. This confirms that there are no significant autocorrelations at these lags.

2.4 Forecast

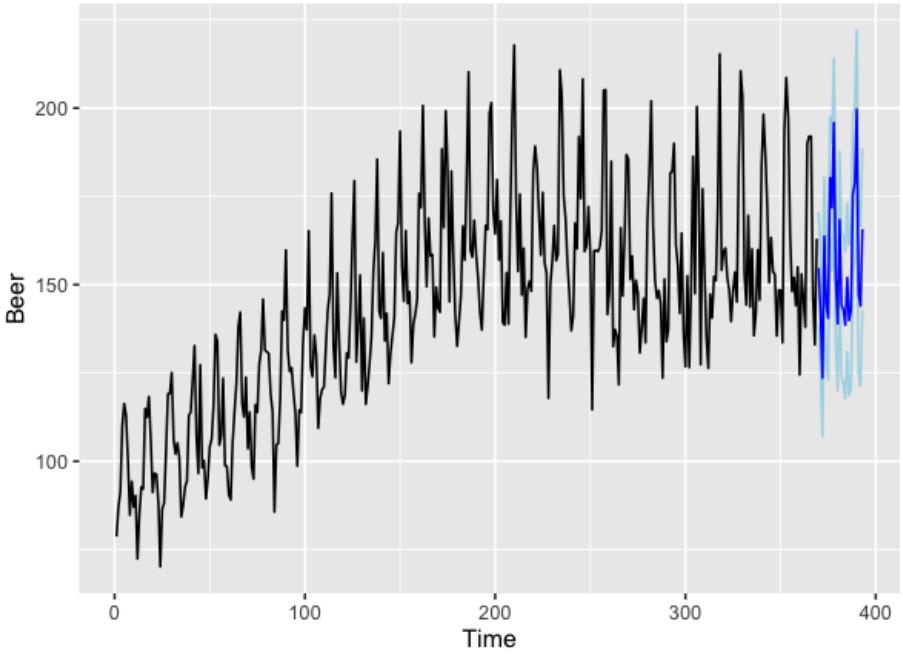

Figure 8: 2 Years Forecast of Australian Beer Production

The dark blue line is the forecast of Australian beer production 2 years ahead. The light blue curves encloses the 95% prediction interval of the forecast.