Do U.S. Imports of Chinese Goods Depend on the Average Joe?

Shiqi Liang, Ingrid Wijaya, Jessica Wong

February 15, 2022

1 Introduction

We would like to investigate whether U.S. imports of Chinese goods depend on the consumption ability of the average American. Here, we define consumption ability as the ability and inclination to consume products, with a focus on products imported from China. We approximated this by including two independent variables: average hourly earnings of production and nonsupervisory employees in the private sector, and industrial production of non-durable nonenergy consumer goods.

For the first independent variable, the focus on production and nonsupervisory employees may better represent the average American because the hourly wage is no longer skewed by a select number of high-earning professionals. At the same time, most Americans today work in the private sector [Lab21]. The second independent variable serves as an approximate of domestic production and was chosen because if more production is happening at home, there may be less of an incentive to import from elsewhere. Both independent variables relate to this consumption ability, since individuals earning lower amounts or having access to domestic goods are likely to shop and consume less imported goods from China.

This is an important area to study because China is still the U.S.'s biggest trade partner, and a large share of consumer goods sold in the U.S. are still manufactured in China. Additionally, the consumption of the average American is a powerful force that can shape many financial dynamics. We hypothesize that higher average earnings and lower domestic production are associated with more Chinese goods being imported to the U.S. These data sets can be found through Federal Reserve Economic Data (FRED). We have also included links to each individual data set in the reference section.

2 The Data

The three variables we have investigated are:

1. **U.S. Imports of Goods by Customs Basis from China**: Monthly, non-seasonally-adjusted data on U.S. imports of Chinese goods by customs basis, in millions of dollars [BE22].
2. **Average Hourly Earnings of Production and Nonsupervisory Employees, Total Private**: Monthly, non-seasonally-adjusted data on average hourly earnings of production and nonsupervisory employees in the U.S. private sector, in dollars per hour [Lab22].
3. **Industrial Production: Non-Durable Nonenergy Consumer Goods**: Monthly, non-seasonally-adjusted industrial production index for non-durable nonenergy consumer goods in the U.S., as an index based upon the 2017 value being 100 [Gov22]. (We have confirmed that in this case, this index is acceptable to use for the project.)

Below are further discussions of each variable based on analyzing their time plots and seasonal box plots.

Variable Short Name	Description	Training Period	Testing Period	Frequency	Source
IMPCH	United States imports of goods by customs basis from China, not seasonally adjusted	Jan 1992 - Dec 2005	Jan 2006 - Dec 2006	Monthly	FRED
CEU0500000008	Average hourly earnings of production and nonsupervisory employees in the private sector of the United States, not seasonally adjusted	Jan 1992 - Dec 2005	Jan 2006 - Dec 2006	Monthly	FRED
IPB51210N	United States industrial production of non-durable nonenergy consumer goods, not seasonally adjusted	Jan 1992 - Dec 2005	Jan 2006 - Dec 2006	Monthly	FRED

Table 1: Table of variable descriptions

U.S. Imports of Goods by Customs Basis from China

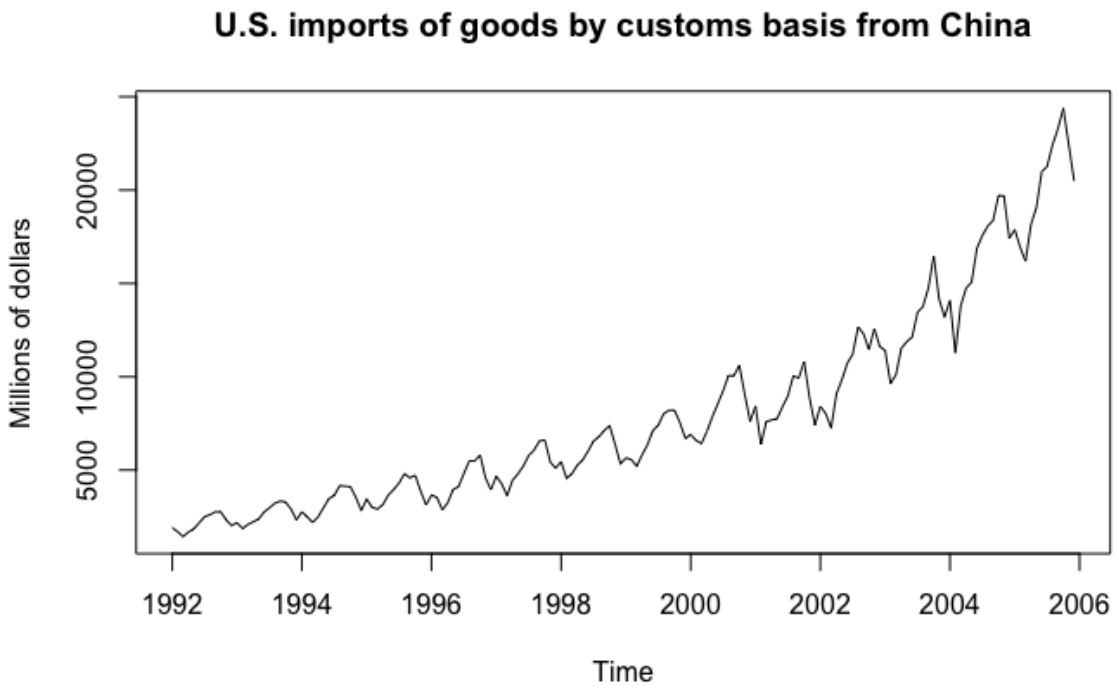

Figure 1: Time plot of U.S. imports of goods by customs basis from China

Referring to Figure 1, the time plot of U.S. imports of goods by customs basis from China displays a strong upward trend, which can be expected due to inflation over time and the U.S.'s growing trade deficit with China. There also seems to be a visibly increasing seasonal component, which we looked

into further using the seasonal box plot.

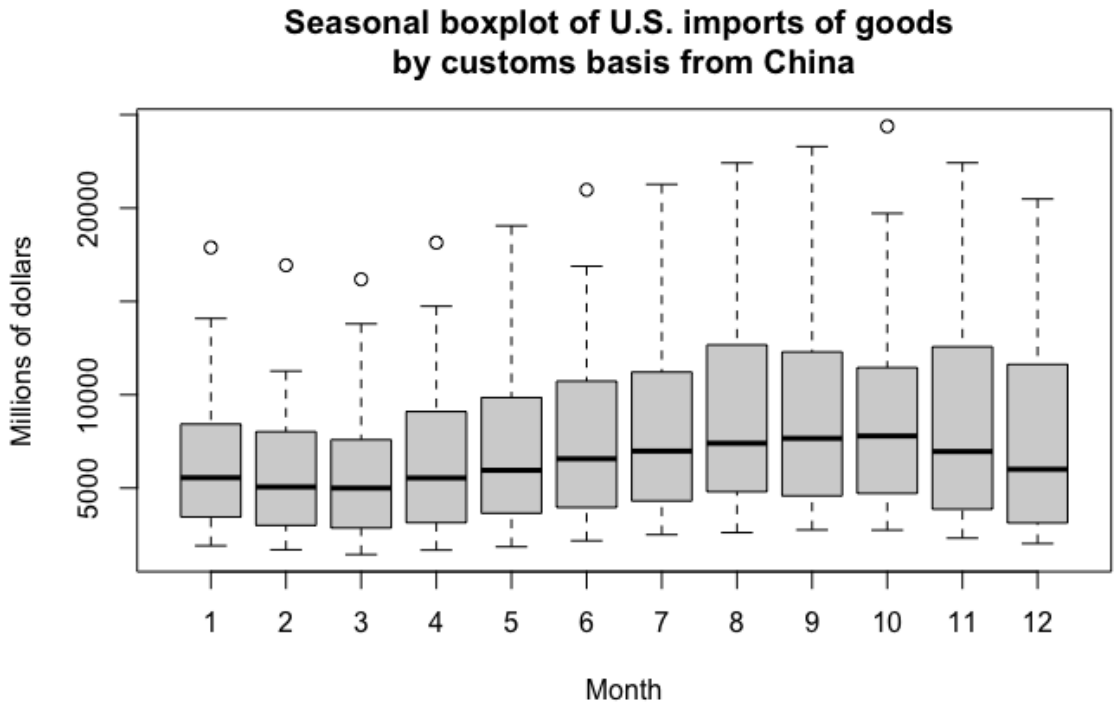

Figure 2: Seasonal box plot of U.S. imports of goods by customs basis from China

Referring to Figure 2, we can see a seasonal component in U.S. imports of goods by customs basis from China, although all of the box plots for the different months overlap with one another, suggesting that the seasonal component is not a dramatic one. It makes sense that the demand for goods would not show drastic shifts throughout the year but is perhaps subject to consumer purchasing patterns, such as suppliers ordering more goods in the fall to prepare for customers' holiday spending. We can observe that October and September generally see the highest dollar amounts of U.S. imports of goods from China with relatively high variability (among the highest variability if the outlier for October is considered), while March and February generally see the lowest dollar amounts of U.S. imports of goods from China with some of the lowest variability.

Average Hourly Earnings of U.S. Private Sector Production and Nonsupervisory Employees

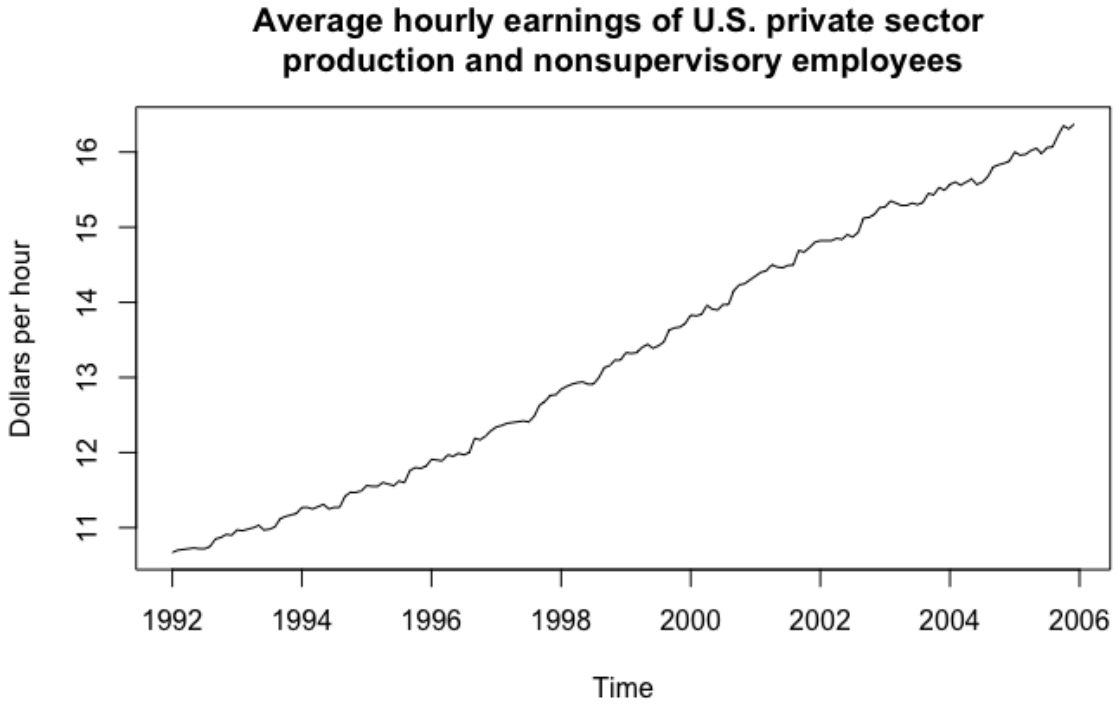

Figure 3: Time plot of average hourly earnings of U.S. private sector production and nonsupervisory employees

Referring to Figure 3, we can see a strong upward trend in the time plot of average hourly earnings of U.S. private sector production and nonsupervisory employees, which could be expected due to inflation and increases in the minimum wage over time. No obvious seasonal component is present, but we used the seasonal box plot to double check.

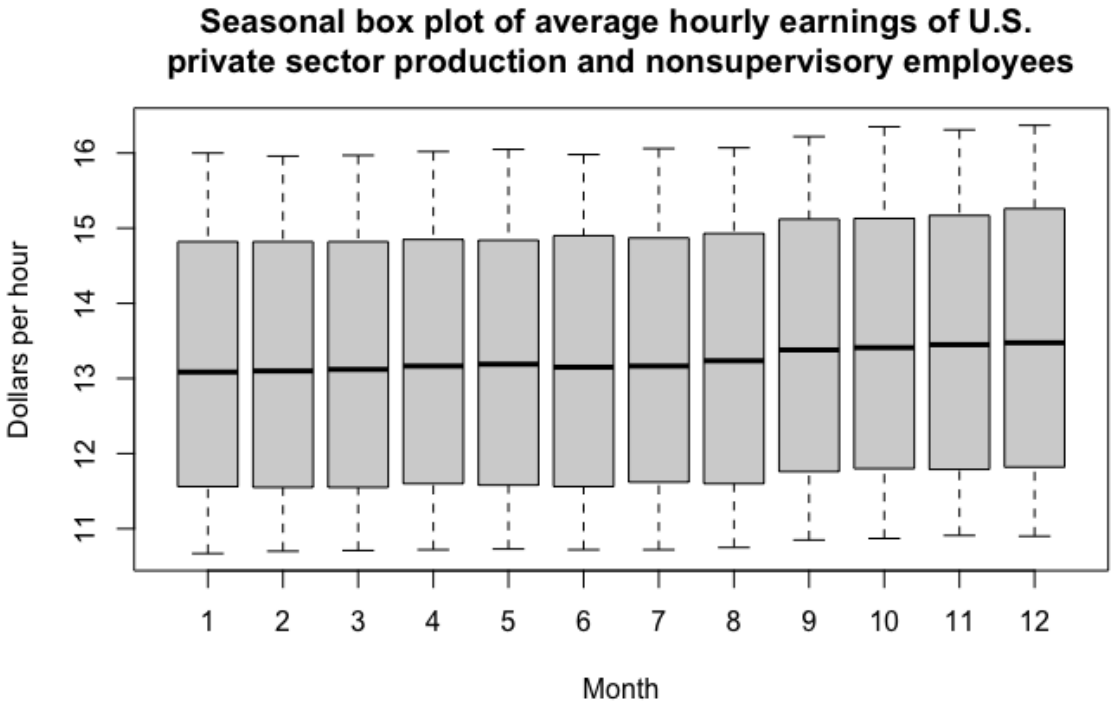

Figure 4: Seasonal box plot of average hourly earnings of U.S. private sector production and nonsupervisory employees

Referring to Figure 4, we can also see that there is not a noticeable seasonal component present for the average hourly earnings of U.S. private sector production and nonsupervisory employees. Based on the seasonal box plot, the boxes for each month look similar to one another in terms of both median and variability, slightly increasing for the last four months of the year. This would be expected since hourly wages are expected to remain relatively stable throughout the year, and the demand for labor leading up to the holiday season with its increased consumer activity may be associated with higher hourly earnings.

U.S. Industrial Production of Non-Durable Consumer Goods

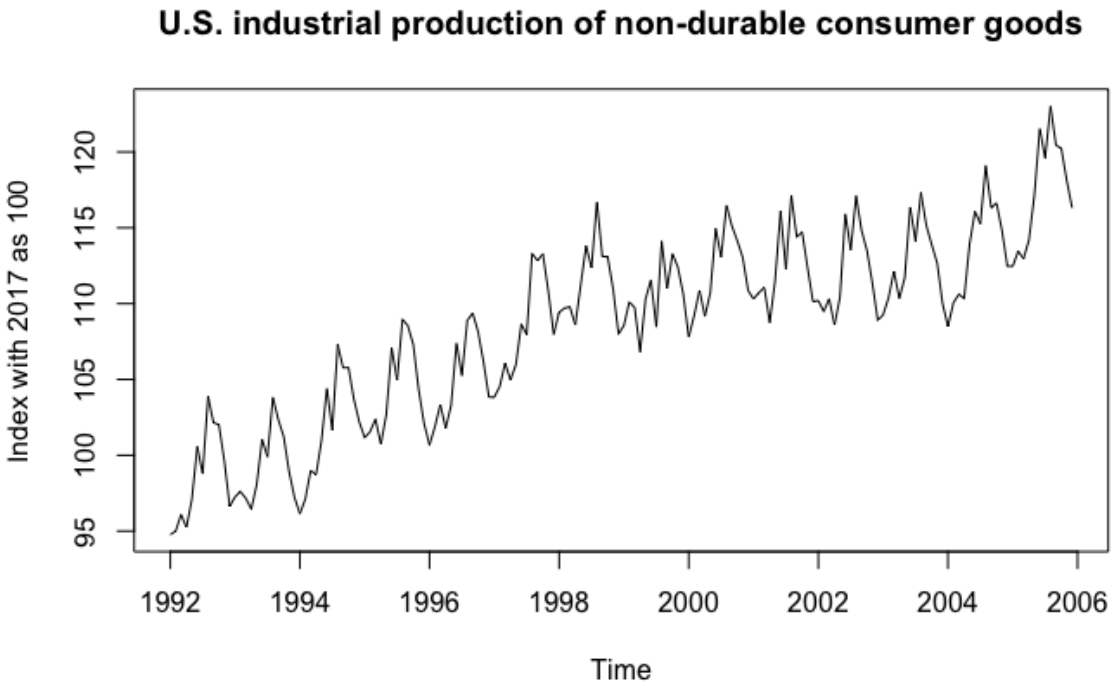

Figure 5: Time plot of U.S. industrial production of non-durable consumer goods

Referring to Figure 5, we can see a general upward trend in the time plot of U.S. industrial production of non-durable consumer goods, which could be expected based on increases in demand from U.S. and international consumers over time, as well as advancements in manufacturing technology to increase the efficiency of production. However, the time plot shows a plateau around 2000-2003, which could perhaps relate to the early 2000s recession. There also seems to be a seasonal component that does not show noticeable increases in seasonal swings over time, and we looked into it further using the seasonal box plot.

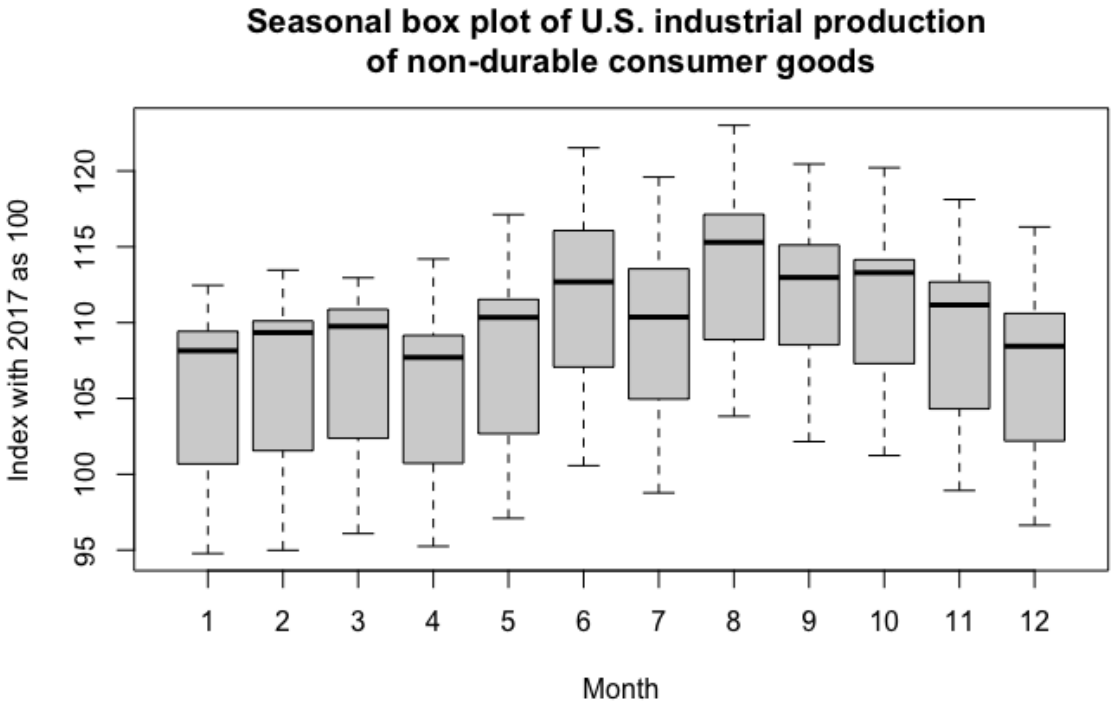

Figure 6: Seasonal box plot of U.S. industrial production of non-durable consumer goods

Referring to Figure 6, we can see a seasonal component in U.S. industrial production of non-durable consumer goods. Variability appears to remain relatively consistent between the different months of the year, but in general August, October, and June display the highest levels of U.S. industrial production of non-durable consumer goods, while April and January display the lowest levels of U.S. industrial production of non-durable consumer goods.

2.1 Decomposition

In this section, we will perform decomposition on U.S. imports of goods by customs basis from China and suggest an appropriate model based on its random component after analyzing the ACF and PACF plot.

We chose our dependent and independent variables to be the following, based on our group's interests:

- Dependent variable
  - U.S. imports of goods by customs basis from China
- Independent variables
  - Average hourly earnings of U.S. private sector production and nonsupervisory employees
  - U.S. industrial production of non-durable consumer goods

From the time plot of U.S. imports of goods by customs basis from China, we are able to notice both trend and seasonality present, which could obscure the signal in the random component without decomposition. We decided to perform decomposition to make the random term stationary, and we believed multiplicative decomposition would be preferred over additive decomposition due to increasing seasonality being visible in the time plot. We then performed both additive and multiplicative decomposition in order to confirm this belief.

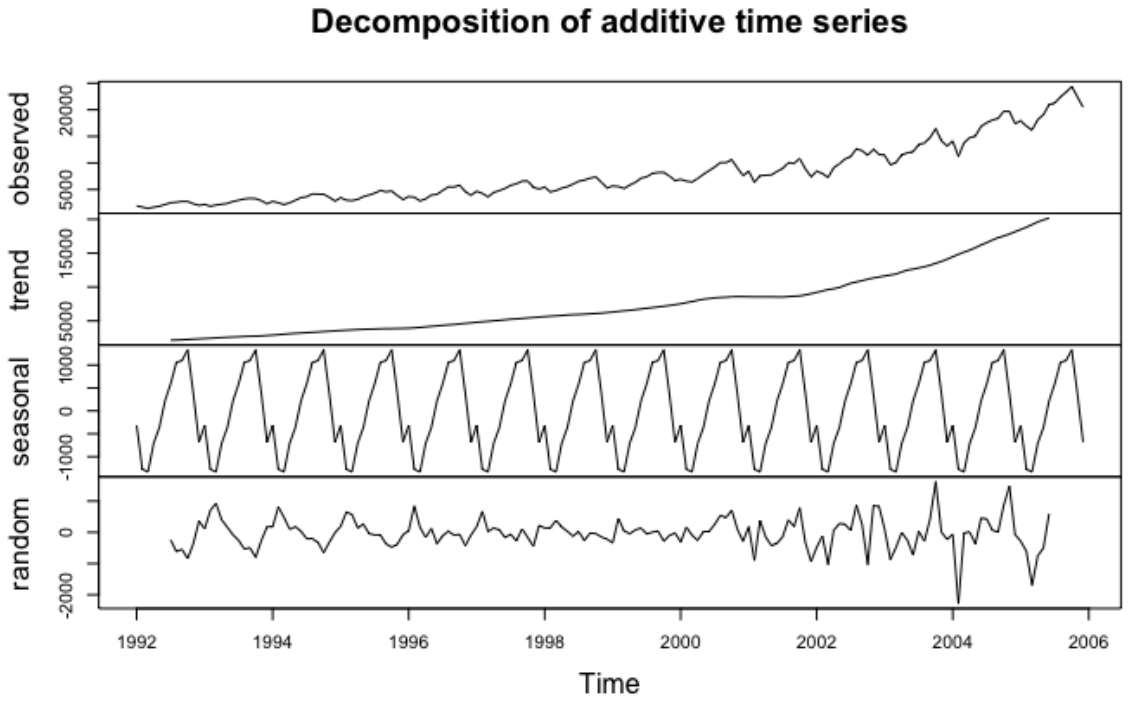

Figure 7: Additive decomposition of U.S. imports of goods by customs basis from China

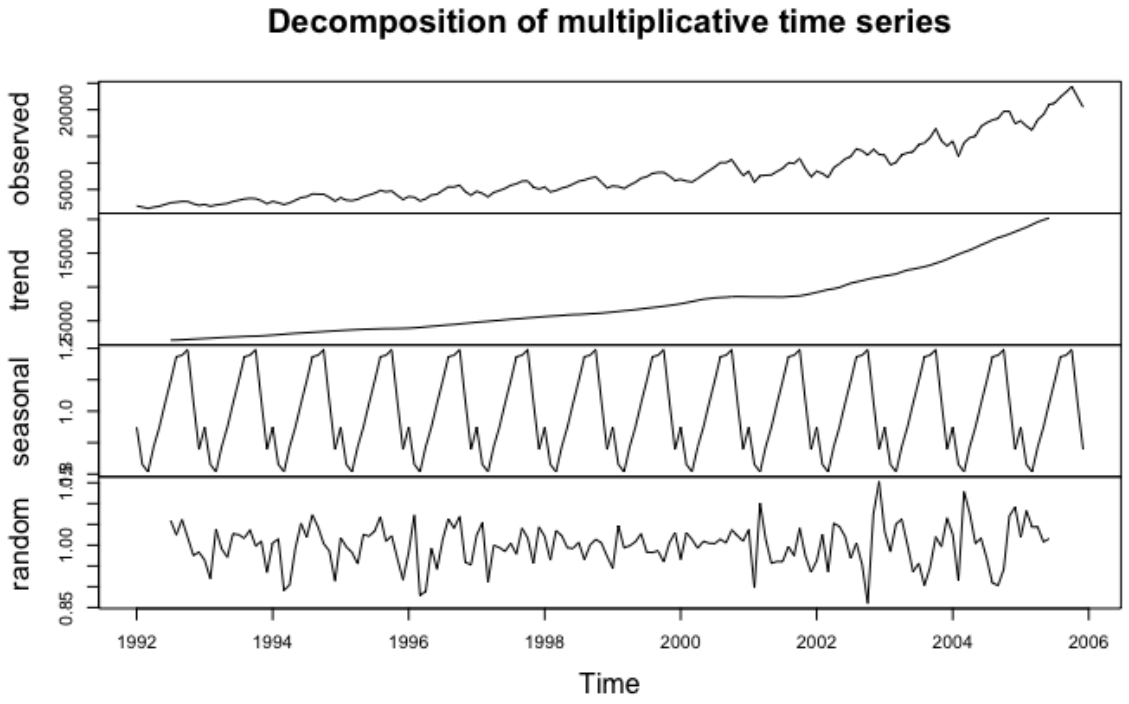

Figure 8: Multiplicative decomposition of U.S. imports of goods by customs basis from China

We performed additive decomposition (Figure 7) at first to have a basis for comparison, and confirmed that it was relatively ineffective compared to multiplicative decomposition (Figure 8) at

making the random term stationary. With additive decomposition, the random term displayed a significant change in variability starting in 2002 as compared to before 2002, while there was not such a pronounced change in variability for multiplicative decomposition. Thus, multiplicative decomposition was preferable for making the random term stationary, so we chose it to proceed.

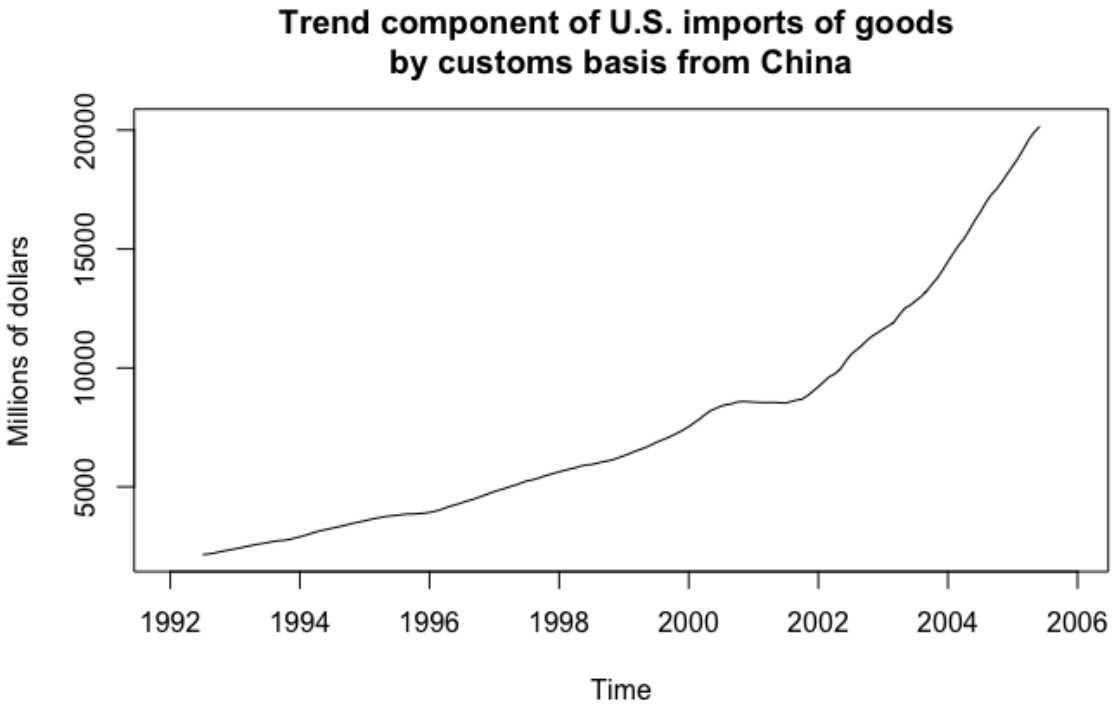

Figure 9: Trend component of U.S. imports of goods by customs basis from China

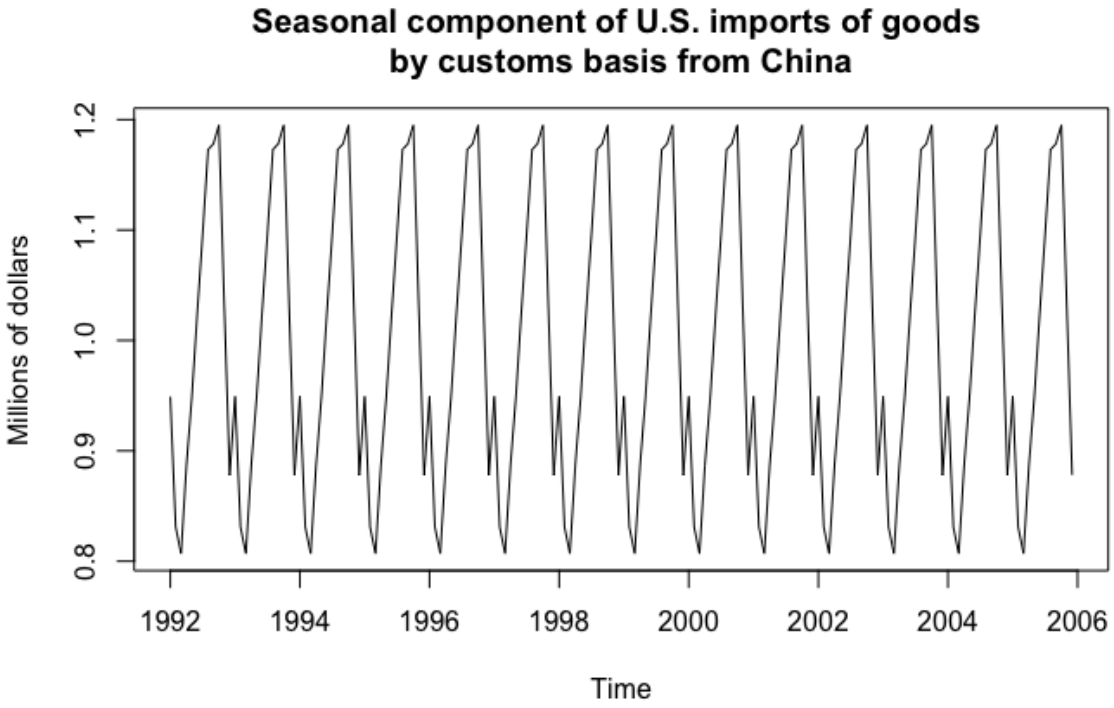

Figure 10: Seasonal component of U.S. imports of goods by customs basis from China

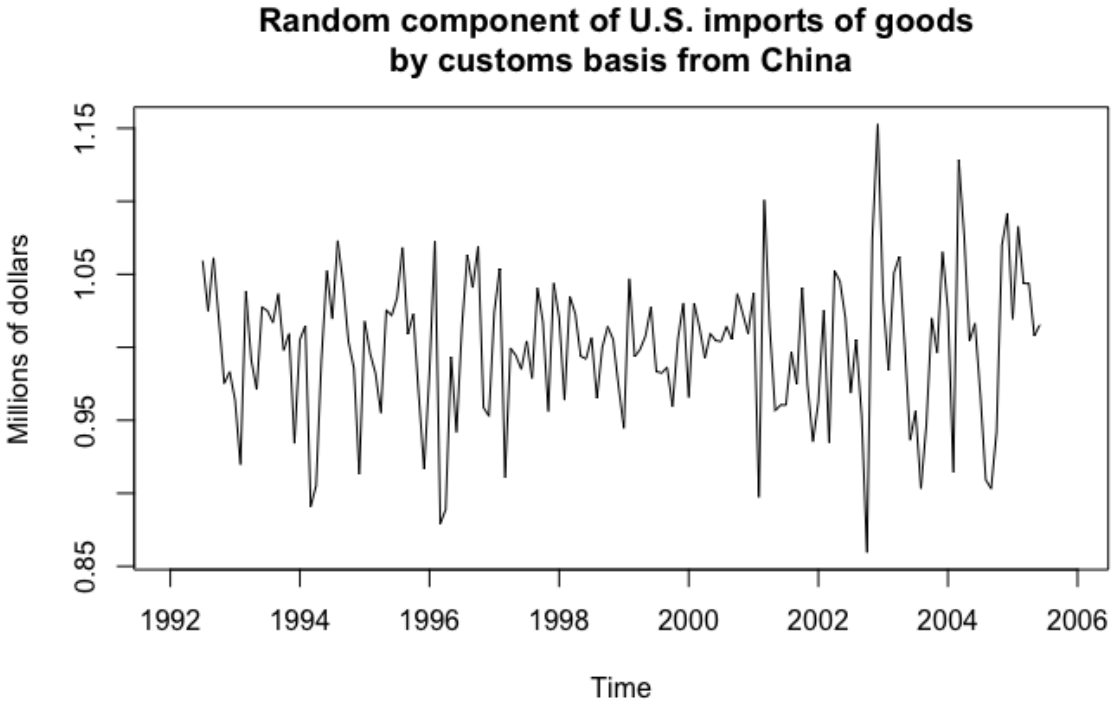

Figure 11: Random component of U.S. imports of goods by customs basis from China

From Figures 8, 9, 10, and 11 above, some notable points are:

- Throughout the years, the plot of the trend component is always increasing, except for a stagnant period in U.S. imports from China between 2000 and 2002, perhaps relating to the early 2000s recession.
- The plot of the random component shows fluctuations around a constant with less drastic changes in variability, indicating that multiplicative decomposition was the appropriate choice for making the random term stationary. There are still some changes in variability, suggesting that there is more information that can be learned from the random component.

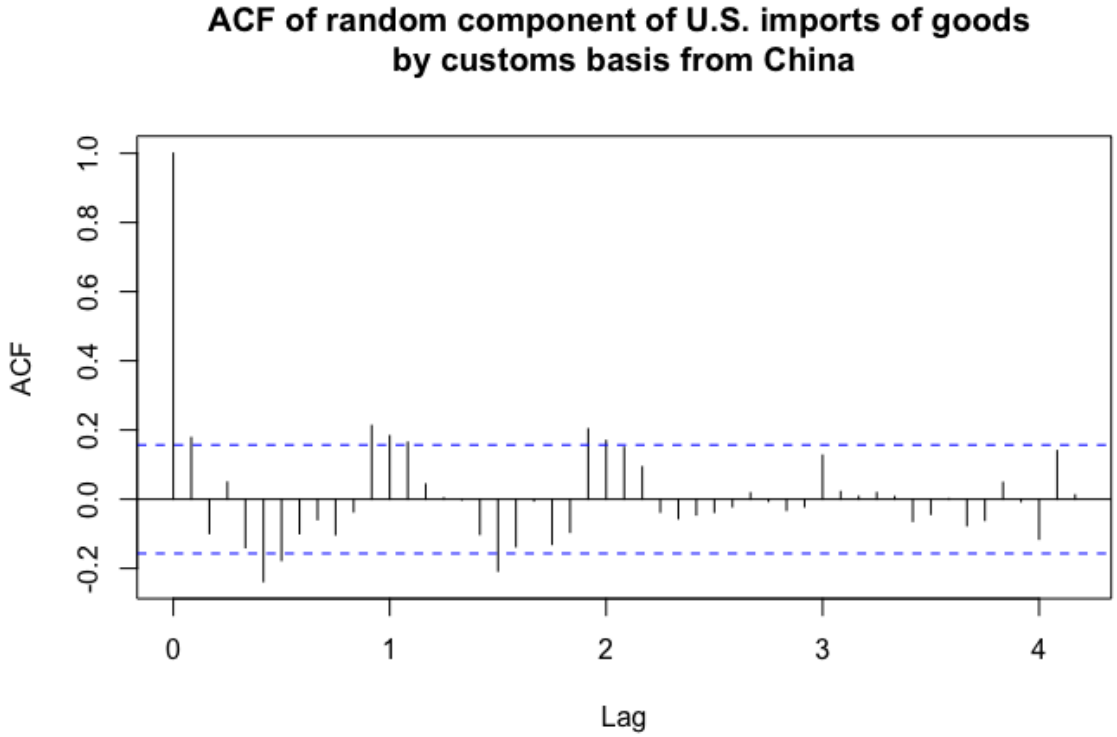

Figure 12: ACF of random component of U.S. imports of goods by customs basis from China

Referring to Figure 12, the sample ACF plot of the random component shows several statistically significant autocorrelations up to a moderately large number of lags, and these significant autocorrelations occur in a smooth, cyclical pattern, suggesting that an AR model may be appropriate. The autocorrelations decrease with increasing lag as they follow this smooth, cyclical pattern, indicating the presence of trend and seasonality. The significant autocorrelations up to a moderately large number of lags, which appear in this noticeable periodic pattern, suggest that the time series' random component is nonstationary.

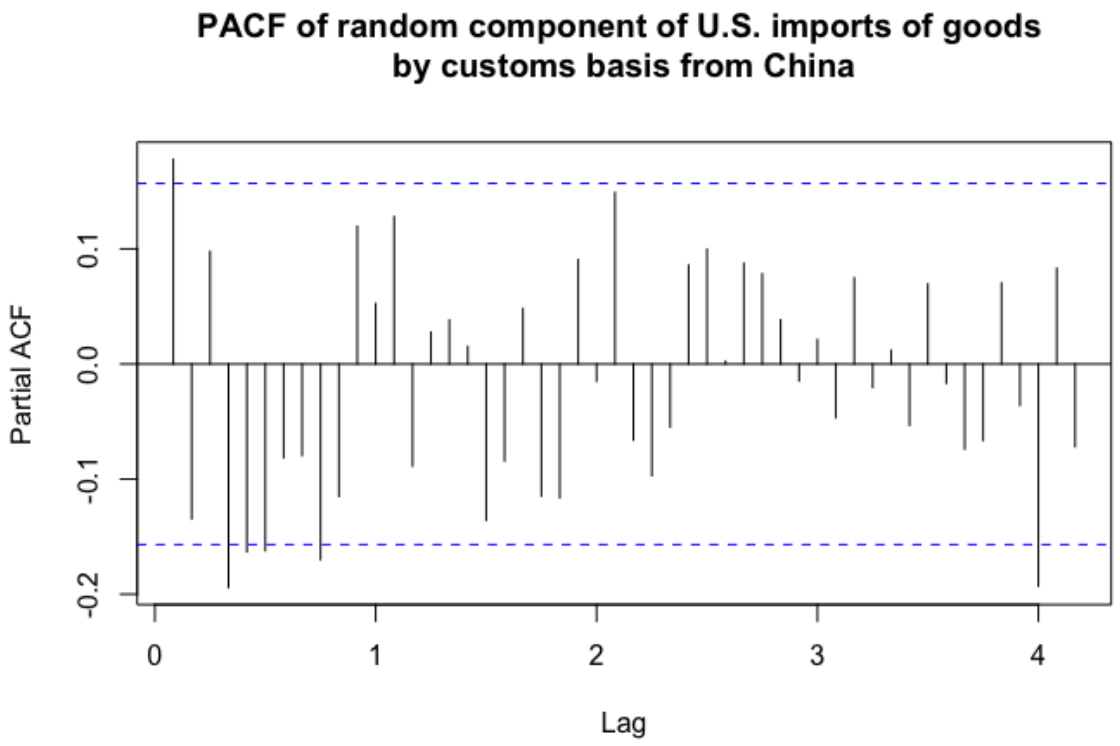

Figure 13: PACF of random component of U.S. imports of goods by customs basis from China

Referring to Figure 13, based on the PACF plot of the random component, we would identify this as an AR(1) model because the first partial autocorrelation is statistically significant, and it is not immediately followed by other significant partial autocorrelations. There are some significant partial autocorrelations at small lags – the fourth, fifth, and sixth partial autocorrelations are significant – but because the second and third are not significant, we believe an AR(1) model would be appropriate. (We later expand on this when identifying a model using ARIMA notation in Section 3.3.)

3 ARIMA Modeling and Forecasting

3.0 Updates from First Draft

Our project has been updated with a different date range (starting from 1992 instead of 1985) and a new independent variable (industrial production of non-durable consumer goods instead of unemployment rate). Both changes were made to avoid having a time series where different models would be more suitable to different time intervals. The different behavior of "U.S. imports of goods from China" time series during the 1980s could be attributed to several economic factors, including the early 1980s recession, fluctuations in the price of crude oil, and the U.S.-Japan trade war. Additionally, the large spikes in unemployment rate were likely impacted by economic recessions in the early 1990s and early 2000s. In comparison, the new time range and new independent variable show less volatility, allowing the models later in this project to achieve a better fit.

3.1 Pre-Transformations

In this section, we will discuss an alternate approach to obtain the random term, that is differencing, to prepare for ARIMA modeling. Referring to Figure 1, the time plot of US import of goods from China shows an existing trend and increasing variability. Hence, we will conduct two types of pre-differencing transformation – the logarithmic and square root transformation to the dependent variable – to choose the appropriate transformation to stabilize the variance.

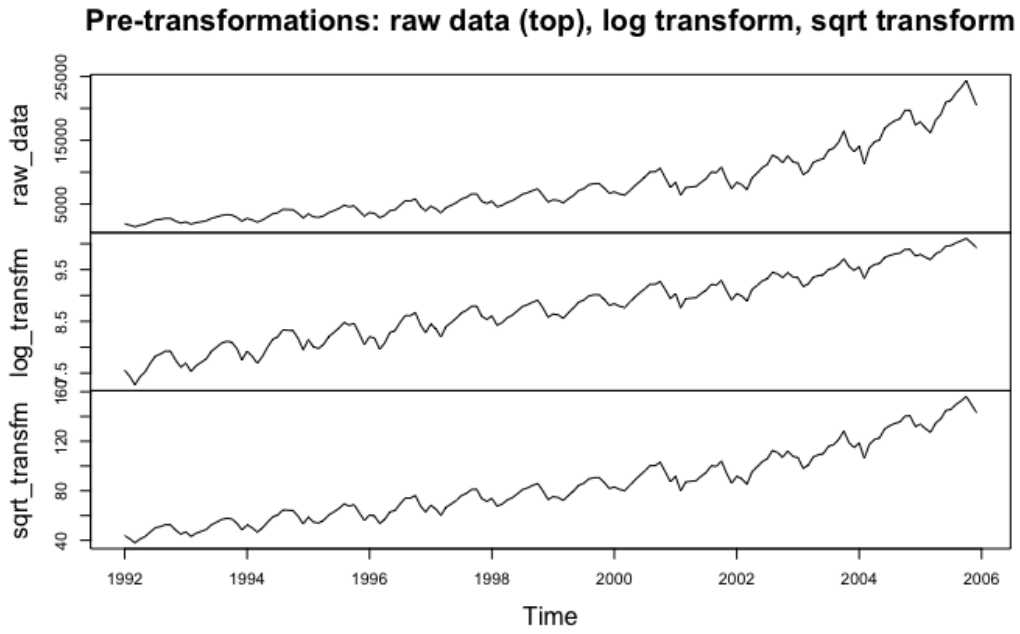

Figure 14: U.S. imports of goods by customs basis from China, raw data compared with logarithmic and square root transformation

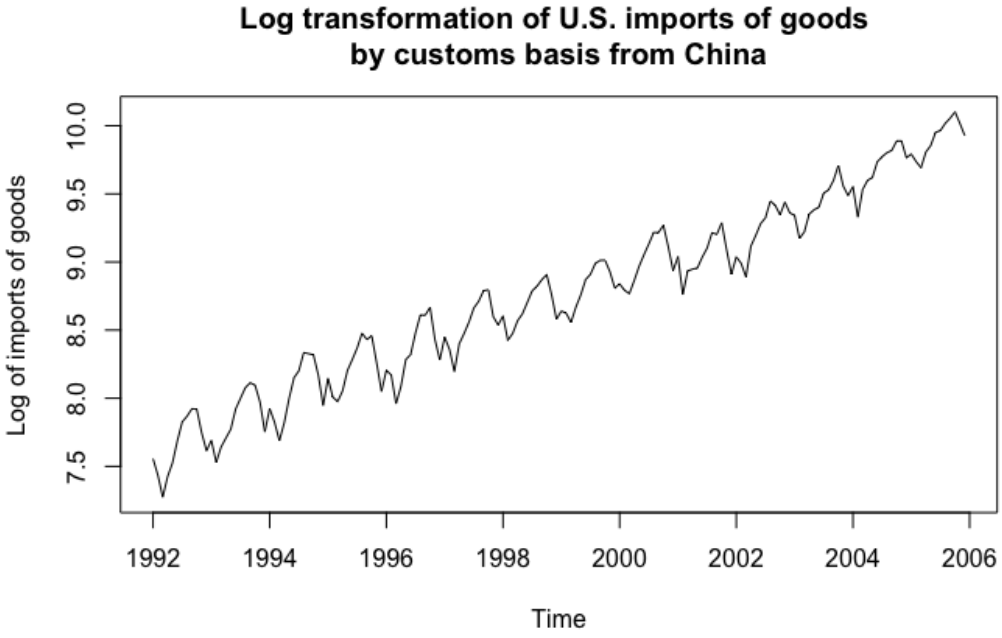

Figure 15: Time plot of logarithmic transformation of U.S. imports of goods by customs basis from China

Referring to Figure 14, the logarithmic transformation shows a greater improvement in terms of stabilizing the variability, bringing it relatively closer to constant variability as compared to the square root transformation. Figure 15 shows the time plot of the logarithmic transformed time series has variance that is relatively close to constant. With that being said, we chose to proceed with the logarithmic transformed time series of U.S. imports of goods from China, as our pre-transformed time series. To perform and decide between these pre-transformations, we referred to resources from Week 3: lecture segment 1 from January 18, as well as the discussion during class time from January 20.

3.2 Assessment of Mean Stationarity

Our goal was to make the pre-transformed time series stationary, so we explored three types of differencing: first-order regular differencing only, seasonal differencing only, and seasonal differencing of the first differencing. We used both the time plot and the ACF plot from the differencing of the pre-transformed time series when determining which differencing would successfully make the pre-transformed time series stationary. First and foremost, we attempted to account for trend by performing the first regular differencing.

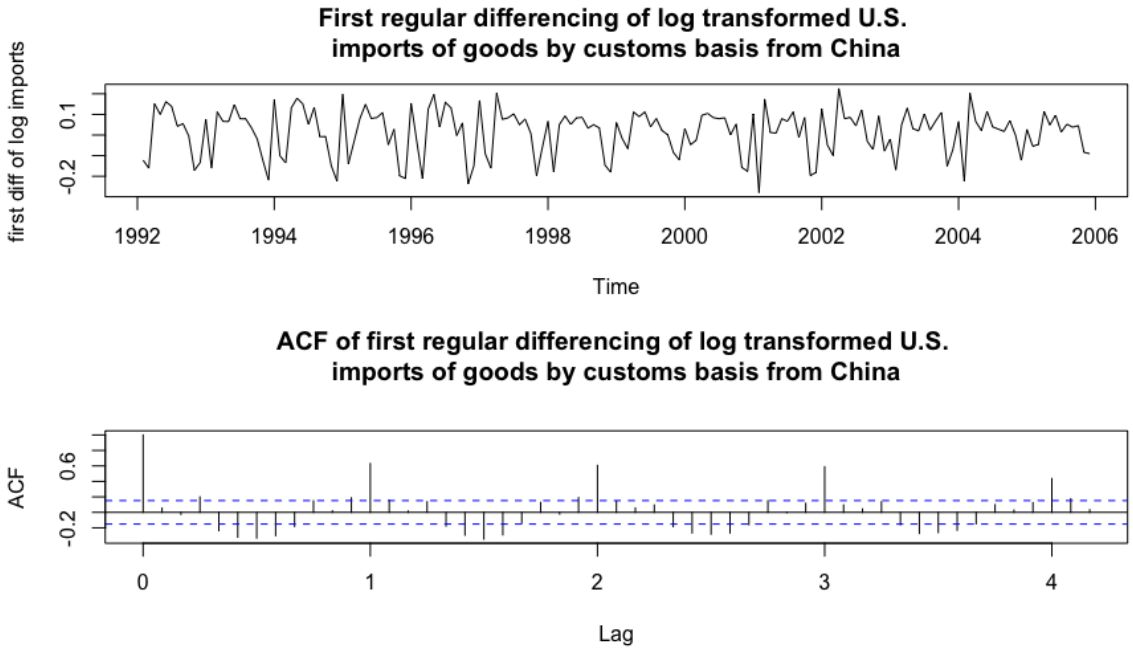

Figure 16: Time plot and ACF plot of first regular differencing of logarithmic transformed U.S. imports of goods by customs basis from China

Referring to Figure 16, after performing the first regular differencing to account for trend, we observed an improvement from the time plot of the pre-differencing log transformation, as the time plot after regular differencing appears to fluctuate around a constant rather than follow an increasing trend. However, it appears that seasonality is likely still present, as indicated by patterns of increases and decreases at relatively regular intervals. Furthermore, the ACF plot shows a smooth pattern of high autocorrelations up to large lags, and this cyclical pattern indicates the presence of seasonality. This reinforces the conclusion from the time plot. This periodic pattern of numerous significant autocorrelations up to large lags also suggests that the time series is nonstationary. Next, we attempted seasonal differencing to account for seasonality.

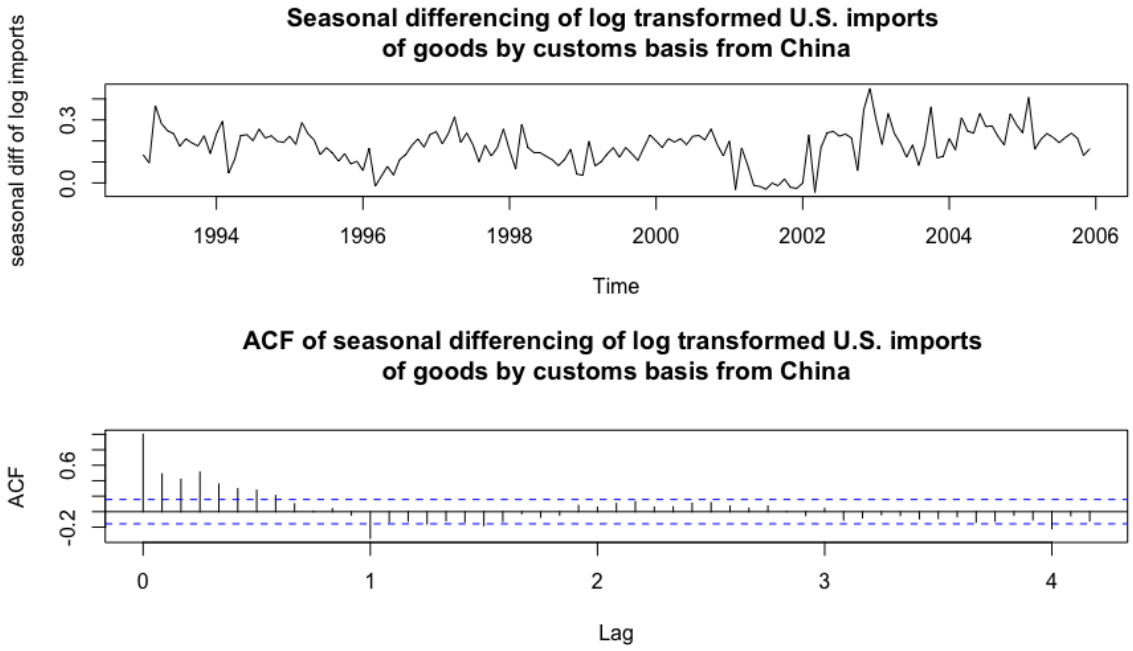

Figure 17: Time plot and ACF plot of seasonal differencing of logarithmic transformed U.S. imports of goods by customs basis from China

Referring to Figure 17, we see an improvement from the time plot of the pre-differencing log transformation, which had displayed noticeable seasonality at relatively regular intervals. However, the mean of the time series may not be adequately stabilized at a given constant, which could suggest the presence of an existing trend. From the ACF plot, we observe that the first seven autocorrelations (excluding a lag of 0) are significant, with significance at the seasonal lag (twelfth autocorrelation) as well. We also observe a few significant autocorrelations at large lags, but this can be expected behavior at a 5% significance level, as approximately 5% of the autocorrelations will be significant due to random chance. This would suggest a stationary time series, as does the ACF's resemblance to that of an MA process, since all MA processes are stationary. However, the time plot may still suggest that the mean requires further stabilization. We next attempted first regular and seasonal differencing to account for both trend and seasonality.

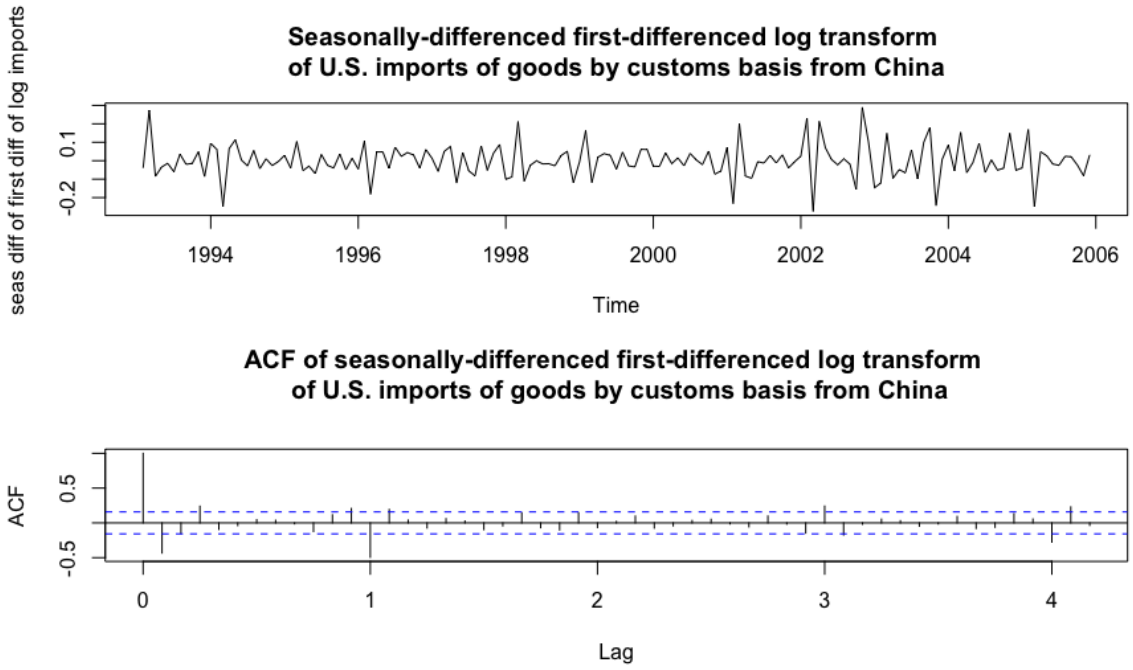

Figure 18: Time plot and ACF plot of seasonally-differenced first-differenced logarithmic transformed U.S. imports of goods by customs basis from China

Referring to Figure 18, after performing the first regular and seasonal differencing to account for both trend and seasonality, the time plot shows the time series fluctuating around a constant. This is an improvement from the time plot of the pre-differencing log transformation as the mean and variance of the time series appear to be sufficiently stabilized, and there does not appear to be an existing trend or existing seasonality. From the ACF, as only a small number of autocorrelations at large lags are significant, which can be expected considering that at a 5% significance level, approximately 5% of the autocorrelations will be significant due to random chance. This sufficiently small number of significant autocorrelations with no clear pattern in the random component's ACF suggests that the time series is stationary.

The seasonal differencing of the first-order regular differencing is the most suitable among the three options in order to arrive at a stationary time series. Moving forward, we identified the appropriate model by analyzing the ACF and PACF plots of the random term left below.

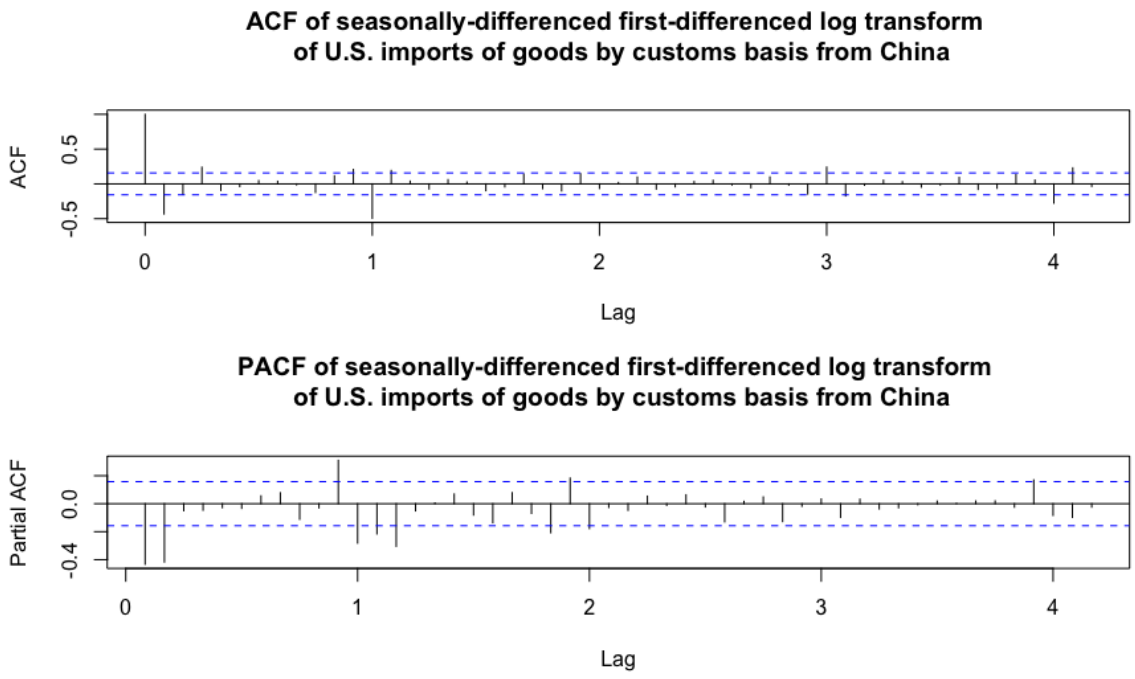

Figure 19: ACF and PACF plot of seasonally-differenced first-differenced logarithmic transformed U.S. imports of goods by customs basis from China

Referring to Figure 19, the ACF plot shows that the autocorrelation at lag 1, r_1, is significant and and is not immediately followed by other significant autocorrelations. Furthermore, the PACF plot reveals a decaying pattern as the lag increases. Both of these features are distinct characteristics of an MA(1) model. Additionally, at large lags, the significant autocorrelations and partial autocorrelations are few enough to be reasonably expected under a 5% significance level. We thus came to a conclusion that an MA(1) model would be a more fitting model than an AR(1) model. Notably, we identified a different model than in Section 2.1. After multiplicative decomposition we chose an AR(1) model, while after pre-transformation and differencing we chose an MA(1) model. (We expanded on both of these when identifying a model using ARIMA notation in the following section.)

3.3 Identification

In this section, we identified that a suitable first tentative model is $ARIMA(2,1,0)(0,1,1)_{12}$ with reference to Figure 19 because:

- Firstly, we performed the seasonal differencing of the first-order regular differencing to arrive at a stationary time series, hence d = D = 1.
- For the stationary part at low lags, we chose p = 2 indicating an AR(2). The ACF plot reveals a decaying cyclical pattern involving several statistically significant autocorrelations r_k; hence we suspected that an AR model would be fitting. At small lags, we observe that significance in the ACF plot dies away quickly, which can be expected for an AR model. To confirm that an AR model would be suitable and select the order of the model, we looked at the PACF plot. We can observe that among low lags, only the first 2 partial autocorrelations are significant, and significance cuts off afterward; hence we believe an AR(2) model would be a fitting model for the regular (low lags) part.
- For the stationary part at seasonal lags, we chose Q = 1 indicating that we suspect an MA(1) would be appropriate because the ACF shows statistically significant seasonal lags that cut off quickly, and the PACF shows statistically significant seasonal lags that die away quickly. The autocorrelations at the 11th, 12th, and 13th lags were statistically significant, with the 12th having an especially large absolute value around 0.5, and significance cutting off after the 13th

lag. For further inspection, referring to the PACF, we can observe that there is structure around the statistically significant seasonal lags that begins at the 11th lag and dies away quickly after the 14th lag; hence we chose an MA(1) model for the seasonal part.

- For the frequency, we chose F = 12 since we are working with monthly data.

3.4 Fit and Diagnose

In this section, we will perform a diagnosis of our identified model by exploring its residuals with the ACF plot and the Ljung Box test. In order to ensure true statistical inference, we will check for normality in the residuals of the model fitted by looking at the histogram. Last but not least, we will also attempt to fit another model and select the better model.

We first analyse the ACF and histogram of the residuals of our identified model in 3.3, that is, $ARIMA(2,1,0)(0,1,1)_{12}$.

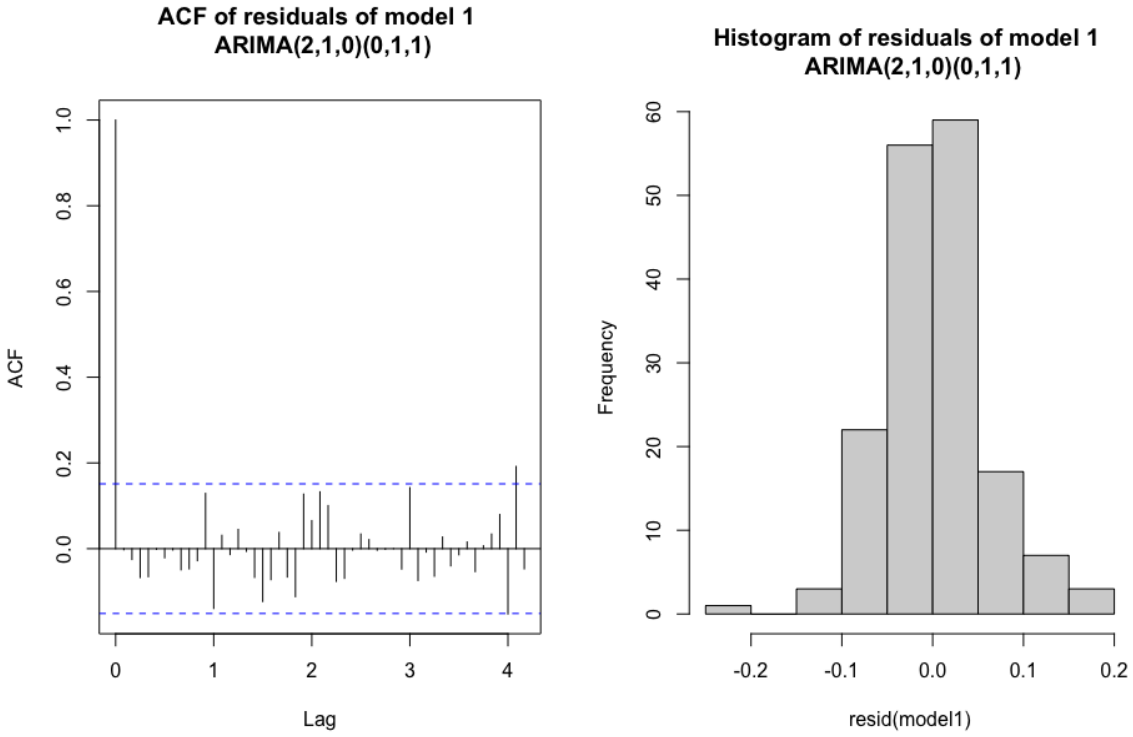

Figure 20: ACF and Histogram of residuals of model 1

Referring to Figure 20, we observe the following:

- The ACF of the residuals shows that the residuals of model1 fit are white noise. We can observe a small amount of statistically significant autocorrelations occurring at higher lags and they appear randomly scattered; this is probably due to the 5% significant autocorrelations occurring by chance.
- Referring to the histogram of the residuals to check the normality of the residuals, it looks relatively close to normal; however, it is slightly left-skewed.

Upon performing the Ljung-Box test for lower and larger lags (12, 24, 50), we get p-values of 0.678, 0.5641, and 0.2823 respectively, which are greater than 0.05 for all the tests conducted. Hence, we fail to reject the null hypothesis that the residuals are white noise, providing evidence suggesting that the residuals are white noise. From the above observations, we can conclude that this is a fairly good model.

However, we would like to explore more models to see if we can improve our tentative model. So, we will inspect the ACF and PACF of the model with seasonal differencing of the first regular difference in Figure 19 again. The PACF plot displays several statistically significant seasonal lags before the significant lags die away. Hence, we attempted to model the seasonal component with an MA(2) model for model 2, that is, $ARIMA(2,1,0)(0,1,2)_{12}$.

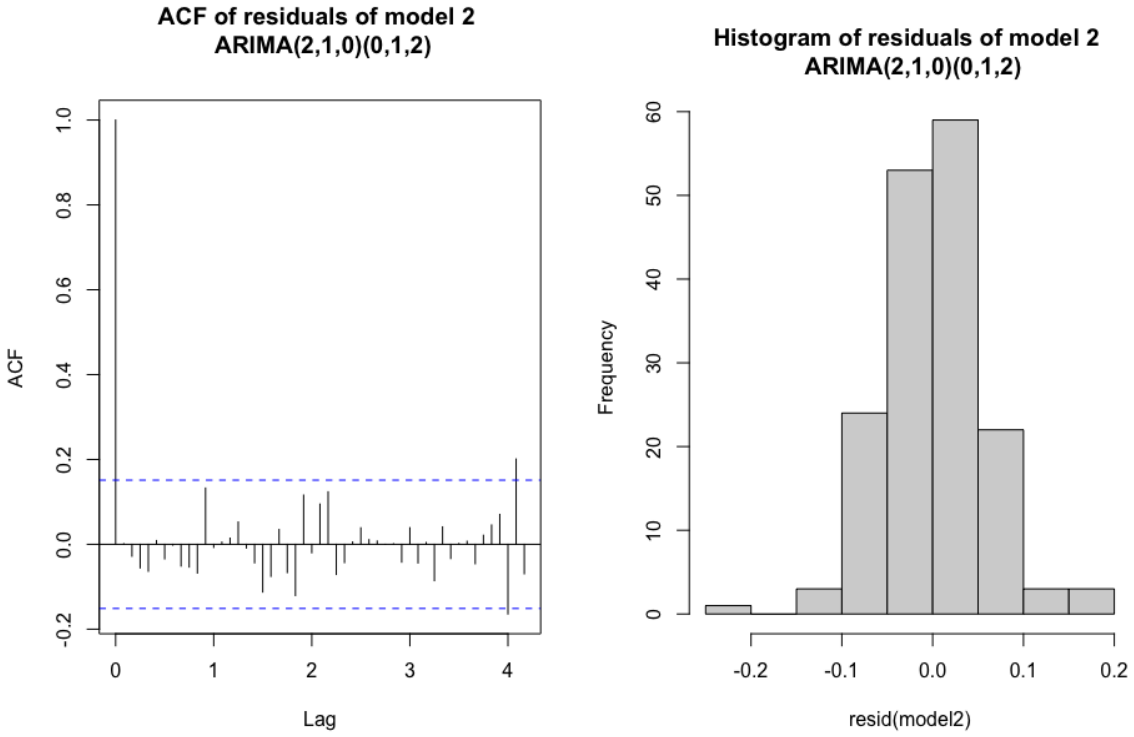

Figure 21: ACF and Histogram of residuals of model 2

Referring to Figure 21, we observe the following:

- Referring to the ACF of the residuals of model 2, there is a slight improvement in comparison to model 1 where we considered an MA(1) model for the seasonal component. The autocorrelations at moderate to high lags are generally lower in absolute value than for model 1, and there are still a small amount of significant autocorrelations at high lags, but those are few enough to be expected in the 5% of significant lags by random chance under a 5% significance level.
- Referring to the histogram of the residuals, the histogram showed a slight improvement in terms of normality of the residuals in comparison to that of model 1.

Upon performing the Ljung-Box test for lower and larger lags (12, 24, 50), we get p-values of 0.8822, 0.8162, and 0.5829 respectively, which are greater than 0.05 for all the tests conducted. Hence, we fail to reject the null hypothesis that the residuals are white noise, providing evidence suggesting that the residuals are white noise. From the above observations, we observe an overall improvement from model 1. Both model 1 and model 2 are relatively good models, but to choose the better model, we will be using the AIC as a measure for model selection.

	RMSE	AIC
Model 1	0.05868446	-423.8037
Model 2	0.05702505	-429.4200

Table 2: Table of RMSE and AIC of models

As expected, as we add more parameters, the RMSE will always decrease. Model 2 has the lowest AIC of the models, indicating that we are maximizing the information we can obtain from the data.

Thus we select model 2 as our final model.

Since we performed the log transformation to the dependent variable, $y_t^* = log(y_t)$. The polynomial form of the final model is:

$$(1 + 0.5883B + 0.4387B^2)(1 - B^{12})(1 - B)y_t^* = (1 - 0.8062B^{12} + 0.2315B^{13})w_t$$

The root of the polynomial for the AR regular lag model is 1.509789, larger than the unit circle, so the model is stationary. The root of the polynomial for the MA seasonal model is 2.078378, larger than the unit circle, so the model is invertible. As such, we can confirm that the model is stationary and invertible. The final forecasting equation is:

$$y_t^* = 0.4117y_{t-1}^* + 0.1496y_{t-2}^* + 0.4387y_{t-3}^* + y_{t-12}^* - 0.4117y_{t-13}^* - 0.1496y_{t-14}^* - 0.4387y_{t-15}^* + w_t - 0.8062w_{t-12} + 0.2315w_{t-13}$$

3.5 Forecasting

In this section, we will proceed to forecast the test period, that is, Jan 2006 - Dec 2006 using the ARIMA model we chose in 3.4. The table 3 below summarises the test data, forecast, prediction interval and standard error for the test period.

Month	Y (Test Data)	Forecast	Prediction Interval	Standard Error
Jan 2006	21382.51	21498.92	(19225.43, 24041.27)	0.05703
Feb 2006	17905.35	18235.60	(16159.47, 20578.47)	0.06167
Mar 2006	20531.26	19755.91	(17416.70, 22409.30)	0.06430
Apr 2006	21459.07	22087.60	(19116.76, 25520.14)	0.07370
May 2006	22317.64	22838.74	(19570.78, 26652.40)	0.07879
Jun 2006	23989.70	24771.11	(21068.18, 29124.85)	0.08261
Jul 2006	24632.01	26121.29	(21984.99, 31035.80)	0.08795
Aug 2006	26713.34	27354.06	(22822.15, 32785.90)	0.09242
Sep 2006	27570.62	28122.28	(23284.91, 33964.60)	0.09630
Oct 2006	29388.60	29856.26	(24519.43, 36354.69)	0.10047
Nov 2006	27775.08	28357.58	(23111.01, 34795.22)	0.10438
Dec 2006	24109.19	25683.18	(20782.48, 31739.50)	0.10802

Table 3: Table of imports test data, forecast, prediction interval, and standard error for Jan 2006 - Dec 2006

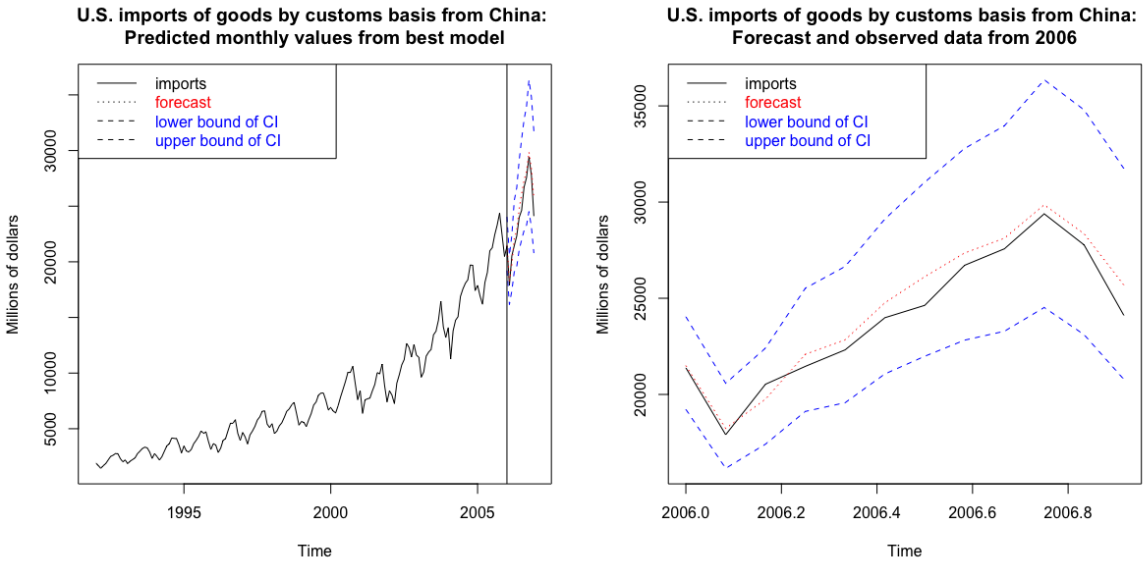

Figure 22: U.S. imports of goods by customs basis from China: test and forecast data

Referring to Figure 22, we can observe that our model is able to forecast the imports test data relatively well. We measured the accuracy of the forecasts using the test root mean square (RMSE) statistic and obtained a value of 814.9264. As we forecasts even further into the future, the standard error of our point forecast increases, which is expected as our accuracy of point forecast will decrease.

4 Multiple Regression

To begin this section, we renamed some variable for simplicity's sake.

1. **y**: our dependent variable, which is U.S. imports of goods by customs basis from China, using a monthly frequency and not seasonally adjusted
2. **x1**: one of our independent variables, which is average hourly earnings of U.S. private sector production and nonsupervisory employees, using a monthly frequency and not seasonally adjusted
3. **x2**: one of our independent variables, which is U.S. industrial production of non-durable consumer goods, using a monthly frequency and not seasonally adjusted
4. **time**: a dummy variable that indicates time. The value 1 is assigned to the first time series observation, and subsequent values are assigned according to the order in which those observations occurred.
5. **D1, D2, D3 ... D11**: dummy variables that indicate month of the year in order to simulate seasonality. D1 is a binary dummy variable that indicates whether this observation occurred in the first month of the year, D2 is a binary dummy variable that indicates whether this observation occurred in the second month of the year, etc.

4.1 Causal Model Fit

Here, we will be fitting our dependent variable to our independent variables. From observing our data, we also decided to add a few dummy variables to simulate seasonal and temporal elements in the time series. We also transformed our dependent and independent variables to stabilize their variances. After some adjustment and exploration, our regression model becomes

$$y^{\frac{1}{4}} = x_1^{-1.2} + x_2^{\frac{1}{2}} + time + D1 + D3 + D7 + D8 + D9 + D10 + D11$$

We first observe the distribution and ACF of the residuals of this multiple regression model.

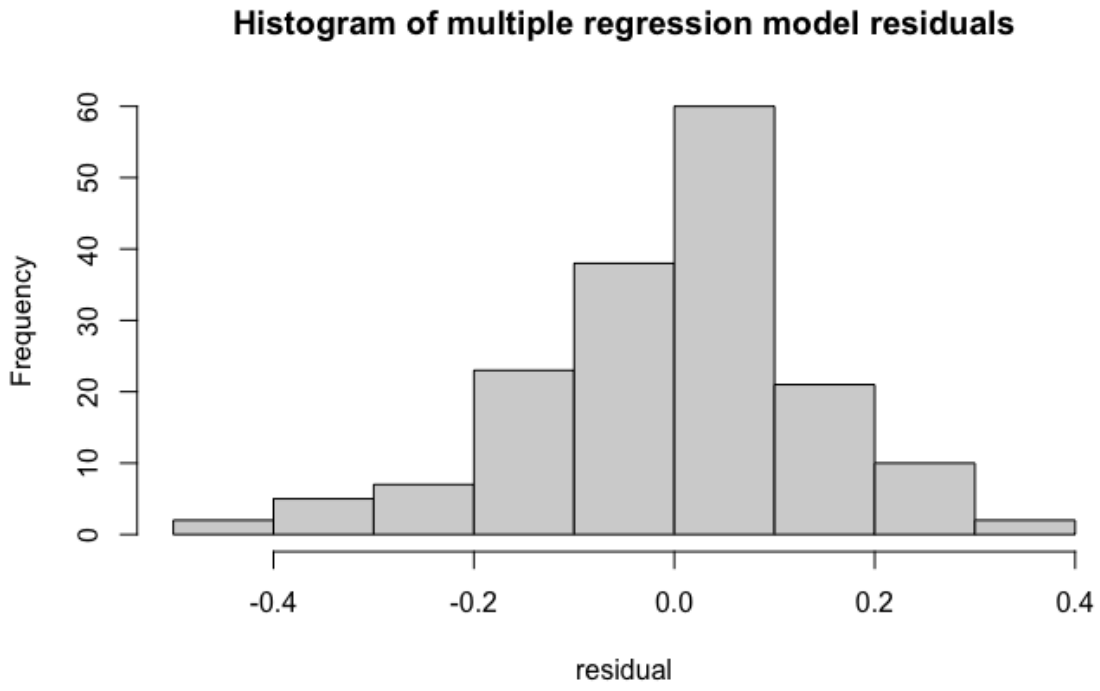

Figure 23: Histogram of the residuals of the multiple regression model

Referring to Figure 23, the histogram shows that our residuals are approximately normally distributed around zero, which satisfies the assumptions in our later analysis.

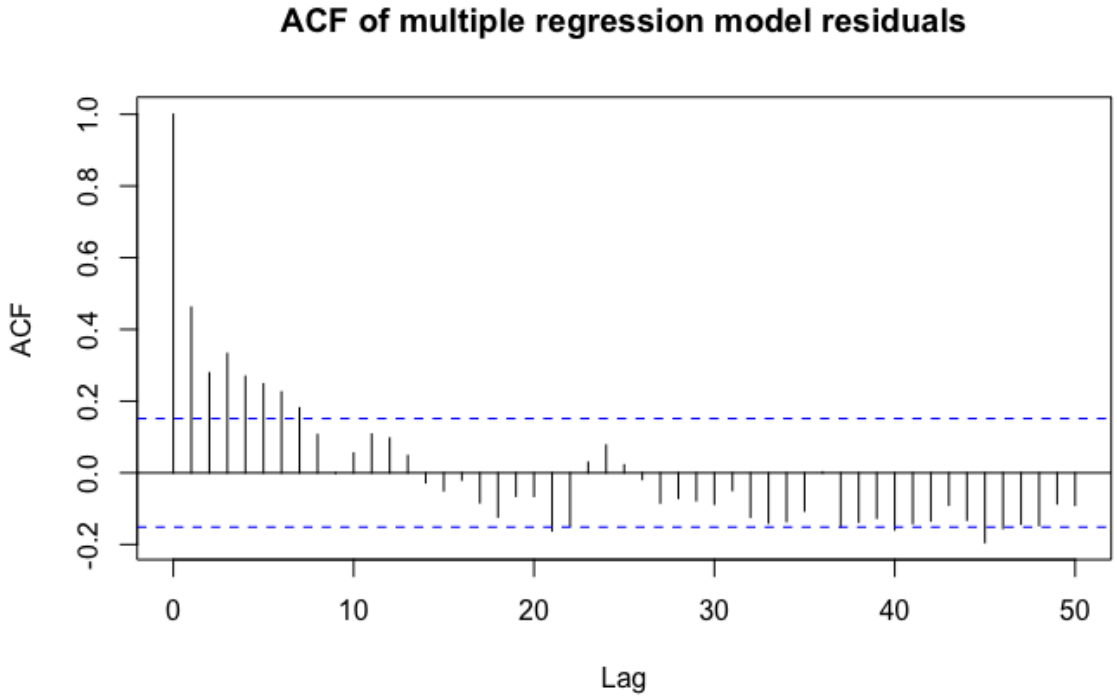

Figure 24: ACF plot of the residuals of the multiple regression model

However, referring to Figure 24, the ACF plot reveals that the residuals are not white noise. The ACF plot shows strong autocorrelations at the lower lags, suggesting an AR model perhaps with a larger coefficient.

We also conducted a Ljung Box test on the residuals to verify our finding. The corresponding p-value for the Ljung Box test at lag 20 is approximately 2.22×10^{-16}, from which we reject our null hypothesis that the residuals are white noise.

Next, we observe the ACF and PACF of the residuals to identify a model.

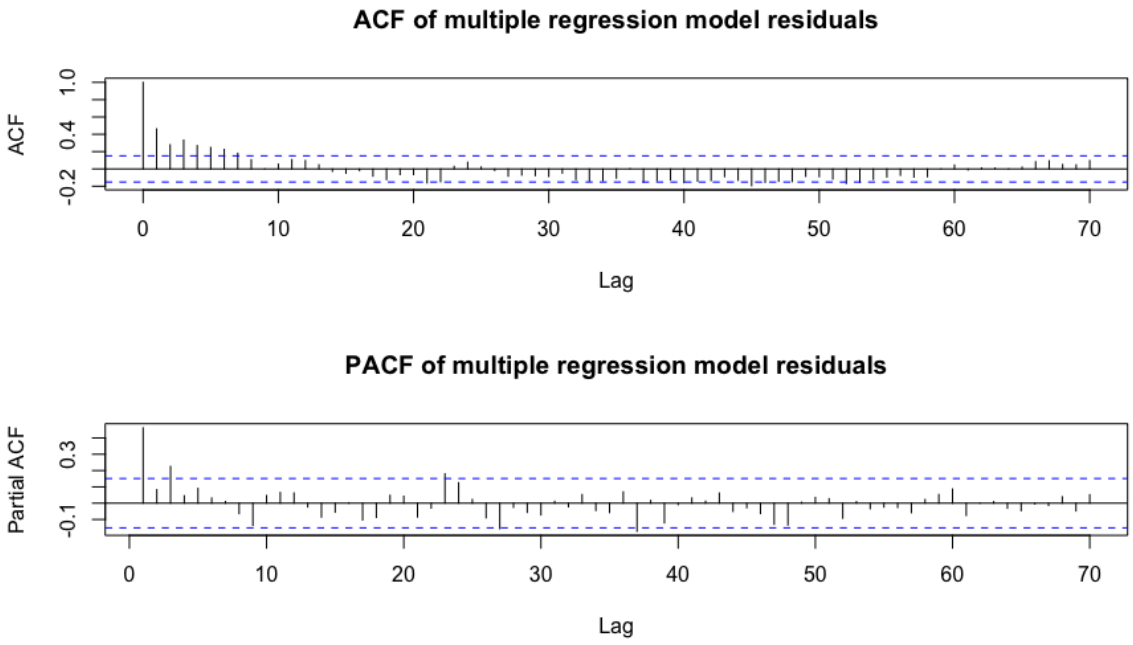

Figure 25: ACF and PACF plots of the residuals of the multiple regression model

Referring to Figure 25, from the ACF and PACF plot, we identified a ARMA(3,0,0) model for our residual. We then proceeded to model the residuals with this ARMA model and observe the residuals of the residuals after fitting ARMA(3,0,0).

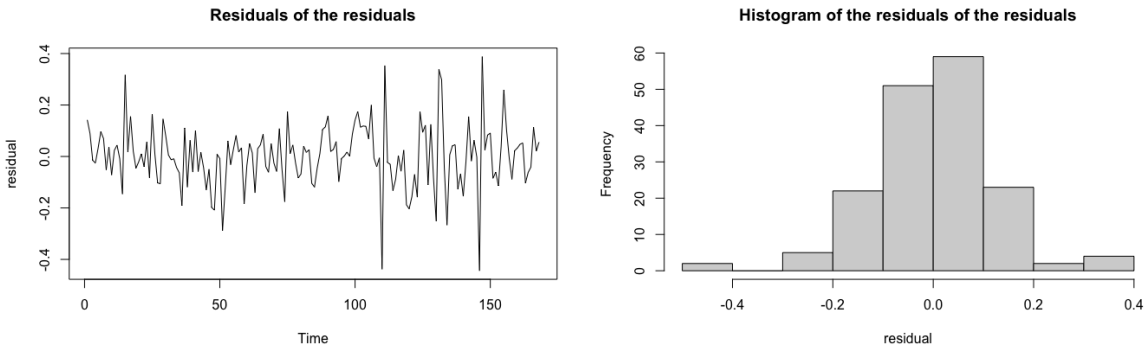

Figure 26: Plot and Histogram of the residuals of the residuals after fitting ARMA(3,0,0)

Referring to Figure 26, we observe that the residuals of the residuals after fitting have no obvious pattern, and that they are approximately normally distributed around zero.

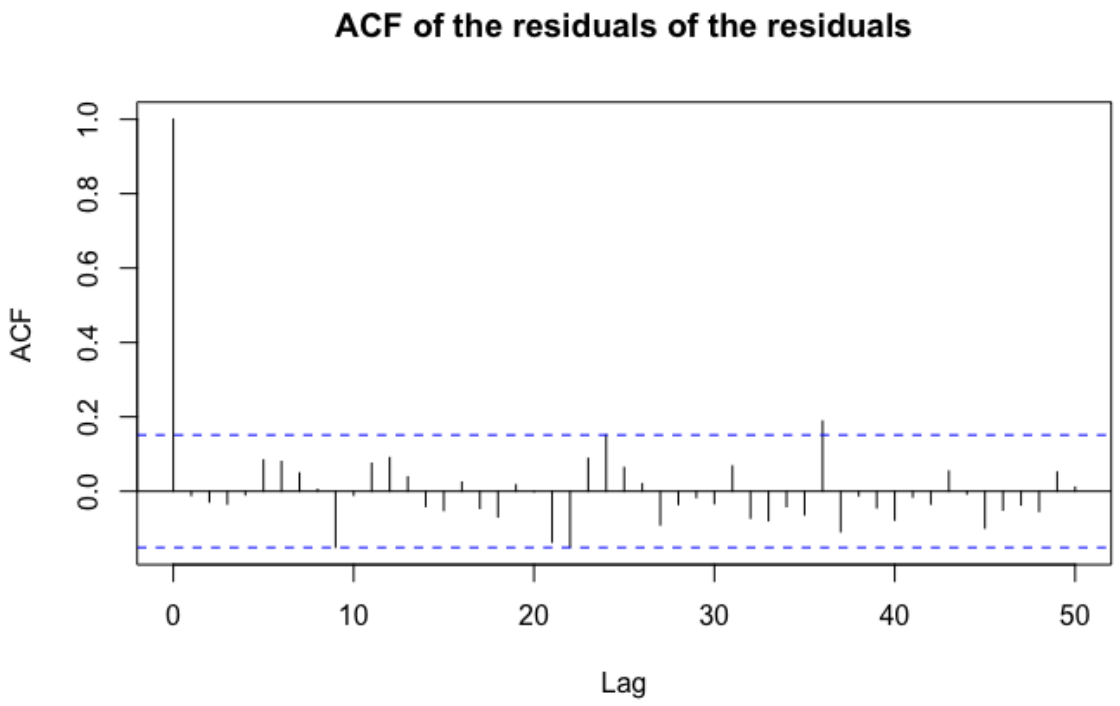

Figure 27: ACF plot of the residual of residuals after fitting ARMA(3,0,0)

Referring to Figure 27, the ACF plot of the residuals of residuals appear as an ACF plot of white noise. We thus conclude that the residuals of residuals after fitting ARMA(3,0,0) are white noise and confirmed this finding with a Ljung Box test. The corresponding p-value to lag 50 is 0.4132, from which we fail to reject the null hypothesis that the residuals of the residuals are white noise.

After getting the appropriate coefficients to our model, our final model using gls() is:

$$\begin{aligned} y^{\frac{1}{4}} = &- 20.73 + 349.13 \times x_1^{-1.2} + 0.66864 \times x_2^{\frac{1}{2}} + 0.079 \times time + 0.272 \times D1 \\ &- 0.211 \times D3 + 0.21 \times D7 + 0.221 \times D8 + 0.469 \times D9 + 0.54 \times D10 + 0.273 \times D11 \\ &+ 0.404\varepsilon_{t-1} - 0.008\varepsilon_{t-2} + 0.226\varepsilon_{t-3} \end{aligned}$$

We fitted this model to our testing data set and calculated a root mean square error of 1435.621. From the time plot in Figure 28 below, we can see that the model captured the general shape and seasonal component very well, although there could definitely be improvements.

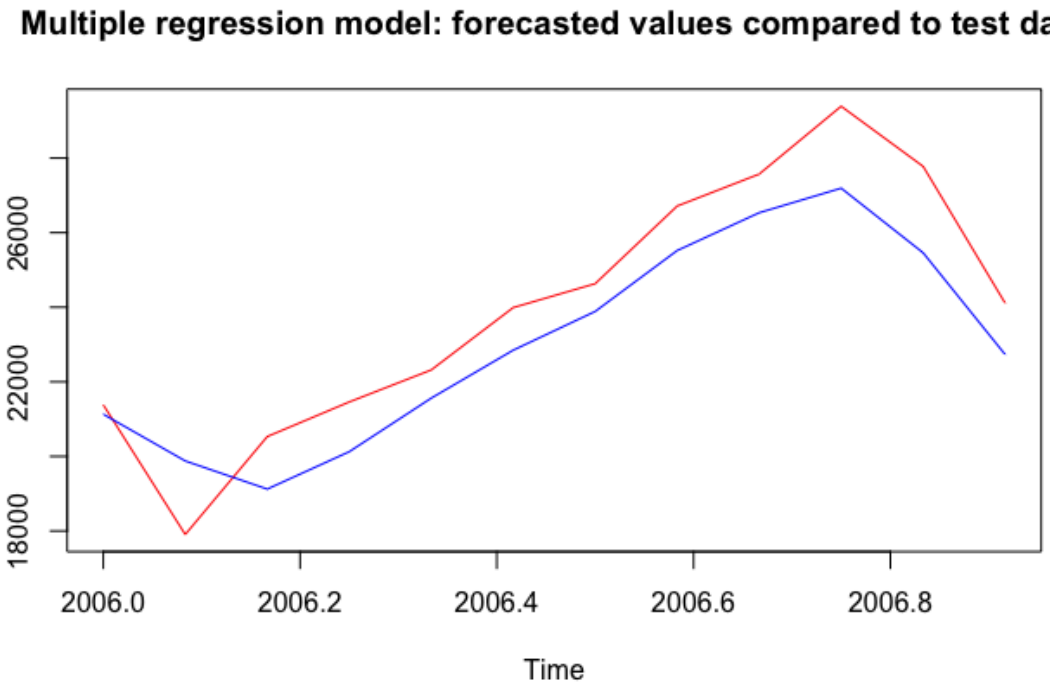

Figure 28: Forecasted values compared to testing data for multiple regression model

4.2 Regression with Dummies and Polynomials Only

For this regression, we are only considering dummy variables. To capture more information, we also introduced two additional variables:

1. **timesq**: the original time variable but squared.
2. **timecb**: the original time variable but cubed.
3. **tr**: a binary variable that indicates the start of a new trend. After observing our dependent variable, we concluded that a new trend started around observation 108 (January of 2001). Thus, this binary variable codes observations before January 2001 as 0, and everything after it is coded as 1.

The model we have is:

$$\log(y) = time + timesq + timecb + tr + D1 + D2 + D3 + D4 + D5 + D6 + D7 + D8 + D9 + D10 + D11$$

We observe the residuals. The residuals are approximately distributed around 0, so that satisfies the assumptions we had for our dummy regression model - see Figure 29.

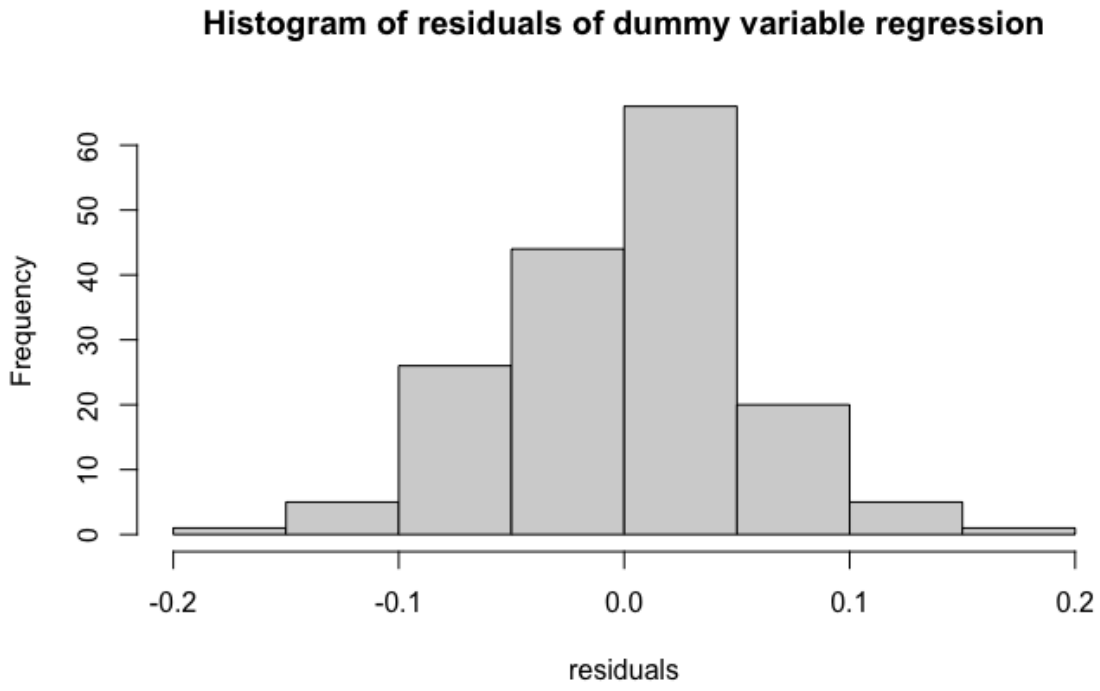

Figure 29: Histogram of residuals after dummy variable fitting

After observing the ACF and PACF plot in Figure 30, we can also confidently say that the residuals are not white noise. Our Ljung Box test confirms this with a p-value of 5.426×10^{-10}. Looking at our ACF, we see significant autocorrelations at lower lags, but it eventually dies off at higher lags. Our PACF indicates that we have a few significant lags at high lags, so we think a hybrid MA and AR model might be best. We then identified a ARMA(3,0,1) model for our residual.

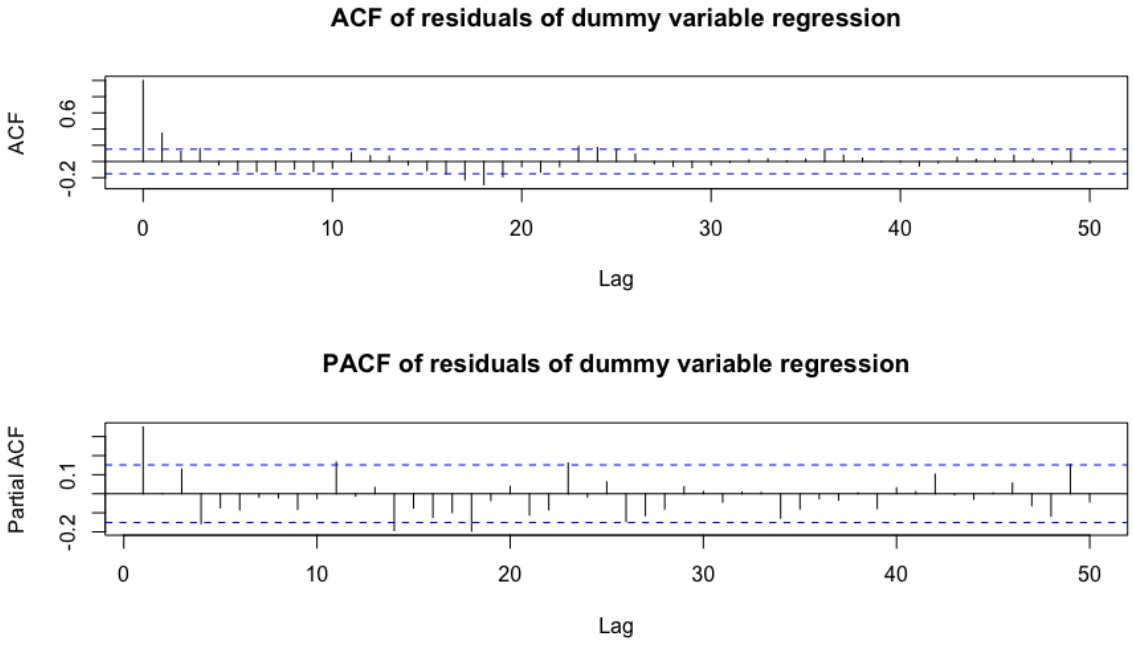

Figure 30: ACF and PACF of residuals of dummy variable regression

After modeling our residuals, we check the validity of our residual model. Referring to Figure 31, We do not observe any significant patterns. Our residuals of residuals are also approximately normally distributed around zero as seen in the histogram.

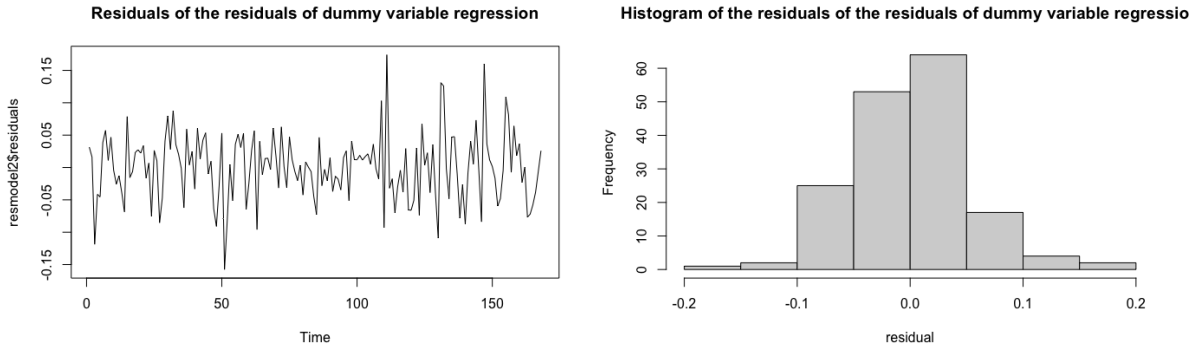

Figure 31: Plot and Histogram of the residuals of the residuals of dummy variable regression

We also observe the ACF of the residuals of residuals in Figure 32 and we see that it is similar to that of a white noise. Our Ljung Box test at lag 40 confirms this with a p-value of 0.07768, failing to reject the null hypothesis that the residuals of residuals are white noise.

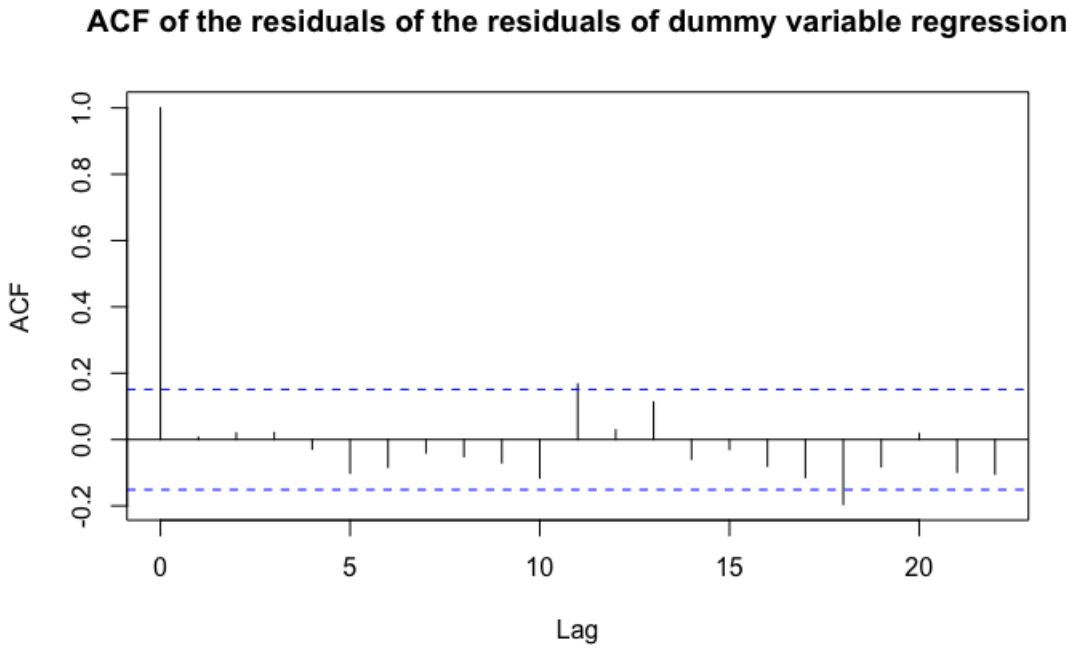

Figure 32: ACF of residuals of residuals of dummy variable regression

After getting the appropriate coefficients to our model, our final model is:

$$\begin{aligned}\log(y) =& 7.409 + 1.91 \times 10^{-2} \times time - 9.09 \times 10^{-5} \times timesq + 4.091 \times 10^{-7} \times timecb - 9.07 \times 10^{-2} \times tr \\ &+ 8.87 \times 10^{-2} \times D1 - 4.55 \times 10^{-2} \times D2 - 8.52 \times 10^{-2} \times D31.40 \times 10^{-2} \times D4 \\ &+ 8.44 \times 10^{-2} \times D5 + 1.65 \times 10^{-1} \times D6 + 2.27 \times 10^{-1} \times D7 + 2.89 \times 10^{-1} \times D8 \\ &+ 2.92 \times 10^{-1} \times D9 - 3.05 \times 10^{-1} \times D10 + 1.61 \times 10^{-1} \times D11 \\ &- 0.18\varepsilon_{t-1} + 0.13\varepsilon_{t-2} + 0.15\varepsilon_{t-3} + 0.540\omega_{t-1}\end{aligned}$$

We fitted this model to our testing data set and calculated a root mean square error of 3316.699. From the time plot in Figure 33 below, we can see that the model captured the general shape, but there are obvious misses, namely that the seasonal components are not very well captured.

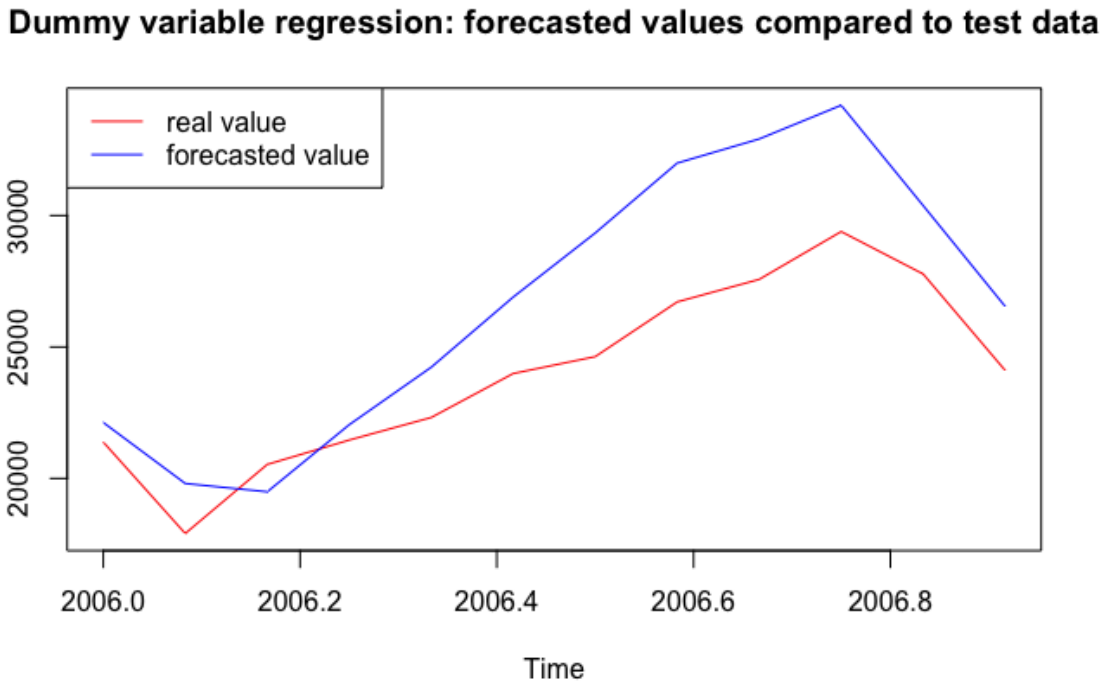

Figure 33: Forecasted values compared to testing data for dummy variable regression model

4.3 Comparing the Two Regressions

Prior to formulating our final model, we tried our multivariate model with just the explanatory variables. We found the resulting model not informative and it did not capture the trend we wish it to capture. Thus, we included dummy variables that indicate time in our causal model fit. We have a RMSE of 1435.621 for our causal model fit. Comparatively, our final model for dummy variables has a RMSE of 3316.699 and we see that our causal fit model is significantly better. However, it is important to understand that without the dummy variables in the causal model, the explanatory variables are not able to perform well on their own. We believe the best approach is to combine the two to maximize results.

Both type of regression are very simple in nature, similar to that of a multivariable regression for any other type of data. However, when working with a time series it's important to understand that time is a very crucial component, and ignoring the time component in your regression for a time series could yield unfavorable results. Dummy variables might be able to capture some of those time components, but could also miss others. Causal model fit is also easy to understand, but can also fail to factor in the crucial time component.

5 Exponential Smoothing

In this section, we will select the appropriate type of exponential smoothing to forecast our dependent variable, U.S imports of goods by customs basis from China. Referring to the time plot in Figure 1, we an observe an upward trend and variance that is increasing proportionally to the trend. The seasonal effect appears to increase with the trend, which suggests that a 'multiplicative' seasonal

component be used in the Holt-Winters procedure. Hence, we decided that the multiplicative seasonal exponential smoothing model would be the most appropriate to forecast our dependent variable.

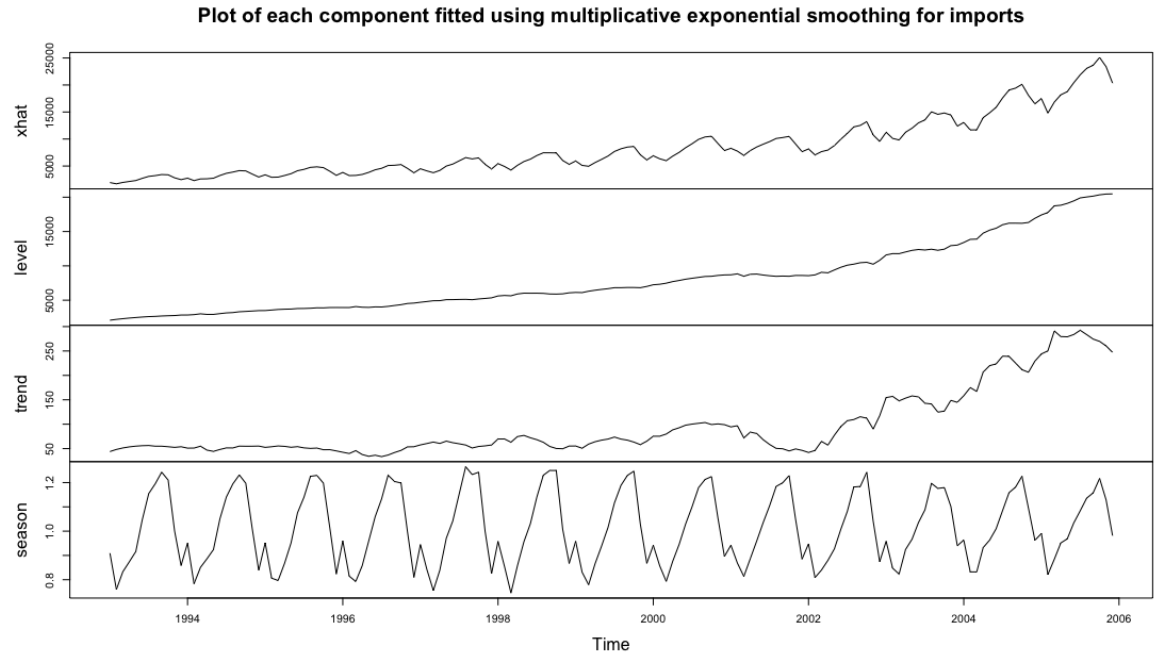

Figure 34: U.S. imports of goods by customs basis from China: fiited values; level; slope(labelled trend); seasonal variation

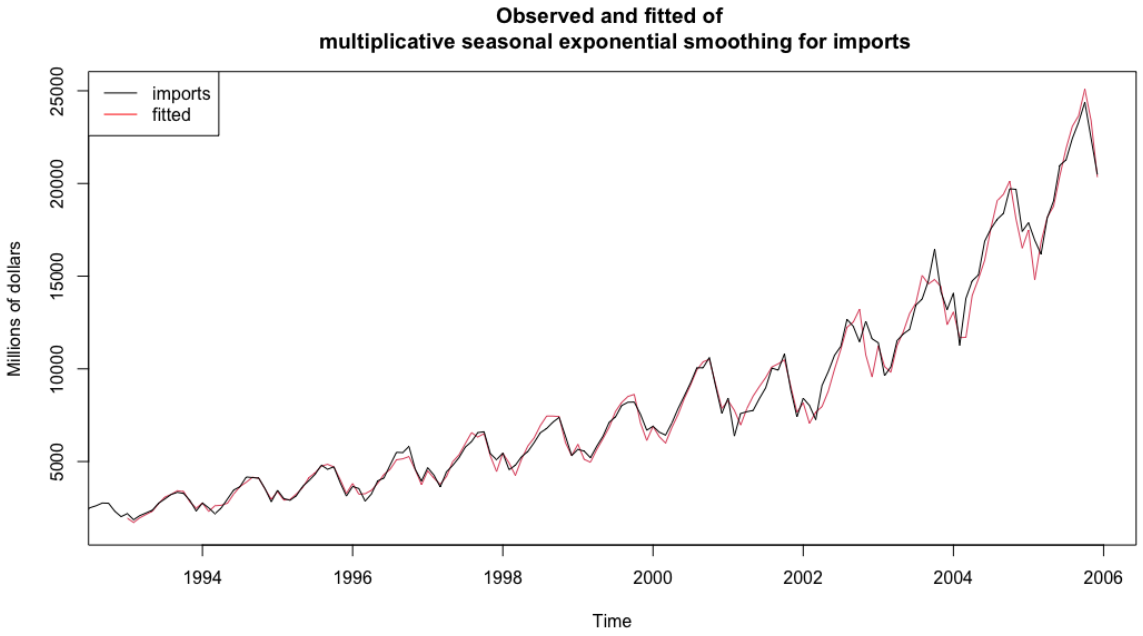

Figure 35: U.S. imports of goods by customs basis from China and Holt-Winters fitted values

Referring to Figure 35, we observe that our Holts-Winters algorithm with multiplicative seasonals is a good choice as we can observe that the training imports data and the fitted values are a close

match. Now, we will forecast 12 months ahead for the U.S. imports of goods by customs basis from China.

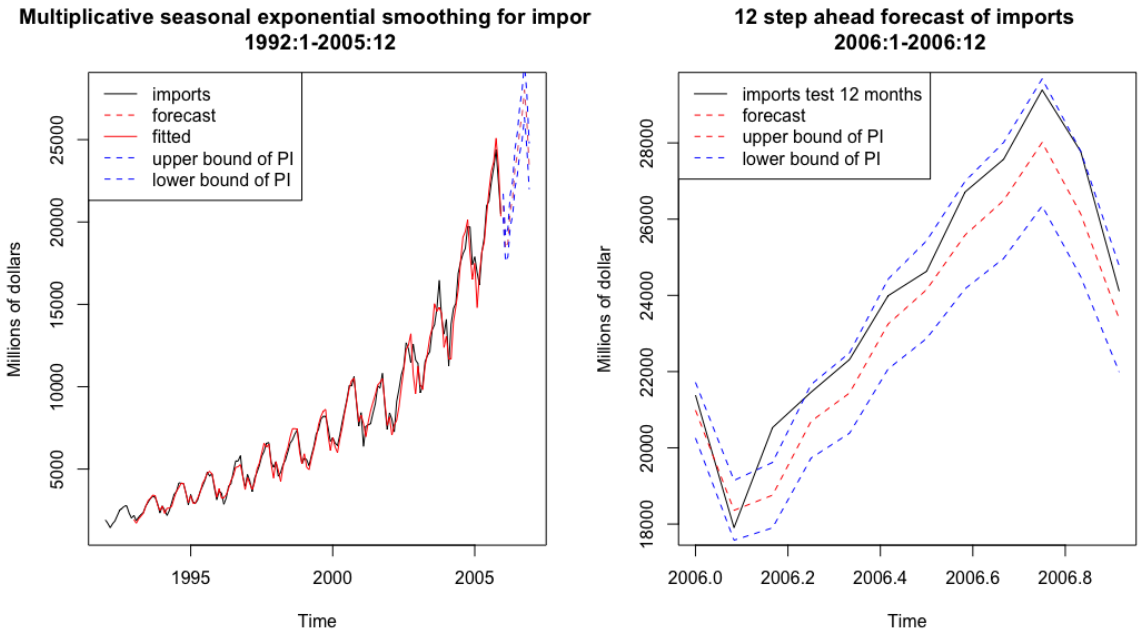

Figure 36: Test data, Holt-Winters forecasts, prediction intervals for imports for Jan 2006 to Dec 2006 shown

The estimates of the model parameters obtained are $\hat{\alpha} = 0.2761983$, $\hat{\beta} = 0.05683906$ and, $\hat{\gamma} = 0.5089468$. Rounded to 3 decimal place, we obtain the Holt-Winters algorithm with multiplicative seasonals is

$$a_t = 0.276(\frac{y_t}{s_{t-12}}) + (1 - 0.276)(a_{t-1} + b_{t-1})$$
$$b_t = 0.057(a_t - a_{t-1}) + (1 - 0.057)b_{t-1}$$
$$s_t = 0.509(\frac{y_t}{a_t}) + (1 - 0.509)s_{t-12}$$

The forecasting equation for y_{t+h} made after the observation at time t becomes

$$\hat{y}_{t+h} = (a_t + hb_t)s_{t+h-12}, \quad h \leq 12$$

Referring to Figure 36, we can see that our forecasts are not perfect but relatively close to the test data. We can observe that it is able to predict the test data of Jan 2006 the best. We measured the accuracy of the forecasts using the test root mean square (RMSE) statistic and obtained a value of 3039.351. It should be noted that the forecasts are based on trends in the period of the training data set. It is possible that unforeseen circumstances could lead to a completely different future values. Furthermore, as we make forecasts even further in the future, the accuracy of the forecast decreases.

6 Forecast Comparison

Below is a table created to compare different model forecasts and their RMSEs, along with a column for the average among the different forecasts:

Date	Raw Data Values	ARIMA Modeling Forecast	Multiple Regression: Causal Model	Multiple Regression: Dummies and Polynomials Only	Exponential Smoothing	Average Forecast
Jan 2006	21382.51	21498.92	21130.56	22126.86	20987.31	21435.91
Feb 2006	17905.35	18235.6	19877.71	19806.82	18360.35	19070.12
Mar 2006	20531.26	19755.91	19123.34	19490.76	18758.09	19282.02
Apr 2006	21459.07	22087.6	20125.38	22042.84	20695.84	21237.92
May 2006	22317.64	22838.74	21561.6	24228.67	21438.42	22516.86
Jun 2006	23989.7	24771.11	22848.14	26898.06	23247.51	24441.2
Jul 2006	24632.01	26121.29	23887.33	29334.69	24150.1	25873.35
Aug 2006	26713.34	27354.06	25520.4	31993.23	25584.88	27613.1
Sep 2006	27570.62	28122.28	26533.73	32912.18	26490.98	28514.79
Oct 2006	29388.60	29856.26	27192.17	34192.86	28014.05	29813.83
Nov 2006	27775.08	28357.58	25458.68	30377.25	26138.62	27583.03
Dec 2006	24109.19	25683.18	22730.81	26543.38	23379.59	24584.24
	RMSE:	**814.93**	**1435.62**	**3316.70**	**1048.20**	**758.18**

Table 4: Table of Forecast and RMSE using different models, and Average Forecast for each date

Using RMSE as our metric of comparison, we see that the average forecast performs the best among all of the forecasts we examined. The second-best forecast according to RMSE was the ARIMA modeling forecast. In the middle were exponential smoothing and multiple regression with the causal model, and the worst-performing forecast was multiple regression with dummy variables and polynomials only. From Section 4.3, we suspect that the two multiple regression models could be prone to not capturing enough of the crucial time component. For this particular time series, the flexibility provided by ARIMA modeling was more effective than exponential smoothing; perhaps if the seasonal effects were more regular, the seasonal exponential smoothing could have outperformed ARIMA. Ultimately, an average forecast appeared to combine the strengths of the various forecasting methods, yielding the lowest RMSE.

References

[Lab21] U.S. Bureau of Labor Statistics. *Table B-1. Employees on nonfarm payrolls by industry sector and selected industry detail*. 2021. URL: https://www.bls.gov/news.release/empsit.t17.htm.

[BE22] U.S. Census Bureau and U.S. Bureau of Economic Analysis. *U.S. Imports of Goods by Customs Basis from China [IMPCH]*. 2022. URL: https://fred.stlouisfed.org/series/IMPCH.

[Gov22] Board of Governors of the Federal Reserve System (US). *Industrial Production: Non-Durable Nonenergy Consumer Goods*. 2022. URL: https://fred.stlouisfed.org/series/IPB51210N.

[Lab22] U.S. Bureau of Labor Statistics. *Average Hourly Earnings of Production and Nonsupervisory Employees, Total Private [CEU0500000008]*. 2022. URL: https://fred.stlouisfed.org/series/CEU0500000008.

Acknowledgments

We consulted Professor Sanchez's Stat 170 class lectures throughout this assignment. To access our variables in R, we used the Quandl API as covered in Week 1 lectures "The Quandl API in R. Demo" and "Using Quandl inside RStudio." To analyze their trend and seasonality using time plots and seasonal box plots, we used concepts from from Week 1 lecture "Introduction to Time Series Part 2. Cleaning. Main features." To generate the plots themselves, we referred to the videos and

accompanying R script for Week 1 lectures "Using Quandl inside RStudio" and "1/4/2021 Video of the discussion during 9:30-10:45 AM."

To select and perform multiplicative decomposition, we referred to the Week 2 lecture "Classical decomposition, part 2. Multiplicative decomposition." For analyzing the ACF of the random component after multiplicative decomposition, we recalled examples from Week 2 lectures "The sample ACF (aka correlogram) of a time series. Interpretation, and conditions under which it makes sense to interpret it" and "Video of the discussion during class time, 1/11/2022, 9:30-10:45." For analyzing the PACF of the random component, we referred to Week 3 lectures "The partial autocorrelation function" and "The theory of the AR stochastic process."

To compare pre-differencing transformations and perform differencing, we applied concepts from Week 3 lecture "ARIMA part 1. First steps of ARIMA modeling. Pretransforming to stabilize variance and differencing to account for trend and seasonality." We also referred to the video and accompanying R script for "1/20/2022 - discussion that took place during 1/20/2022 class time (9:30-10:45)." To analyze the resulting ACF and PACF plots, we referred to Week 3 lectures "The theory of stationary stochastic processes" and "The partial autocorrelation function."

To identify an ARIMA modeling, fit it and model it, we consulted week 4 lecture "ARIMA modeling and forecasting Part 1" and "ARIMA modeling and forecasting Part 2". To diagnose the fit of it and also forecast it, we consulted week 5 lectures "ARIMA modeling and forecasting Part 6".

To model a time series with a causal model fit as well as fitting the residuals, we consulted week 5 lectures "Regression with autocorrelated errors. Generalized lest squares". We consulted the same lecture to model our times series with a dummy variable fit.

To apply exponential smoothing on our time series, we took ideas from week 6 lectures "Simple Exponential smoothing", "Trend-corrected(two parameters), nonseasonal Exponential smoothing", and "Seasonal+trend exponential smoothing".