Project Report - SF2943 Time Series Analysis

Henrik Söderhielm 940524-3115 Alva Liu 950831 Amit Alam 900328-1335
Ghassan Alrheis 950817-5610 Mark Guban 920108

April 27, 2018

Part 1: Non-Financial Time Series

This section consisted of the analysis of a time series of a non-financial nature. The data in this analysis included the US national candy production from January 1972 to October 2017 recorded month to month.

Part 1 a)

The candy production data is expressed as a percentage of the candy production from January 2012. The data is presented in Figure 1 shown below.

Raw Data

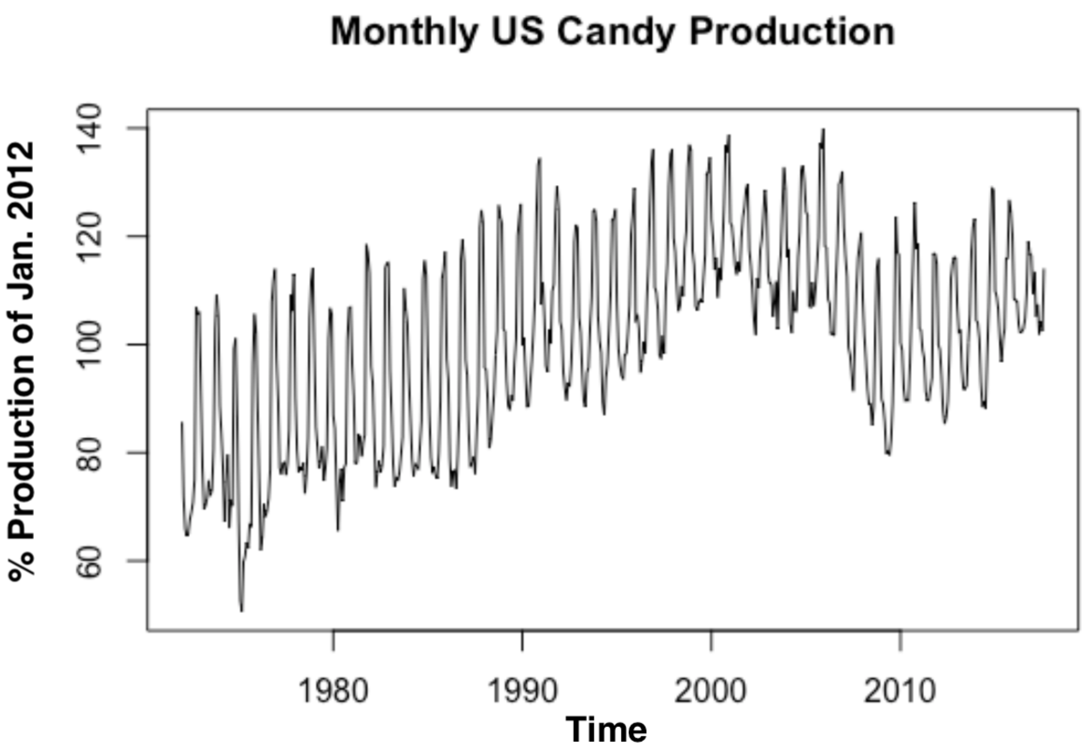

Figure 1: US Monthly Candy Production as a percentage of production recorded in January 2012.

Initial Analysis

The beginning of the analysis is to check whether the data is stationary or not. The property of stationarity of a time-series implies that the mean and variance of the data does not change over the time. A simple check of stationarity is the autocorrelation function (ACF). If the values of the ACF exceed the bounds $\pm 1.96/\sqrt{n}$ then the time-series is not stationarity, and modifications need to be made, such as the removal of trend and seasonal components. The ACF is shown below in Figure 2.

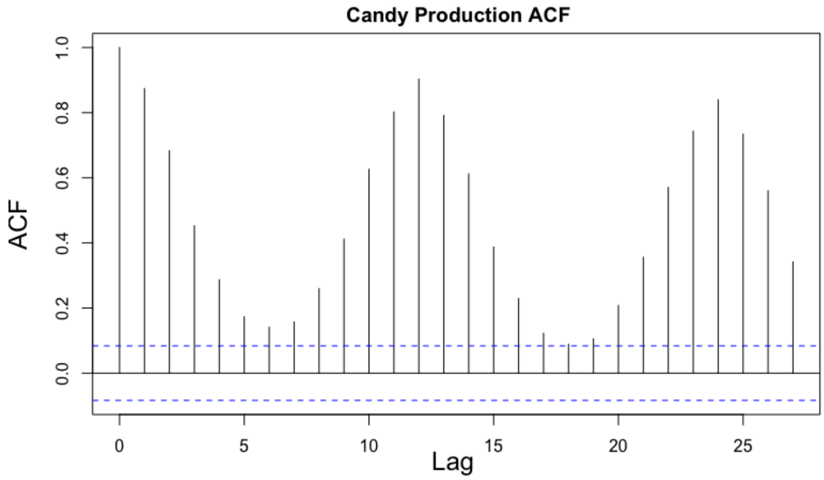

Figure 2: Initial Autocorrelation Function (ACF) for the raw data set.

From Figure 2, it is clear that the time-series is not stationary due to the ACF and its oscillating behaviour, which indicates a changing variance over time. None of the values are within the $\pm 1.96/\sqrt{n}$ bounds shown illustrated in blue in Figure 2. The data requires needs to be processed in order to achieve stationarity for subsequent forecasting.

Trend & Seasonality

In the process of producing stationarity, the main properties of the data have to be identified, i.e. visual patterns. Does the variable increase with time? Does the variable exhibit consistent changes in regular intervals?

Observing Figure 1, the movement in the data has a clear upwards sloping trend. The simplest form is the selection of a trend-line as shown below in red in Figure 3. This trend has to be taken into account when fitting the relevant model.

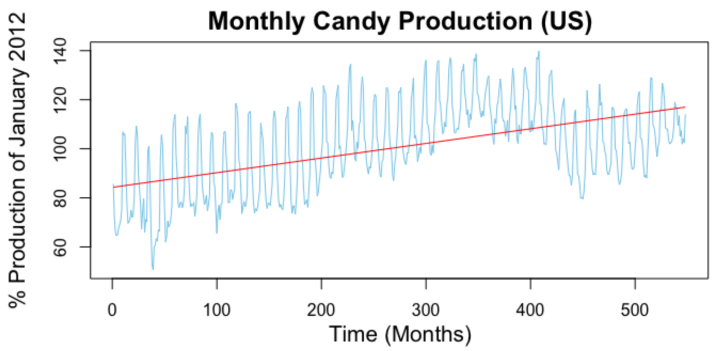

Figure 3: The identified linear trend in the time series.

Another issue to address is seasonality. As we can see in Figure 1, there is a regular pattern of oscillation. This is a clear sign of seasonality. By looking at a more concentrated set of data, it can be seen from Figure 4 that there is a distinct pattern which provides evidence that the data is seasonal.

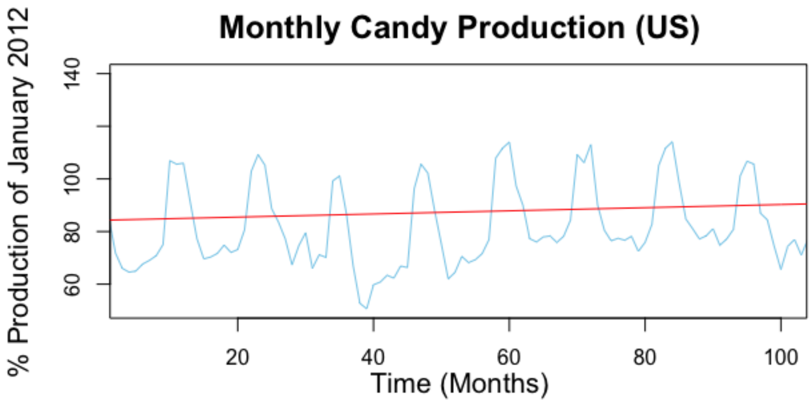

Figure 4: Close-up seasonality for the data.

From Figure 3 and 4 we have concluded that the data exhibits a trend and seasonality properties. These will aid as guidelines for selecting the relevant model to fit to the Candy Production data in Figure 1.

Data Transformation

Does the data require transformation? For example, a logarithmic transformation of the data may be used when the seasonal fluctuations change as time progresses. In Figure 1, such behaviour is not indicated. However, with a consistent upward trend, it may be assumed that the data does have a multiplicative nature. Also, the seasonal fluctuations throughout the data are comparable in size, and as stated before, do not increase/decrease with time. Hence, there is evidence for additivity. The best guess is to say that the data is in between an additive and multiplicative nature. In this analysis, we will proceed with a logarithmic transformation of the data.

Model Selection and Fitting

The ACF of the log-transformed data is typical for a series that is not stationary with seasonality. Thus, it is reasonable to choose a general seasonal ARIMA (**A**utoregressive **I**ntegrated **M**oving **A**verage) model to fit the data. A seasonal ARIMA model has the form ARIMA(p,d,q)x(P,D,Q)[s]. At least one order of differencing is required to adjust for seasonality. In this analysis, the ARIMA process was evaluated in a step by step process, adding relevant p, d, q, P, D, Q and s terms where deemed necessary.

Step 1

The first step was to process seasonality. This step included one differencing step for seasonality. The resulting ACF/PACF are shown in Figure 5.

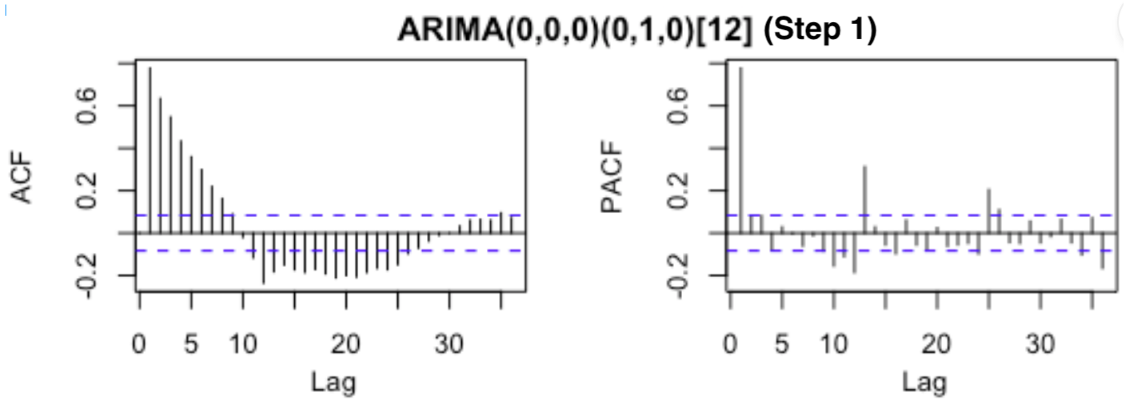

Figure 5: ACF/PACF after non-seasonal differencing (Step 1).

As we can see, the data is not stationary and further steps need to be taken.

Intermediate Steps

- **Step 2; ARIMA(0, 1, 0)(0, 1, 0)**: In step 1, we have, due to strong and consistent seasonality in the data, applied seasonal differencing in order to remove it. The next step is to apply non-seasonal differencing to remove the trend from the data.
- **Step 3; ARIMA(1, 0, 0)(0, 1, 0):** Due to lower AIC (Akaike Information Criterion) and RMSE (Root Mean Square Error) values, this model is a better fit than the one considered in step 2, with the addition of an AR term and the removal of non-seasonal differencing.
- **Step 4; ARIMA(1, 0, 0)(0, 1, 1):** In order to remove a negative spike in the ACF/PACF correlograms at lag 12, an SMA term is added to the model, further improving the AIC and RMSE statistics.

Final manual step: Step 5

To compensate for the slightly negative value at lag 1 in the ACF, we can add an MA term. In terms of AIC and RMSE, this step slightly improves the model and it appears to be a reasonable fit to the data. It is possible to further fine-tune the model by adding more parameters to it. Figure 6 below shows the ACF/PACF correlograms in this step.

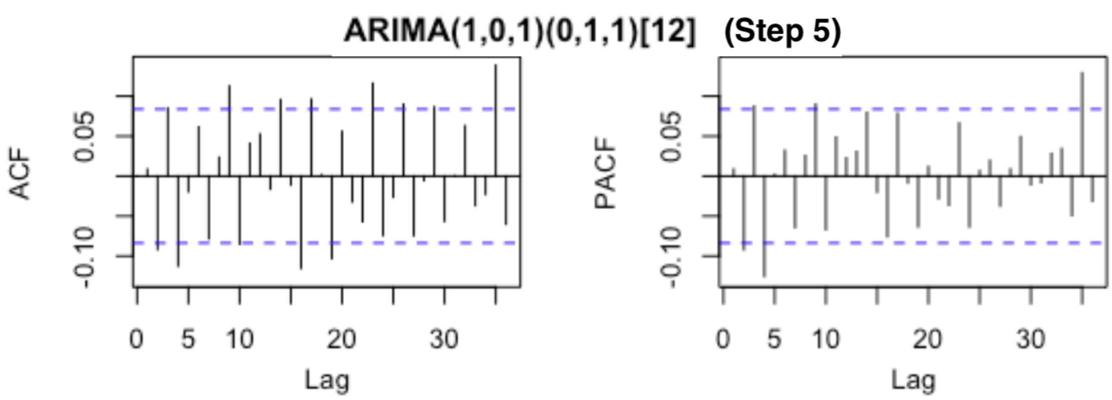

Figure 6: Step 5.

As we can see, the ACF/PACF correlograms show that most of the values are within the limit $\pm 1.96/\sqrt{n}$ which indicates a close to stationary set of data. In order to verify our assumptions in this step-wise process, we use *auto.arima()* in R.

Automatic ARIMA Fit.

The ACF/PACF correlograms for the *auto.arima()* model are shown below in Figure 7. As we can see, this model exhibits a stronger stationarity than before, and it verifies that the assumptions which we have made in the manual analysis were correct.

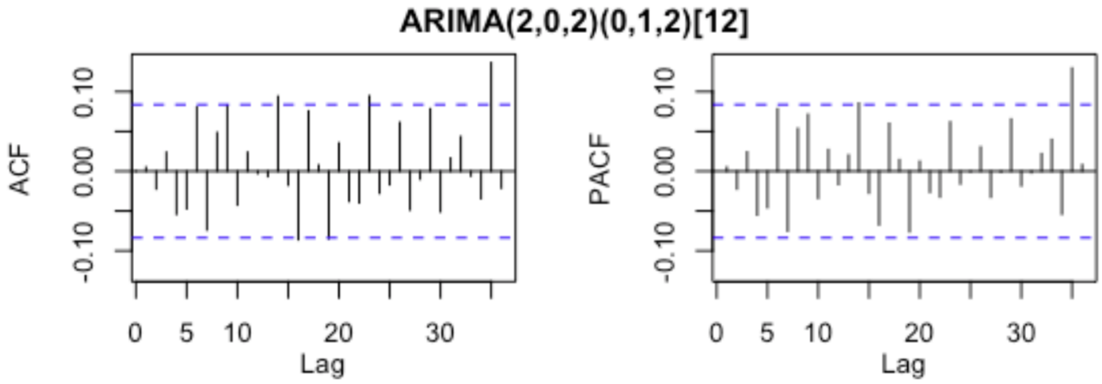

Figure 7: Autmatic ARIMA(2, 0, 2)(0, 1, 2) fit to the data.

Results

Our final model that fits the data best is the ARIMA(1,0,1)(0,1,1)[12] model. We use the *auto.arima()* command in order to check our assumptions, and the function suggests the model: ARIMA(2,0,2)(0,1,2)[12]. Thus, our model is very close to the "optimal" model found by the *auto.arima()* function. If we compare the AIC and RMSE of our model with the auto.arima model, we can see that there difference between them is not very significant. Table summarizes the AIC and RMSE values for each and every step as well as those for the automatic model.

ARIMA(p, d, q)·(P, Q, D)	*AIC*	*RMSE*
(0, 0, 0)·(0, 1, 0)	-1249.887	0.0744
(0, 1, 0)·(0, 1, 0)	-1687.871	0.0493
(1, 0, 0)·(0, 1, 0)	-1750.416	0.0465
(1, 0, 0)·(0, 1, 1)	-1903.030	0.0400
(1, 0, 1)·(0, 1, 1)	-1910.508	0.0397
Automatic ARIMA	*AIC*	*RMSE*
(2,0,2)·(0,1,2)	-1920.020	0.0399

Table 1: Obtained ARIMA models.

Forecasting

Since both of the models are comparable in accuracy, it is useful to also compare the forecast provided by each model. The forecast for 12 steps are shown below in Figures 8 and 9.

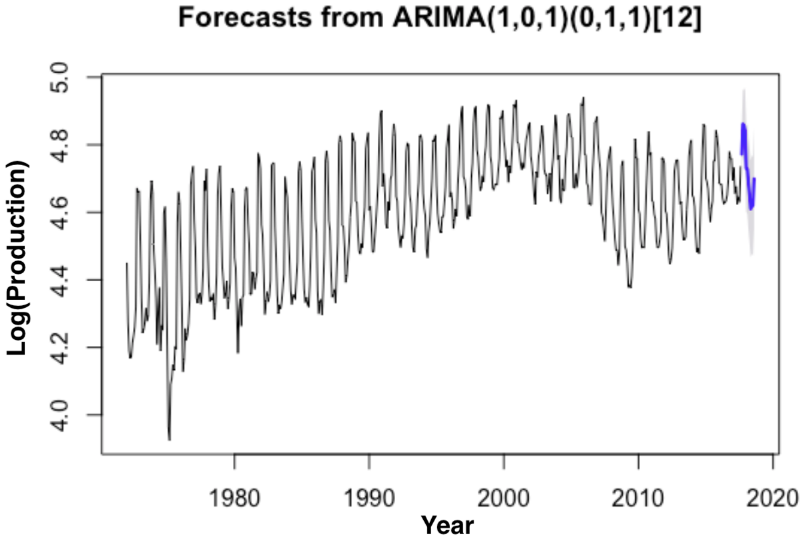

Figure 8: 12 Step Forecast for the manually derived ARIMA model.

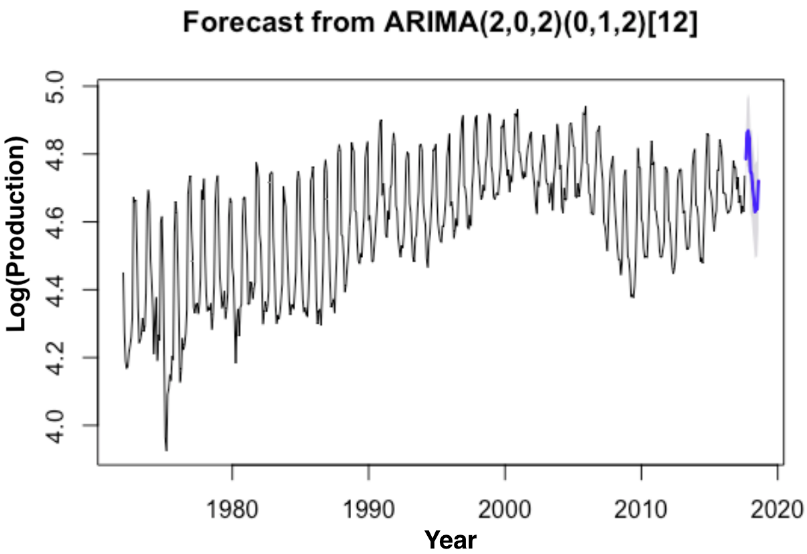

Figure 9: 12 Step Forecast for the automatically determined ARIMA model.

Conclusion

Through the analyses performed in the previous sections, the stationarity of the data has become debatable. We have come close to achieving good values in the ACF and PACF correlograms. From the correlograms in Figure 6 and 7, the values are within the cut-off limits $\pm 1.96/\sqrt{n}$ which suggests that the data has become stationary and we have obtained good fits using a seasonal ARIMA model.