Analyzing California and New York State Exports with ARIMA and VAR Models

Johan Chua, Nicholas Delianedis, Grace Pham

12/19/23

Contents

I Introduction

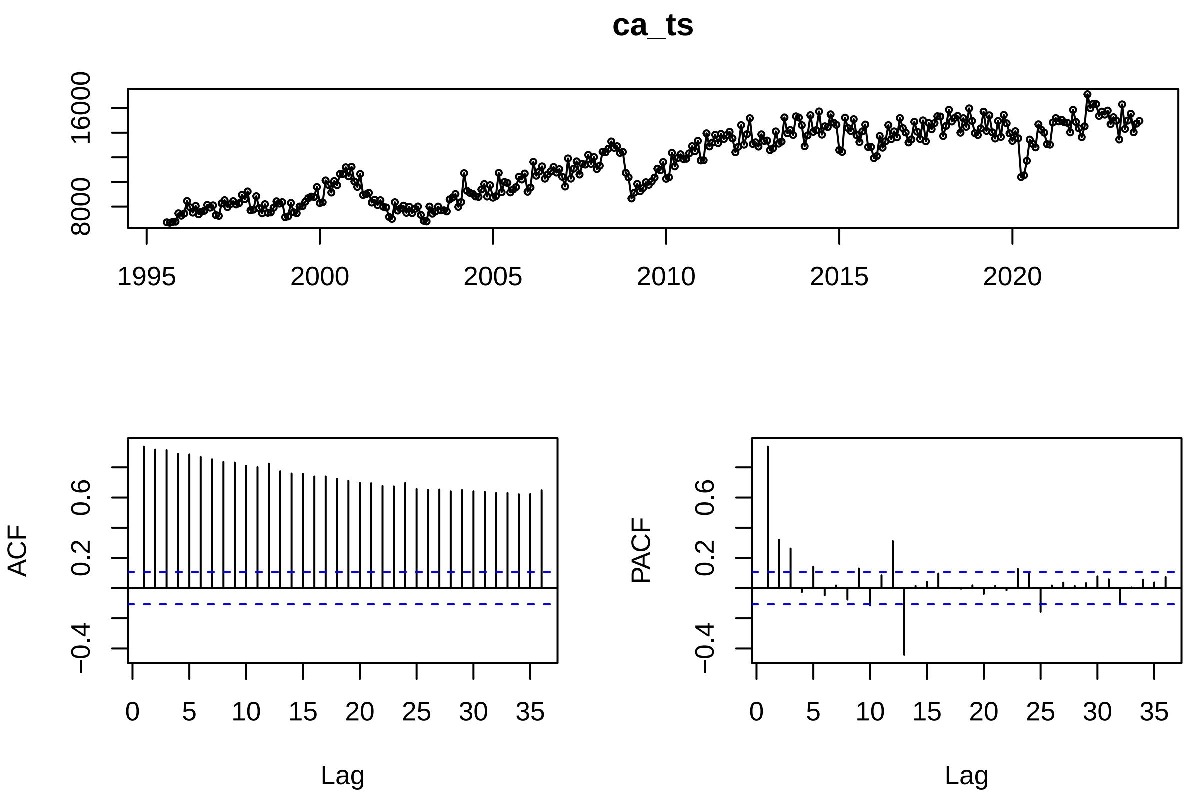
In this project, we model and forecast the exports of goods for New York and California using ARIMA in addition to analyzing causality with VAR. These two datasets contain monthly data, from August 1995 through September 2023, on the exports of manufactured and non-manufactured commodities based on origin of movement in millions of dollars. Analyzing and measuring the export levels of each state is significant as exports help facilitate international trade and stimulate economic activity. Since California and New York are two of the top three states with the highest GDP, we thought it would be beneficial to investigate these states' data in greater detail. Forecasting the exports will be beneficial in examining the future economic climates and help in creating meaningful economic policies in each state.

II Results

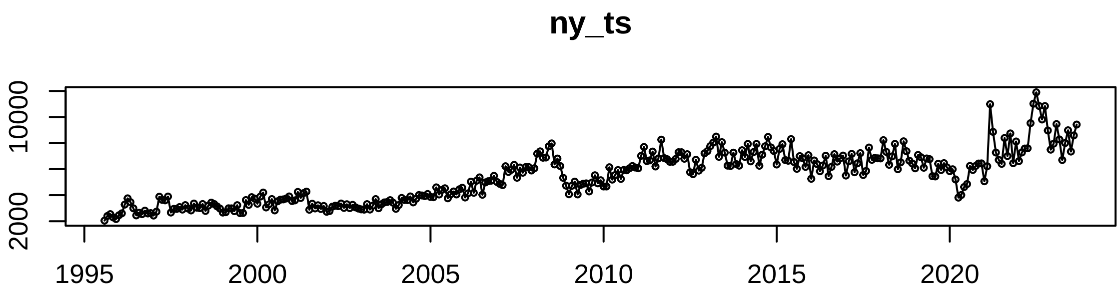
```
ca <- read.csv("CAEX.csv") # load in CA data
ny <- read.csv("NYEX.csv") # load in NY data
ca_ts <- ts(ca[,2], start = c(1995, 8), frequency = 12) # convert CA to time series
ny_ts <- ts(ny[,2], start = c(1995, 8), frequency = 12) # convert NY to time series
```

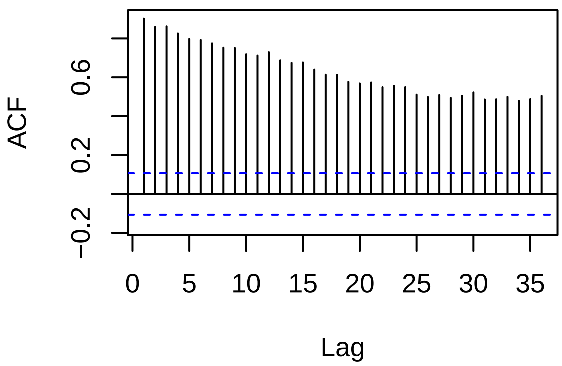
(a) Time-Series Plot, ACF and PACF

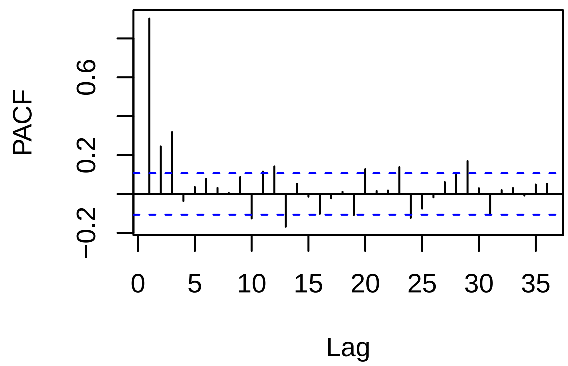
```
tsdisplay(ca_ts) # Plot/ACF/PACF of CA Exports
```




```
tsdisplay(ny_ts) # Plot/ACF/PACF of NY Exports
```

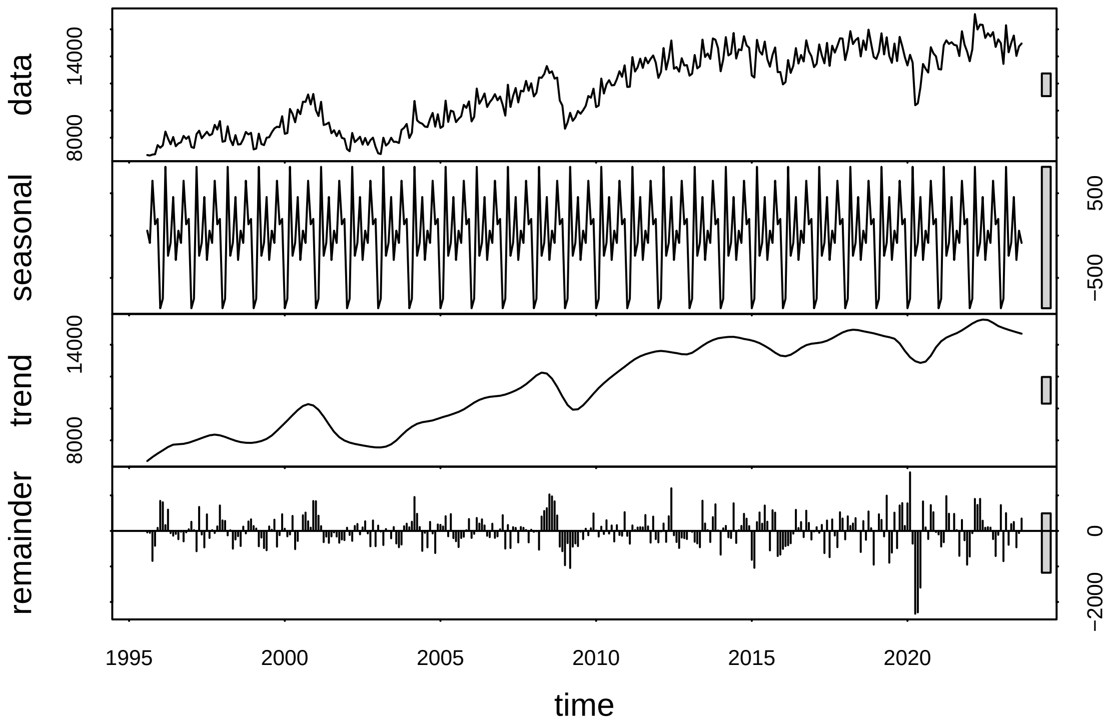




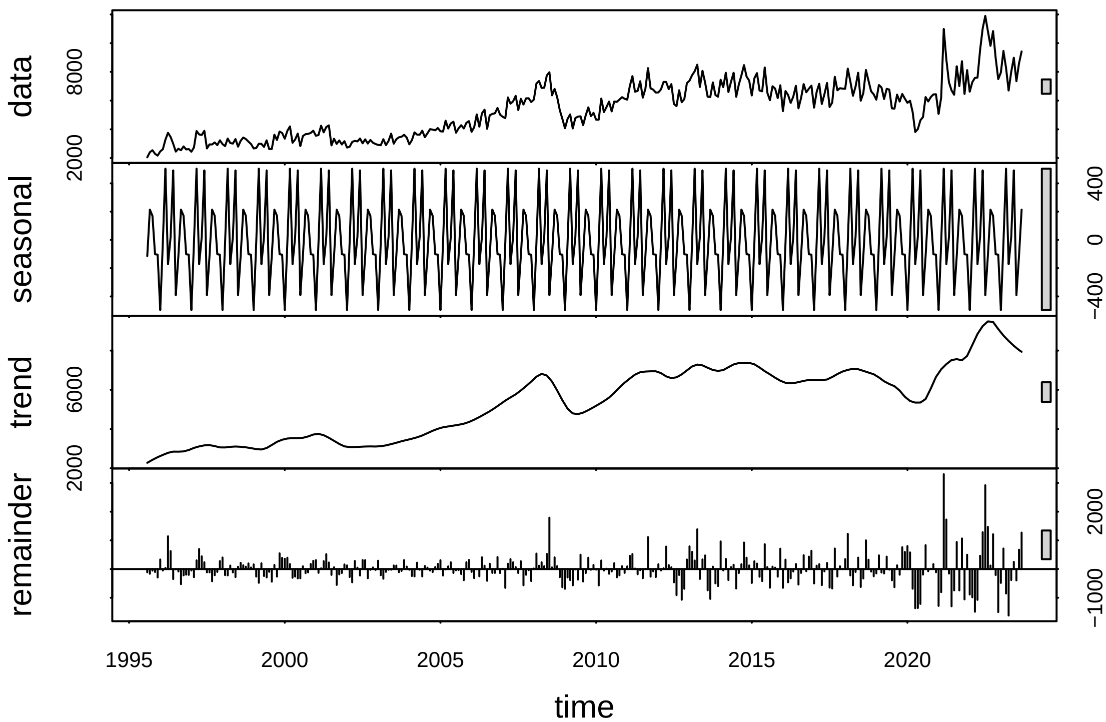


(b) STL Decomposition Plots

```
plot(stl(ca_ts, s.window="periodic")) # stl decomposition plot of CA
```



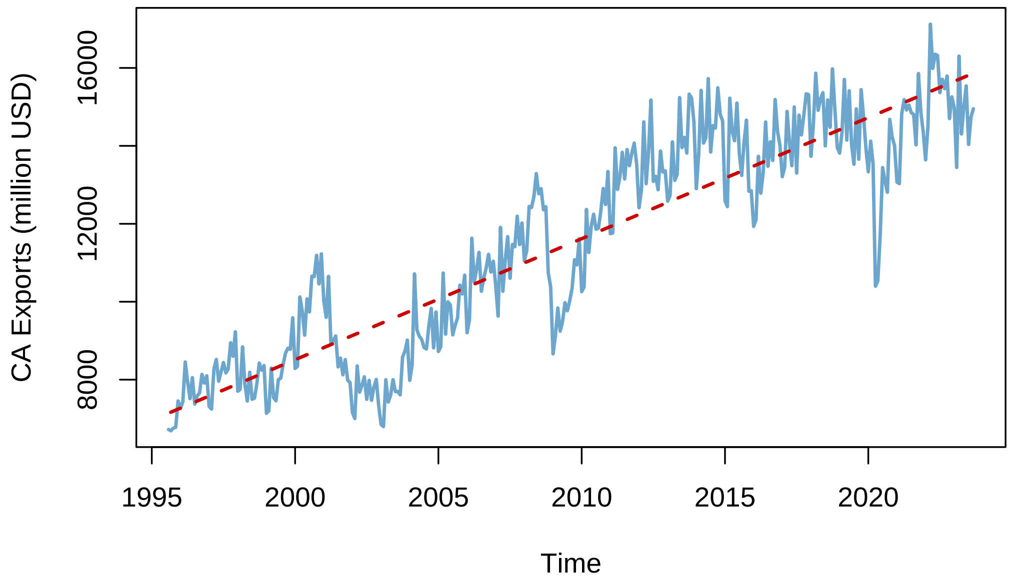
```r
plot(stl(ny_ts, s.window="periodic")) # stl decomposition plot of NY
```



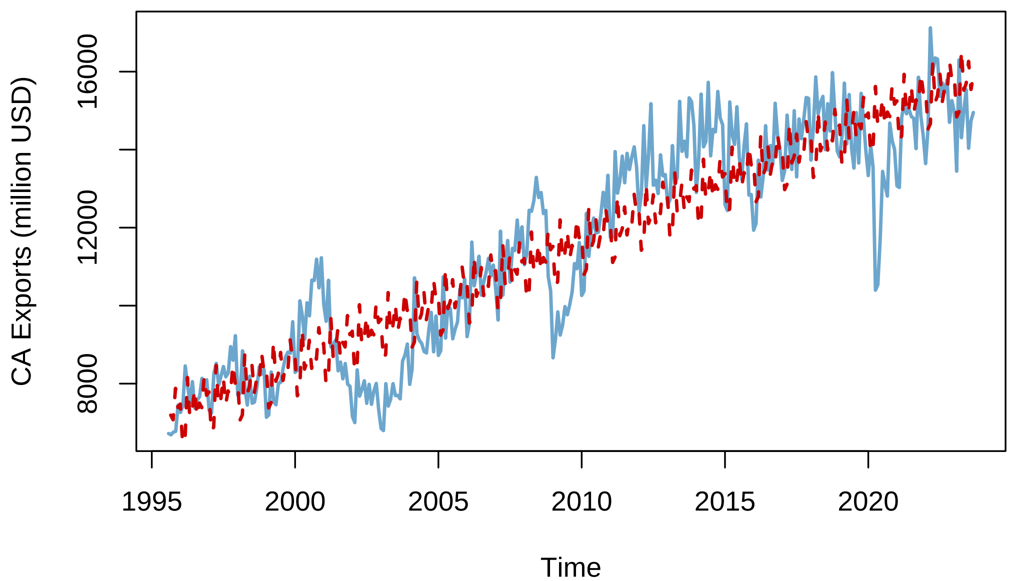
For these decompositions, we see that there is a very significant upward trend in both, as shown by the graphs here and their small confidence interval bar, showing that they have real overall movement. In both plots, we see a strong seasonal component, that, while existing mostly within the confidence interval, seems to explain some components of both datasets. Finally, in the remainder plots, we can see that, in the California plot, most of the remainder is noise, with only a couple strong outliers, most notably near 2008 and 2020, while in the New York plot, we see many more bars outside of the confidence interval band, especially after 2020.

(c) Model with Trend, Seasonality and Cycles

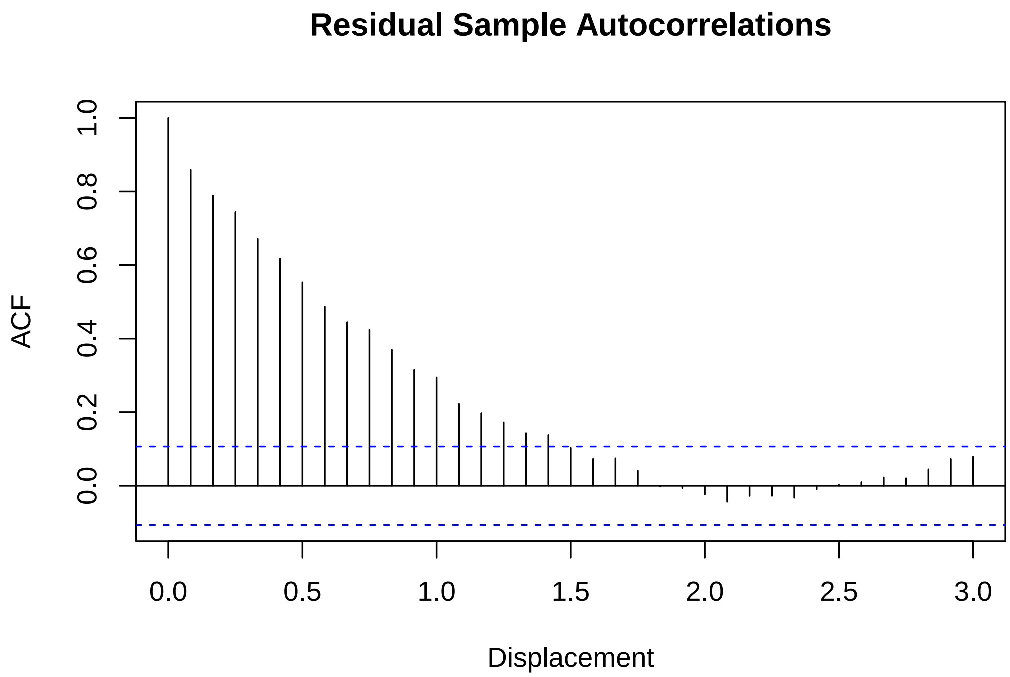
```r
# time dummy variable
t <- seq(1995.66, 2023.75, length = length(ca_ts))

# CA trend
m1_ca <- lm(ca_ts ~ t)
plot(ca_ts, ylab = "CA Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, m1_ca$fit, col = "red3", lwd = 2, lty = 2)
```



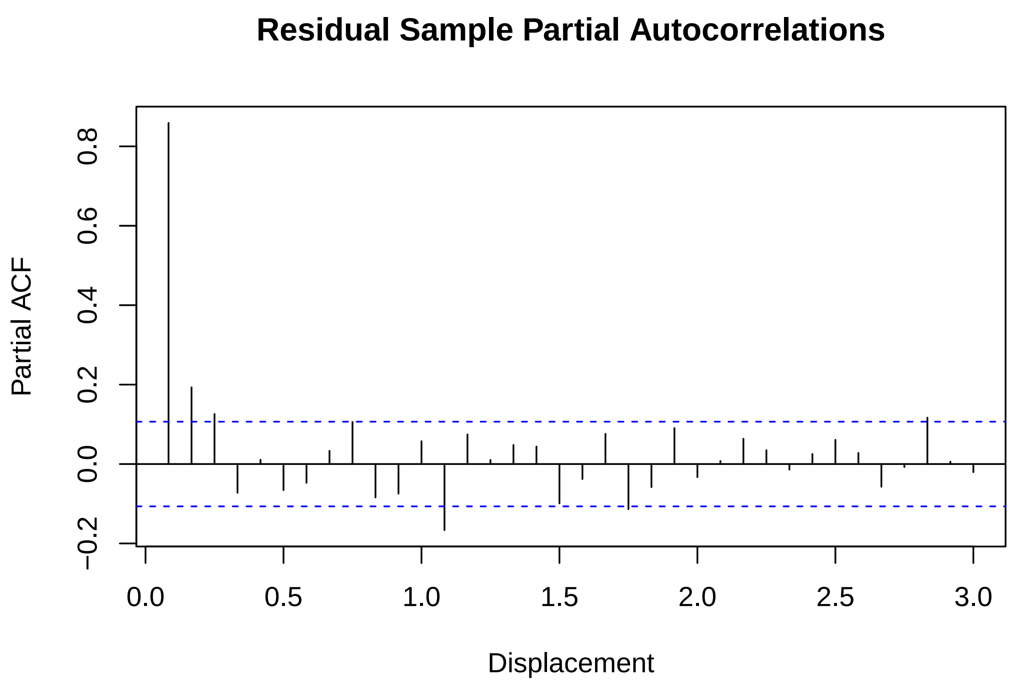

```
# CA trend + seasonality
m2_ca <- tslm(ca_ts ~ trend + season)
plot(ca_ts, ylab = "CA Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, m2_ca$fit, col="red3", lwd = 2, lty = 2)
```



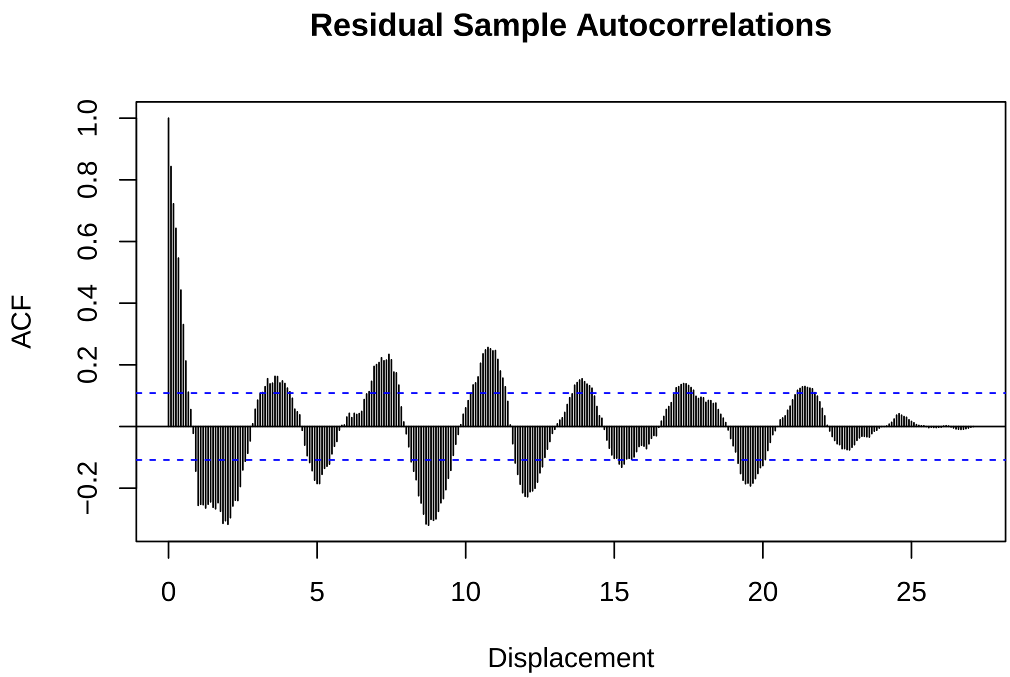

```
# CA trend + seasonality + cycles
acf(m2_ca$res,lag=36,main="Residual Sample Autocorrelations",xlab="Displacement")
```



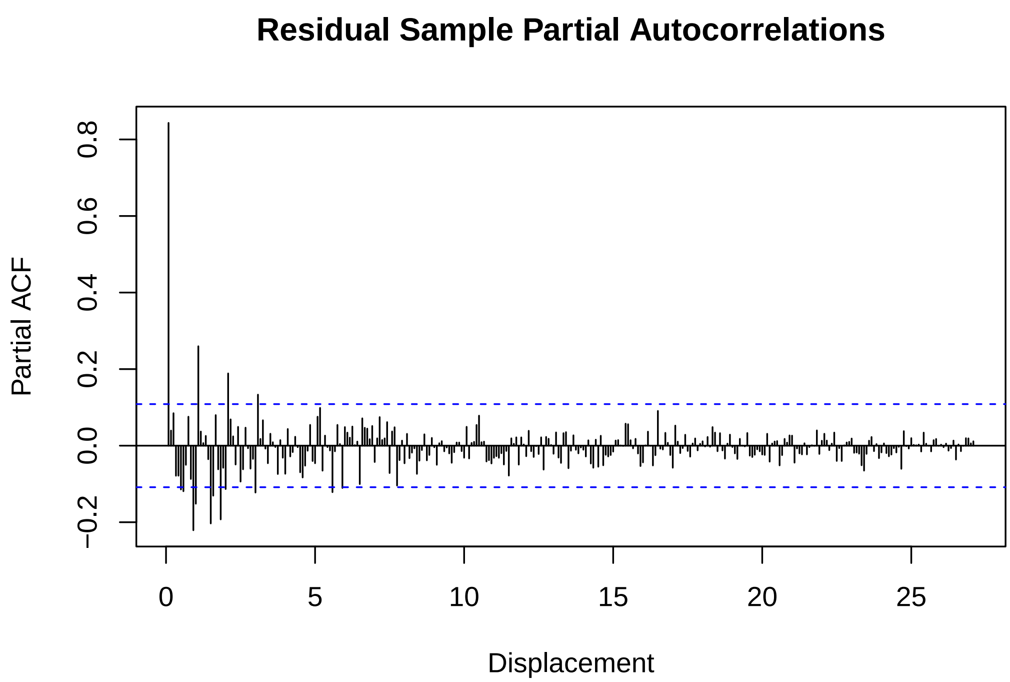

```
pacf(m2_ca$res,lag=36,main="Residual Sample Partial Autocorrelations", xlab="Displacement")
```



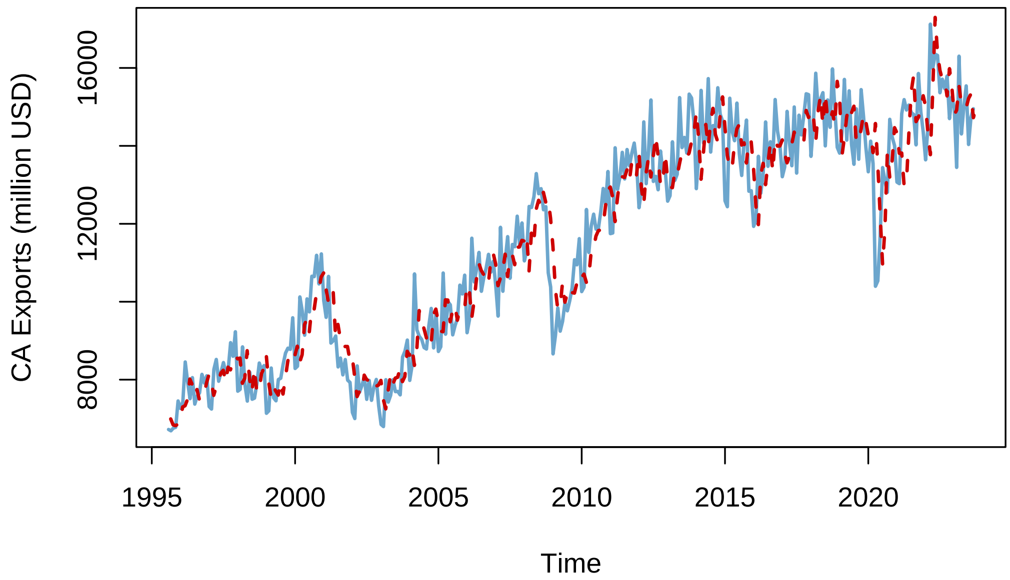

```
## check for seasonal ARMA
acf(diff(ca_ts,12),lag=360,main="Residual Sample Autocorrelations",xlab="Displacement")
```



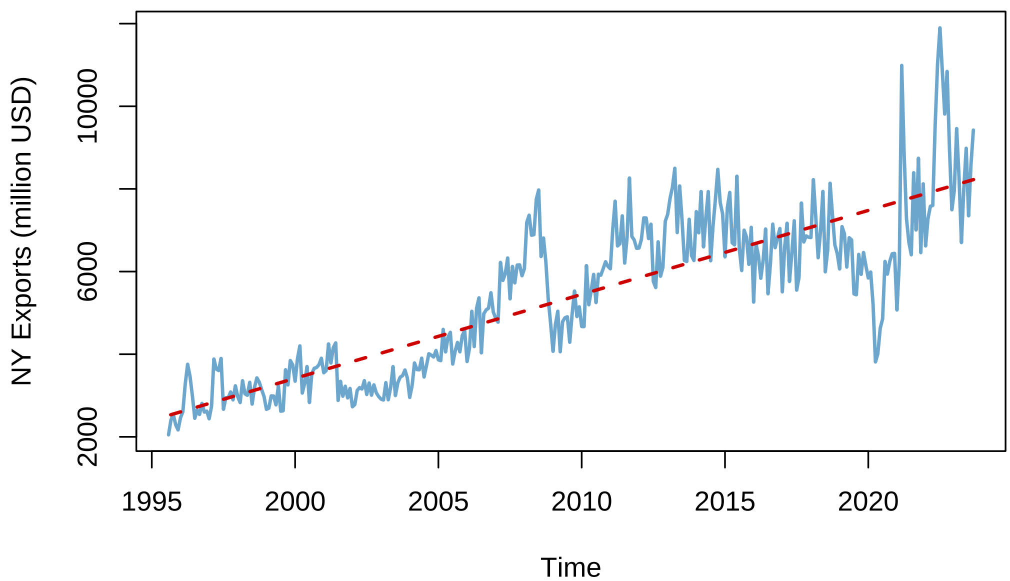

```
pacf(diff(ca_ts,12),lag=360,main="Residual Sample Partial Autocorrelations", xlab="Displacement")
```



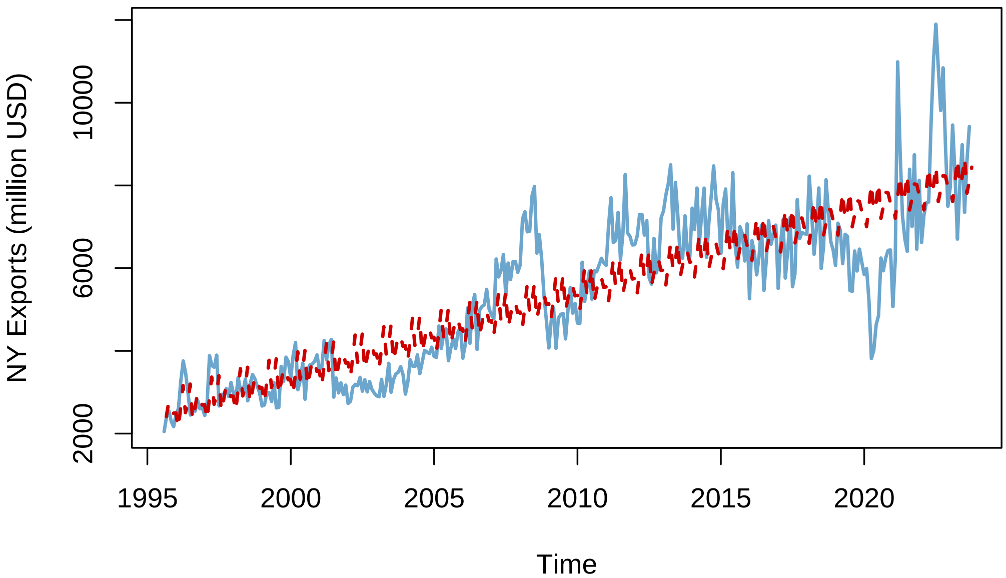

```
## final model
m3_ca <- arima(ca_ts, order = c(2,0,0), xreg = c(t), season = list(order = c(0,0,1)))
plot(ca_ts, ylab = "CA Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, fitted(m3_ca), col="red3", lwd = 2, lty = 2)
```



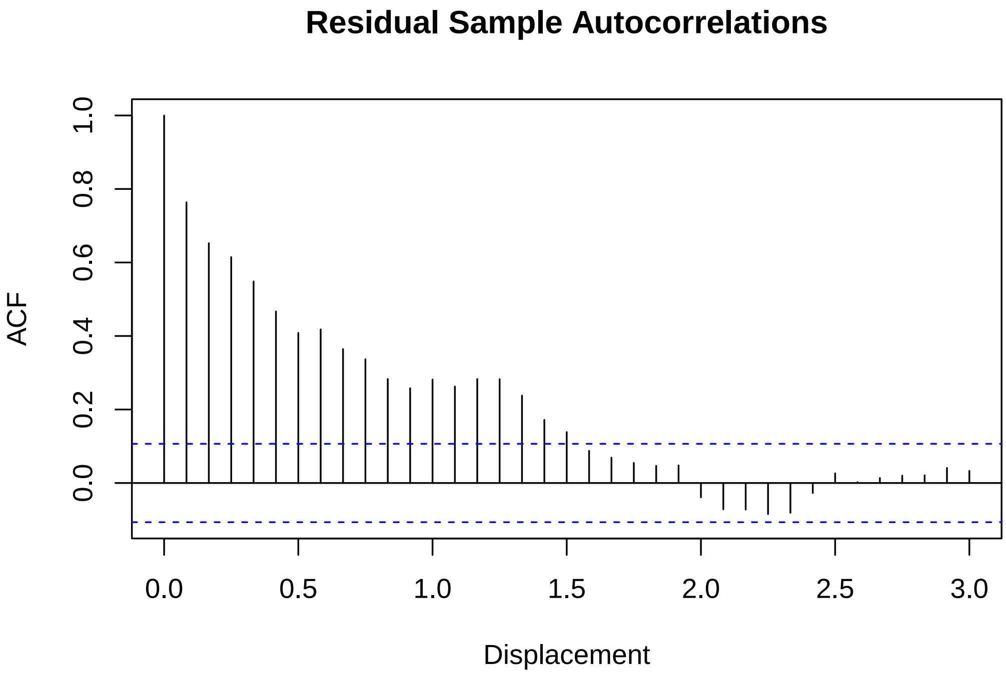

```
# NY trend
m1_ny <- lm(ny_ts ~ t)
plot(ny_ts, ylab = "NY Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, m1_ny$fit, col = "red3", lwd = 2, lty = 2)
```



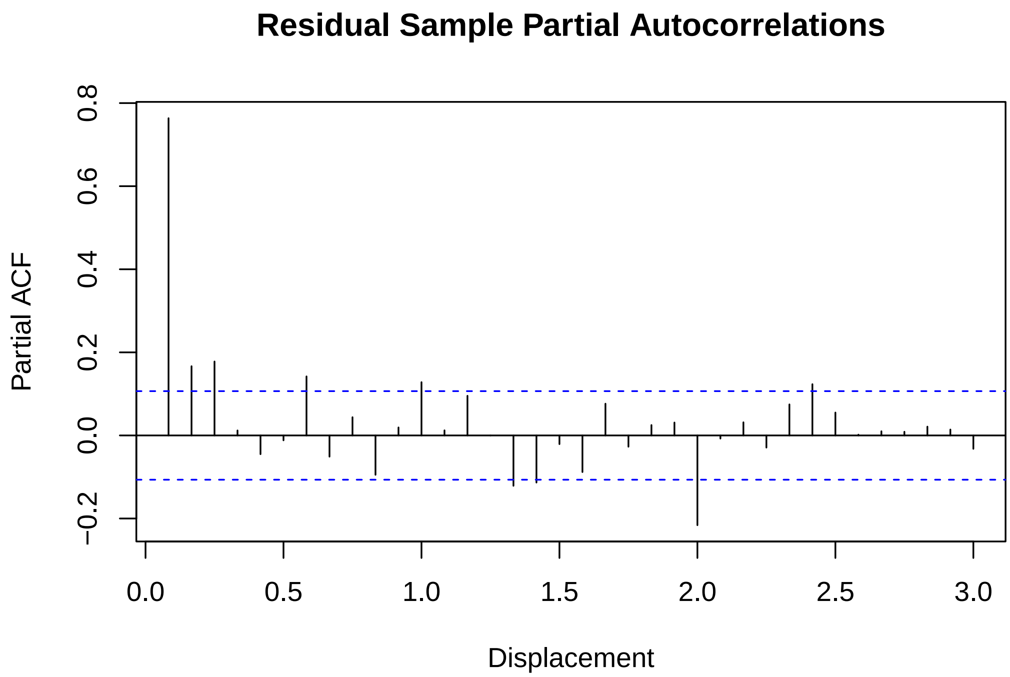

```
# NY trend + seasonality
m2_ny <- tslm(ny_ts ~ trend + season)
plot(ny_ts, ylab = "NY Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, m2_ny$fit, col="red3", lwd = 2, lty = 2)
```



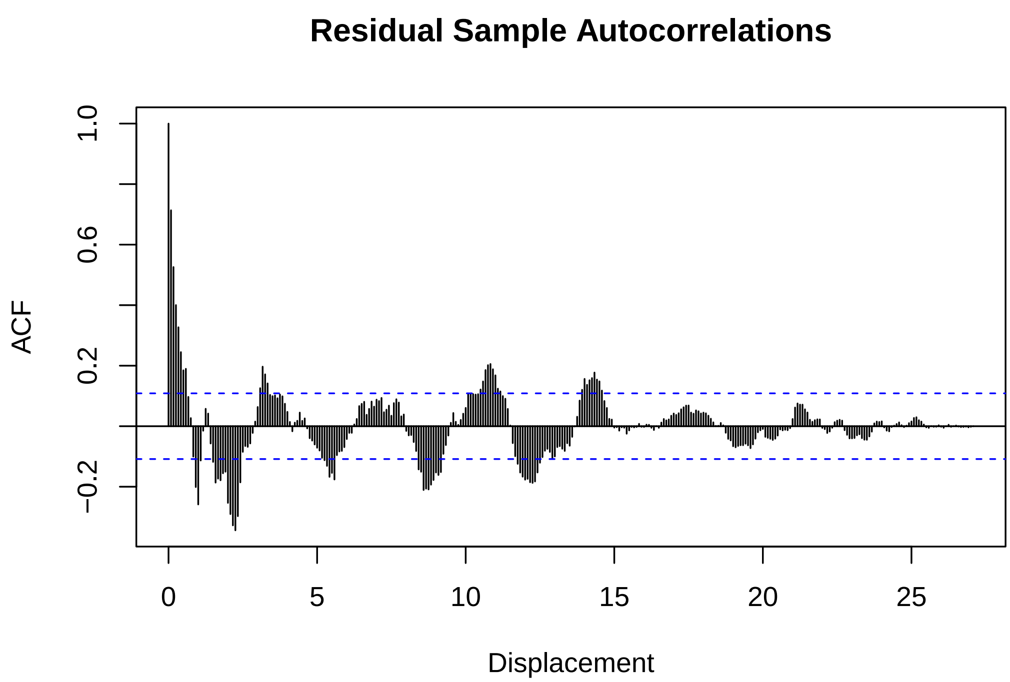

```r
# NY trend + seasonality + cycles
acf(m2_ny$res,lag=36,main="Residual Sample Autocorrelations",xlab="Displacement")
```



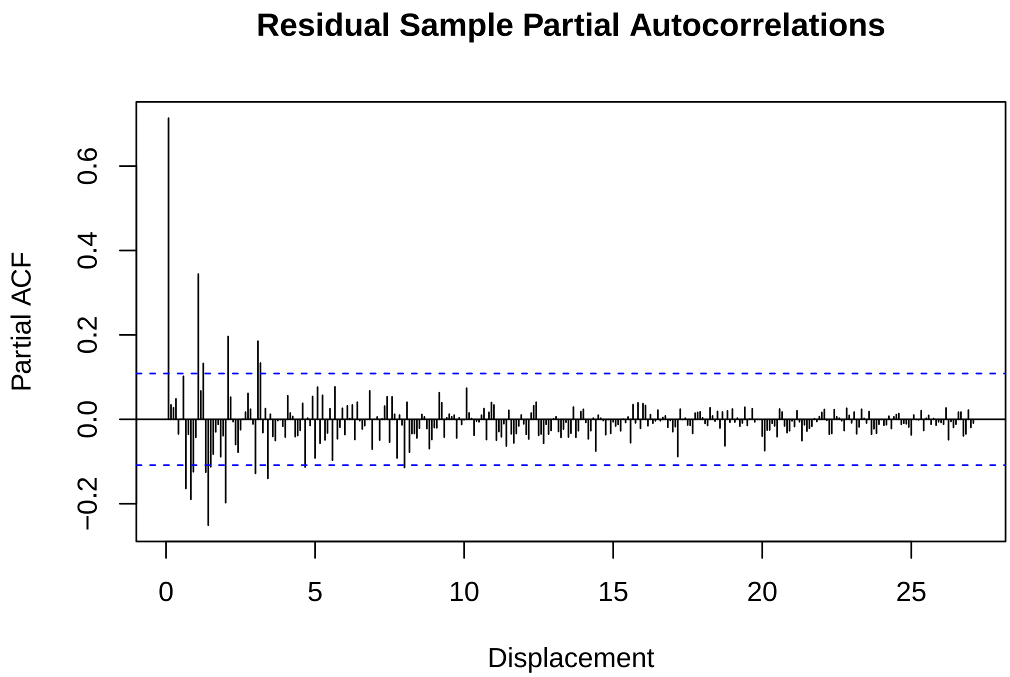

```r
pacf(m2_ny$res,lag=36,main="Residual Sample Partial Autocorrelations", xlab="Displacement")
```



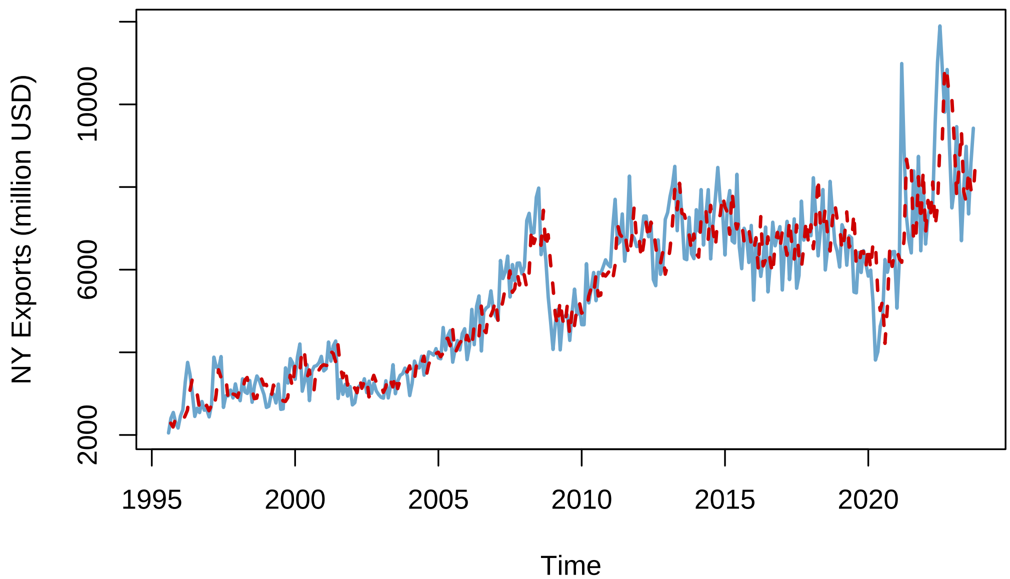

```r
## check for seasonal ARMA
acf(diff(ny_ts,12),lag=360,main="Residual Sample Autocorrelations",xlab="Displacement")
```




```r
pacf(diff(ny_ts,12),lag=360,main="Residual Sample Partial Autocorrelations", xlab="Displacement")
```



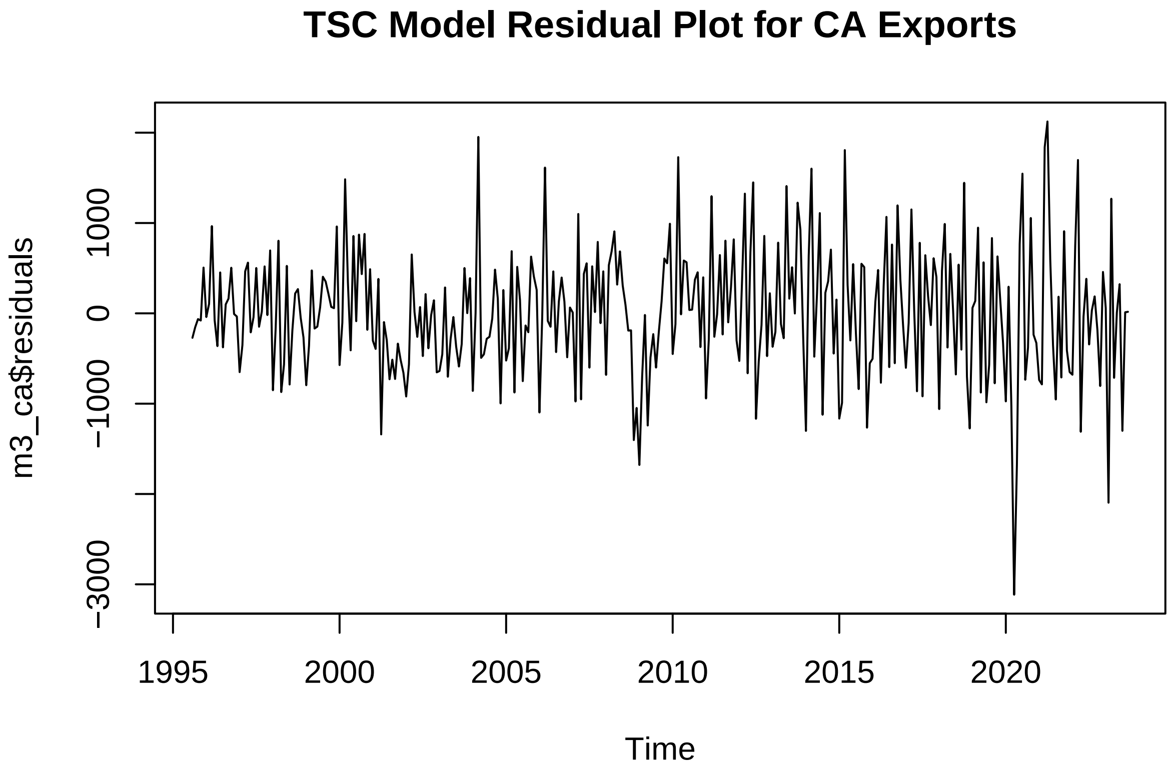

```
## final model
m3_ny <- arima(ny_ts, order = c(3,0,0), xreg = c(t), season = list(order = c(1,0,1)))
plot(ny_ts, ylab = "NY Exports (million USD)", xlab = "Time", lwd = 2, col = 'skyblue3')
lines(t, fitted(m3_ny), col="red3", lwd = 2, lty = 2)
```



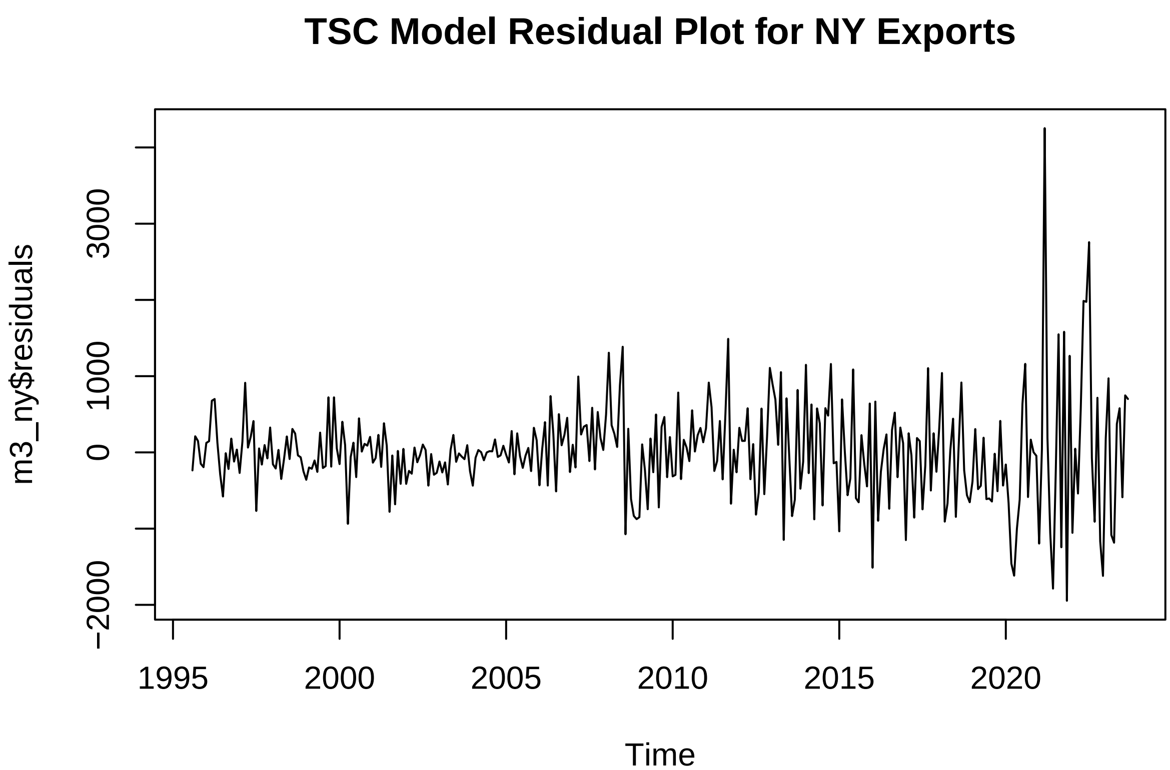
To start our model, we used a linear trend on the data, since the data seemed to most closely follow a linear pattern. Then, after looking at the ACF and PACF graphs for both detrended datasets, we found that the California dataset exhibited a seasonal AR(1) pattern, while the New York dataset exhibited a seasonal ARMA(1,1) pattern after looking at lags spaced 12 apart, since this is monthly data. Finally, to model cycles, we again looked at the ACF and PACF graphs of the two datasets, and concluded that, for the California data, an AR(2) model was best suited for the data, while for the New York data, an AR(3) model was optimal.

(e) Residual vs Fitted

```r
plot(m3_ca$residuals, main = "TSC Model Residual Plot for CA Exports") # CA
```



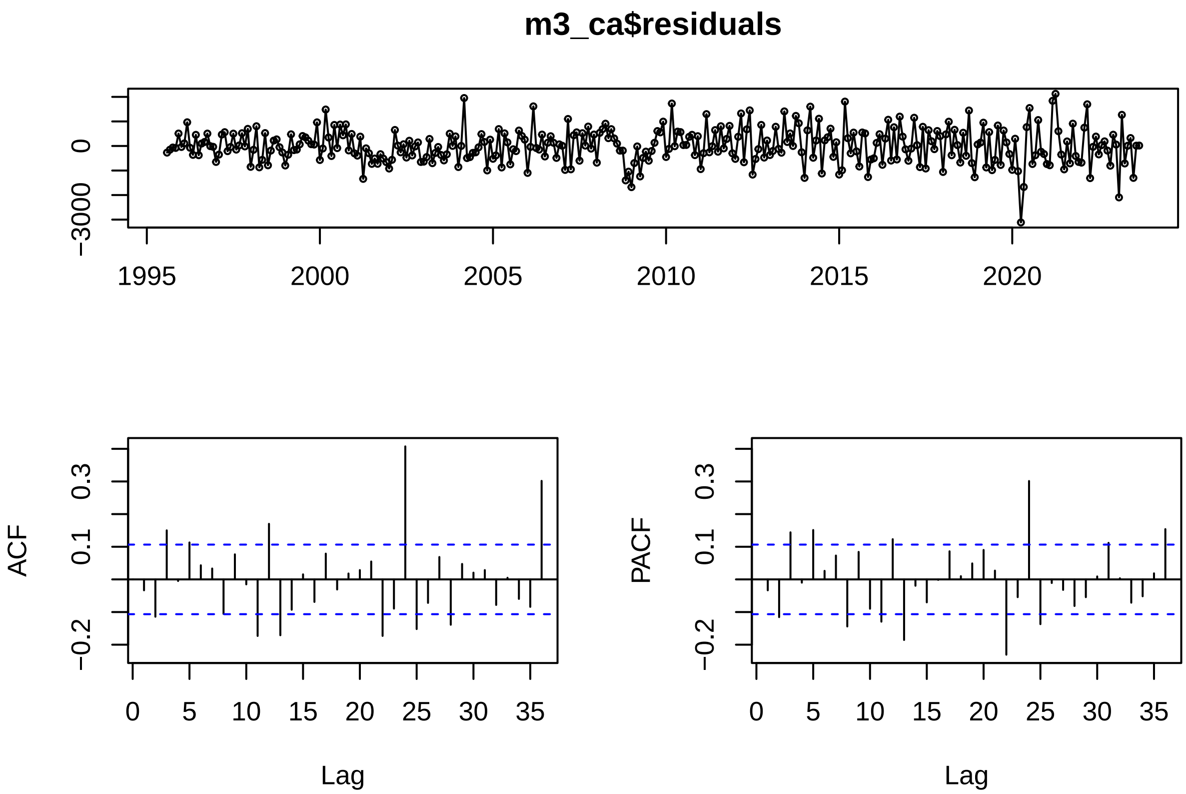

```r
plot(m3_ny$residuals, main = "TSC Model Residual Plot for NY Exports") # NY
```



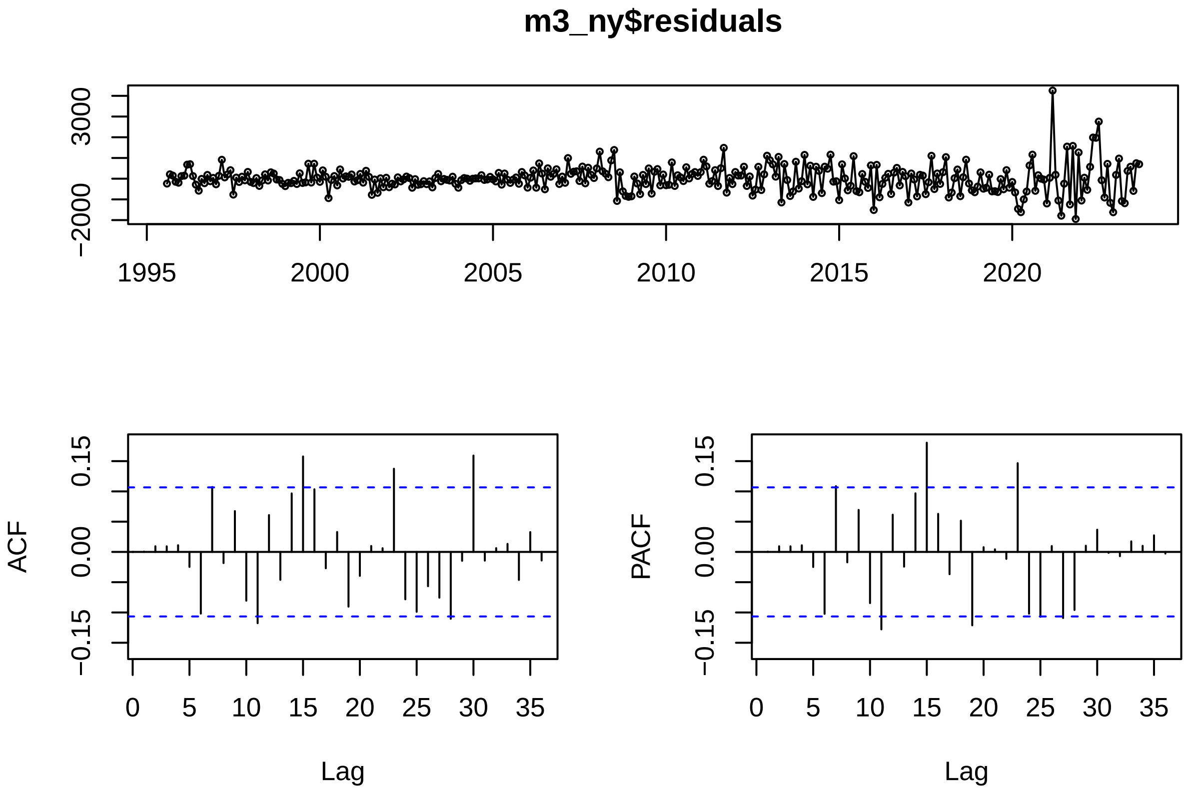
We plot the respective residuals vs. the fitted values for California and New York exports. For California, the variance is not constant and seems to slightly increase. There are a couple significant outliers very far from 0. The residual plot for New York seems to have an even greater variance and obvious increase in the errors. There are also many outliers. Since the residual plots do not resemble white noise, we can suggest further differencing or taking the log of the data to make the residuals covariance stationary.

(f) ACF and PACF of Residuals

```
tsdisplay(m3_ca$residuals) # ACF and PACF for CA
```



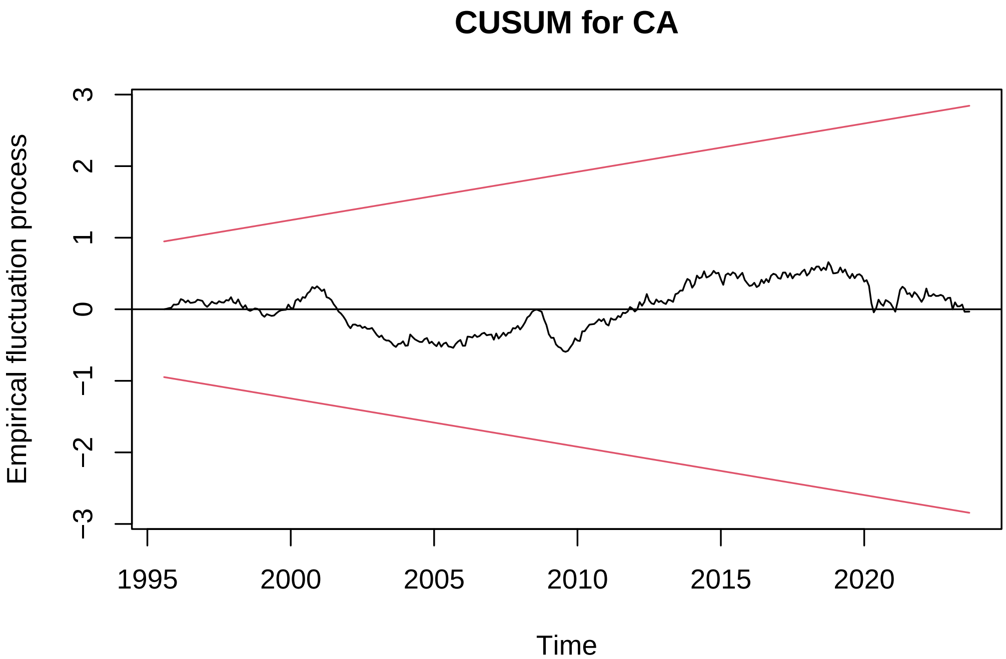

```
tsdisplay(m3_ny$residuals) # ACF and PACF for NY
```



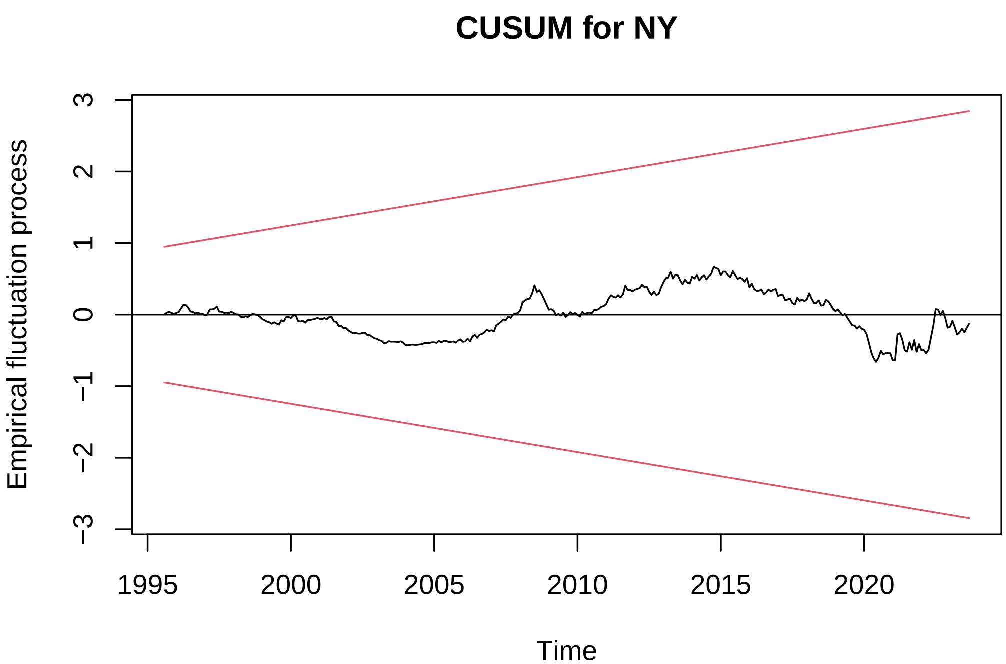
We plot the ACF and PACF of the respective residuals for California and New York. There still seems to be a few statistically significant spikes. This suggests that our final model still needs improvements. We can perform a further formal statistical test to confirm whether or not the spikes are actually significant.

(g) CUSUM

```
plot(efp(m3_ca$res~1, type = "Rec-CUSUM"), main = "CUSUM for CA") # CUSUM for CA
```




```
plot(efp(m3_ny$res~1, type = "Rec-CUSUM"), main = "CUSUM for NY") # CUSUM for NY
```



From the CUSUM plots for the models for both datasets, we can see that there are no structural breaks, as the errors do not go outside or even get near the red lines in each, so the model seems to function well at all time intervals.

(h) Diagnostic Statistics

```
MAPE(.resid = m3_ca$resid, .actual = ca_ts) # MAPE for CA
```

```
## [1] 4.800544
```

```
RMSE(.resid = m3_ca$resid, .actual = ca_ts) # RMSE for CA
```

```
## [1] 704.8028
```

```
MSE(.resid = m3_ca$resid, .actual = ca_ts) # MSE for CA
```

```
## [1] 496747
```

```
MAPE(.resid = m3_ny$resid, .actual = ny_ts) # MAPE for NY
```

```
## [1] 8.028586
```

```
RMSE(.resid = m3_ny$resid, .actual = ny_ts) # RMSE for NY
```

```
## [1] 643.5928
```

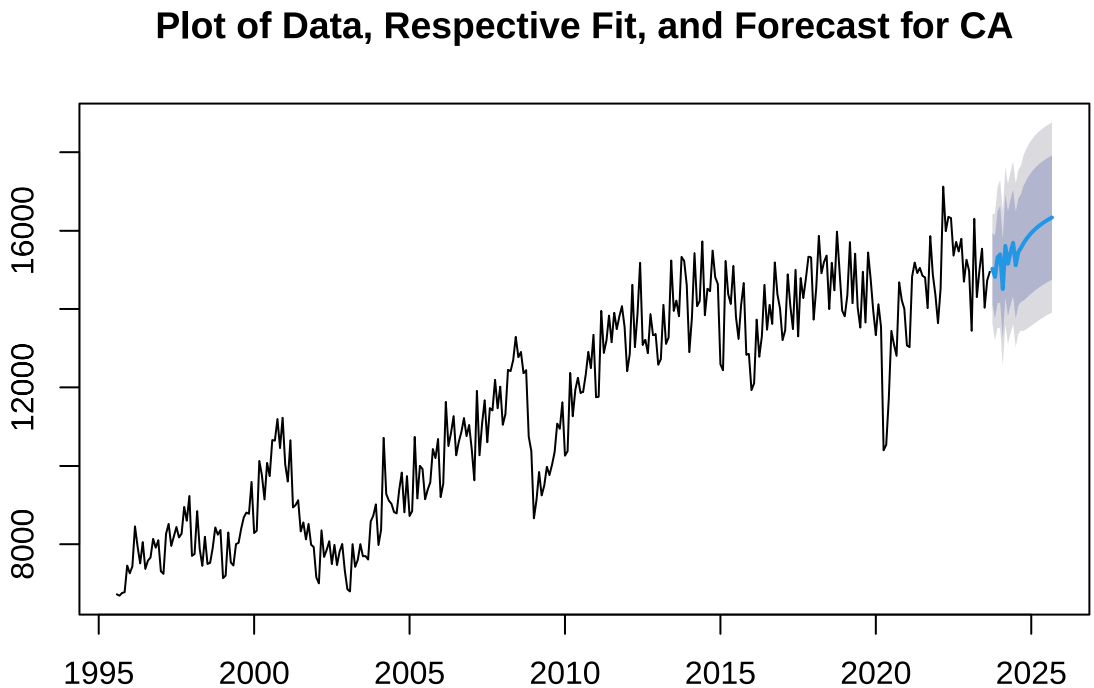
```
MSE(.resid = m3_ny$resid, .actual = ny_ts) # MSE for NY
```

```
## [1] 414211.7
```

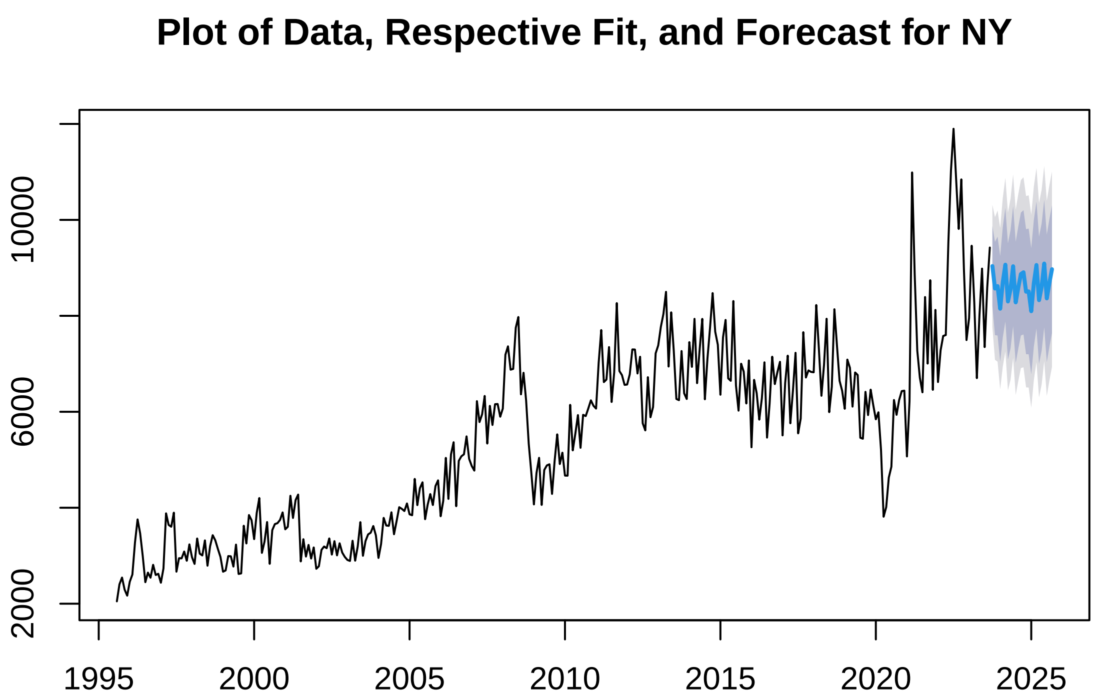
For the diagnostic statistics, we can see that the MAPE, or the mean absolute percentage error, is around 5% for the California model, and 8% for the New York model, which are decent results, but it shows that we could improve both of our models, but especially the New York model. Similarly, the RMSE, or root mean square error, is around the 600-700 range for both models, which indicates that we could improve both models to lower that value.

(i) 12-Steps Ahead Forecast

```
m4_ca <- Arima(ca_ts, order = c(2,0,0), include.drift = TRUE,
               season = list(order = c(0,0,1))) # CA alternate arima model
m4_ny <- Arima(ny_ts, order = c(3,0,0), include.drift = TRUE,
               season = list(order = c(1,0,1))) # NY alternate arima model
plot(forecast(m4_ca, n.ahead = 12), main = "Plot of Data, Respective Fit, and Forecast for CA")
```

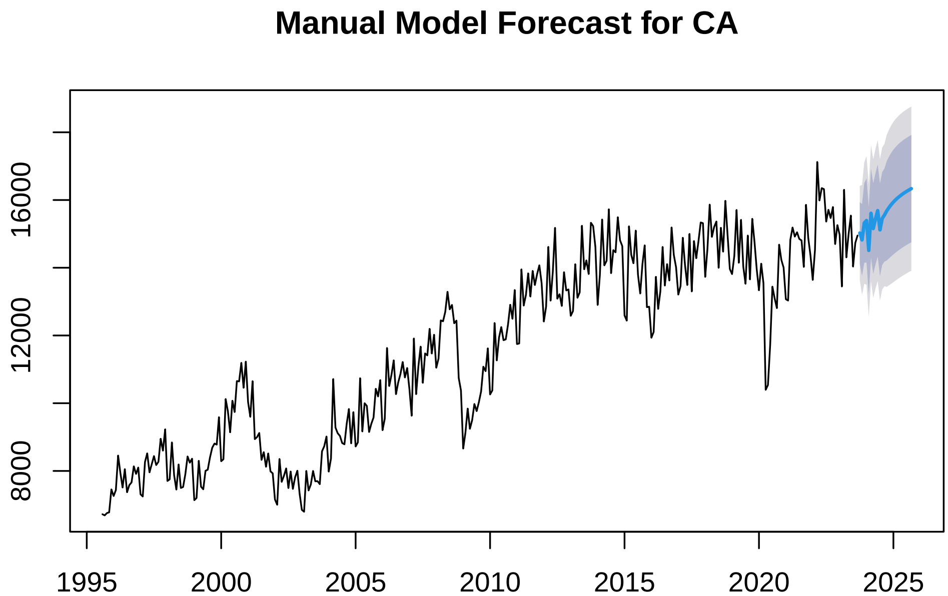



```r
plot(forecast(m4_ny, n.ahead = 12), main = "Plot of Data, Respective Fit, and Forecast for NY")
```

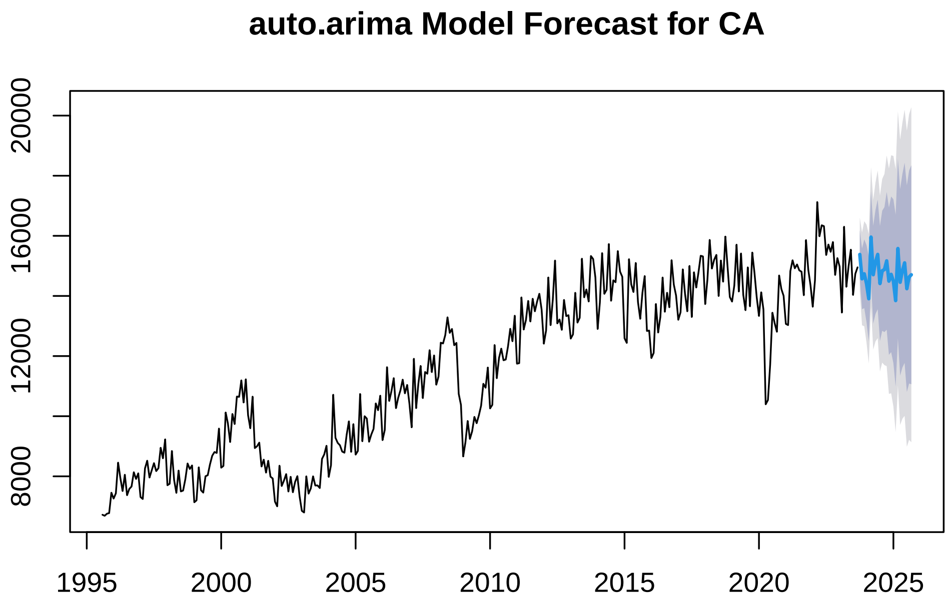


(j) Compare to Auto.Arima

```
# Comparing Forecasts for CA
plot(forecast(m4_ca, n.ahead = 12), main = "Manual Model Forecast for CA") # CA Manual
```

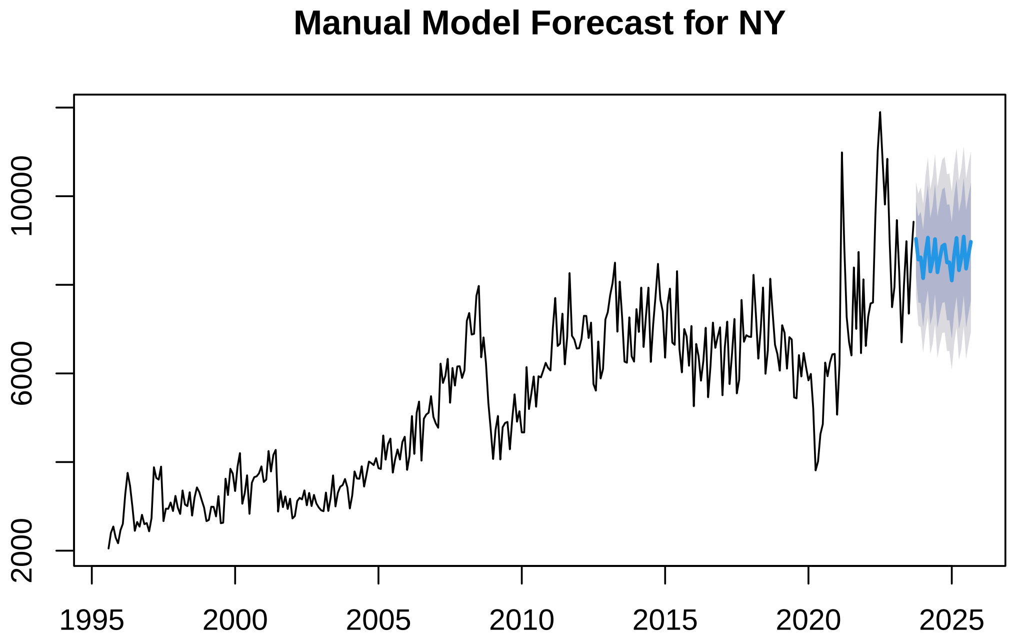



```
plot(forecast(auto.arima(ca_ts), n.ahead=12),
     main = "auto.arima Model Forecast for CA") # CA auto.arima
```

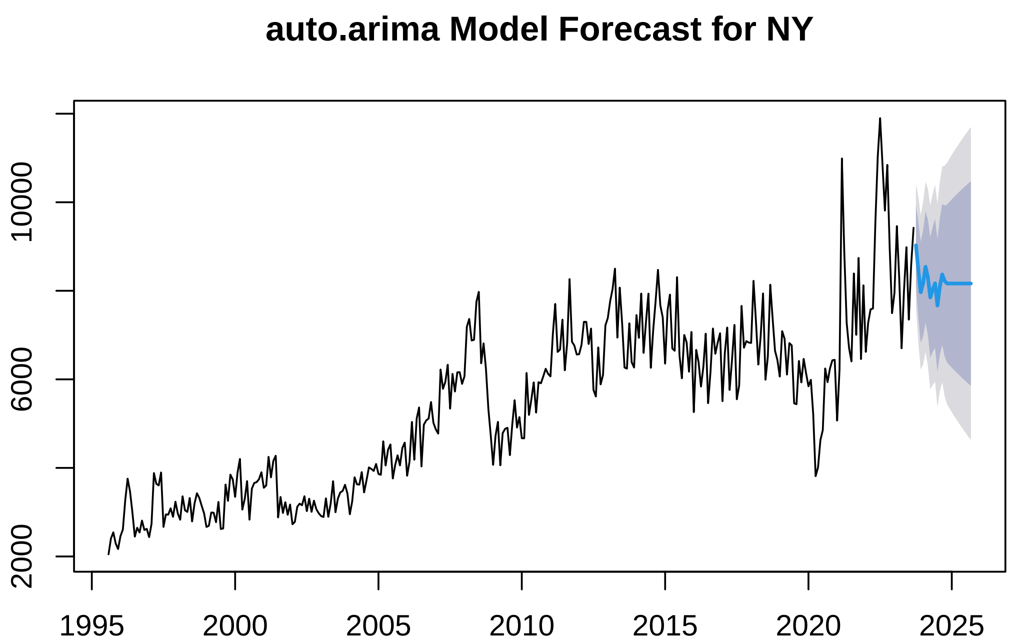


For California exports, our two models have different forecasts. Our manual model's forecast is overall more optimistic, and loses its seasonality and cyclic patterns as time increases. On the other hand, the auto.arima model's forecast is very conservative. The trend is completely straight, predicting that exports will—on average—remain at the level that they currently are at. Looking at the error bars, we find that our model's forecast has smaller error bars, whereas the error bars for the auto.arima model's forecast are exponentially increasing in size. This indicates that the auto.arima model's prediction carries more uncertainty than that from our manual model.

```r
# Comparing Forecasts for NY
plot(forecast(m4_ny, n.ahead = 12), main = "Manual Model Forecast for NY") # NY Manual
```




```r
plot(forecast(auto.arima(ny_ts), n.ahead=12),
     main = "auto.arima Model Forecast for NY") # NY auto.arima
```



For New York exports, the two model's respective forecasts are more similar. They both predict a straight line trend, with exports staying on average at the level they are currently at. The main difference is that the auto-arima model's forecast loses seasonality and cycles as the forecast gets further into the future. In terms of error bars, our manual model has narrower confidence intervals that stay around the same size. The auto.arima error bars increase in width overtime, indicating increasing uncertainty as time goes on.

```r
# Comparing MAPE for CA
MAPE(m3_ca$resid, ca_ts) # manual model
```

```
## [1] 4.800544
```

```r
MAPE(auto.arima(ca_ts)$resid, ca_ts) # auto.arima
```

```
## [1] 3.999594
```

```r
# Comparing MAPE for NY
MAPE(m3_ny$resid, ny_ts) # manual model
```

```
## [1] 8.028586
```

```r
MAPE(auto.arima(ny_ts)$resid, ny_ts) # auto.arima
```

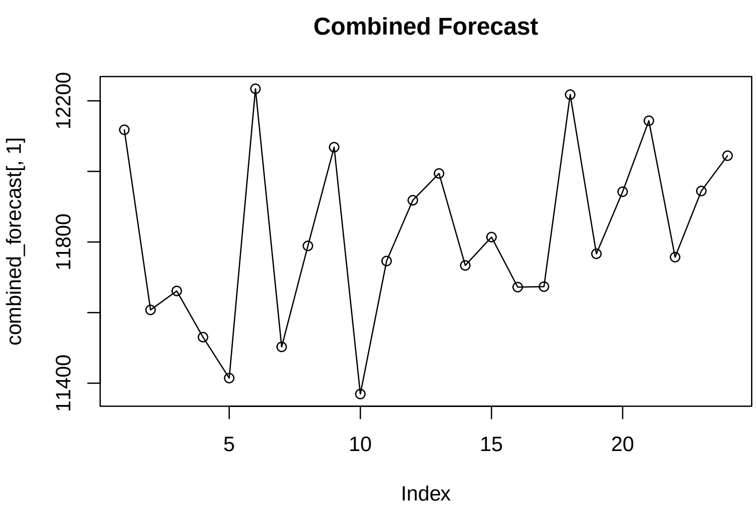
```
## [1] 8.870101
```

For California, the auto.arima model performs better as it has a lower MAPE of 3.999594 compared to an MAPE of 4.800544 for our manual model. This means that the auto.arima model is more accurate than ours.

On the other hand, our model outperforms auto.arima for the New York series with an MAPE of 8.028586 compared to auto.arima's 8.870101. This indicates that our model is more accurate than auto.arima, as determined by MAPE.

(k) Combination Forecast

```r
# combining 4 forecasts
combined_forecast <- (data.frame(forecast(m4_ca, n.ahead = 12)) +
                      data.frame(forecast(m4_ny, n.ahead = 12)) +
                      data.frame(forecast(auto.arima(ca_ts), n.ahead = 12)) +
                      data.frame(forecast(auto.arima(ny_ts), n.ahead = 12)))/4
# plotting combined forecast
plot(combined_forecast[,1], main = "Combined Forecast") + lines(combined_forecast[,1])
```




```
## integer(0)
```

```r
# MAPE for combined forecast (by averaging residuals and comparing to combined CA/NY series)
avgres <- (m3_ca$resid + auto.arima(ca_ts)$resid + m3_ny$resid + auto.arima(ny_ts)$resid)/4
MAPE(avgres, ca_ts + ny_ts)
```

```
## [1] 2.300376
```

Combining the forecasts, we achieve an MAPE of 2.300376 Comparing this to the individual models, we find that

- auto.arima for CA performs worse than the combined forecast
- manual model for CA performs worse with combined forecast
- auto.arima for NY performs worse than the combined forecast
- manual model for NY performs worse than the combined forecast.

(l) VAR Model

```
# creating VAR model
y <- cbind(ca_ts, ny_ts)
VARselect(y)$selection
```

```
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     10      3      3     10
```

```
y_tot <- data.frame(y)
y_model <- VAR(y_tot, p=10)

kable(summary(y_model)$varresult$ca_ts$coefficients) # CA
```

	Estimate	Std. Error	t value	Pr(>|t|)
ca_ts.l1	0.4011988	0.0632475	6.3433109	0.0000000
ny_ts.l1	0.2688340	0.0709682	3.7880930	0.0001827
ca_ts.l2	0.1146863	0.0692511	1.6560936	0.0987246
ny_ts.l2	-0.0741259	0.0799635	-0.9269961	0.3546567
ca_ts.l3	0.3131756	0.0685090	4.5713085	0.0000070
ny_ts.l3	-0.0421051	0.0797166	-0.5281851	0.5977524
ca_ts.l4	-0.0965336	0.0699141	-1.3807449	0.1683618
ny_ts.l4	-0.0761058	0.0802529	-0.9483243	0.3437103
ca_ts.l5	0.2555565	0.0695190	3.6760668	0.0002795
ny_ts.l5	-0.0394104	0.0799952	-0.4926590	0.6226057
ca_ts.l6	-0.1356234	0.0694670	-1.9523433	0.0518062
ny_ts.l6	0.1294135	0.0802389	1.6128518	0.1078042
ca_ts.l7	0.1116509	0.0702820	1.5886127	0.1131772
ny_ts.l7	-0.0085247	0.0804394	-0.1059767	0.9156700
ca_ts.l8	-0.1620157	0.0717724	-2.2573548	0.0246884
ny_ts.l8	-0.0095311	0.0802619	-0.1187503	0.9055508
ca_ts.l9	0.1512269	0.0730860	2.0691644	0.0393672
ny_ts.l9	0.1166741	0.0803708	1.4516977	0.1476068
ca_ts.l10	-0.0055457	0.0654389	-0.0847461	0.9325185
ny_ts.l10	-0.2209529	0.0727326	-3.0378811	0.0025871
const	403.4013238	260.2362973	1.5501347	0.1221394

```
kable(summary(y_model)$varresult$ny_ts$coefficients) # NY
```

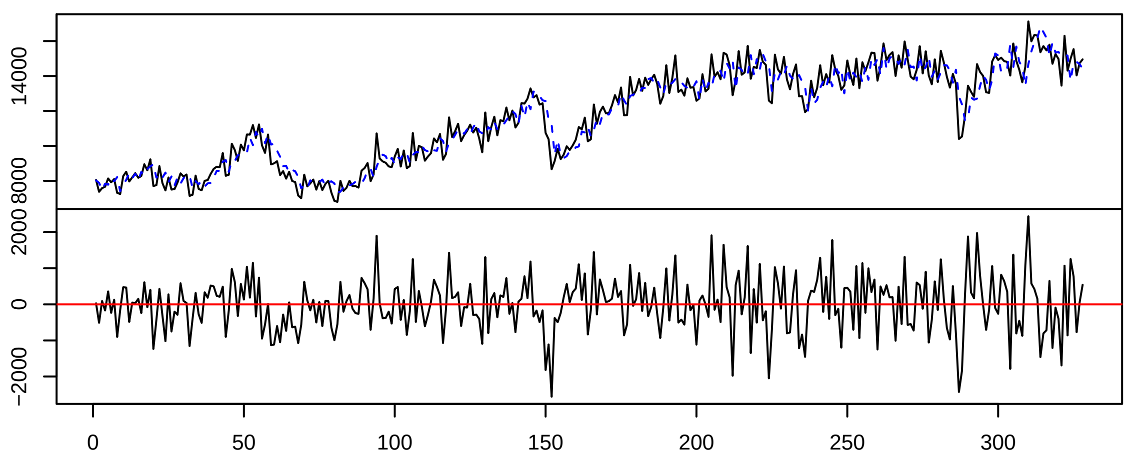
	Estimate	Std. Error	t value	Pr(>|t|)
ca_ts.l1	-0.1090445	0.0569085	-1.9161376	0.0562761
ny_ts.l1	0.6231009	0.0638553	9.7580163	0.0000000
ca_ts.l2	0.1089602	0.0623103	1.7486708	0.0813474
ny_ts.l2	-0.0375124	0.0719491	-0.5213746	0.6024818
ca_ts.l3	0.1623399	0.0616426	2.6335686	0.0088772

	Estimate	Std. Error	t value	Pr(>|t|)
ny_ts.l3	0.1872986	0.0717269	2.6112763	0.0094635
ca_ts.l4	-0.0717941	0.0629069	-1.1412748	0.2546450
ny_ts.l4	-0.0018498	0.0722095	-0.0256175	0.9795790
ca_ts.l5	0.1171056	0.0625514	1.8721509	0.0621363
ny_ts.l5	-0.0789656	0.0719776	-1.0970856	0.2734637
ca_ts.l6	0.0416770	0.0625046	0.6667831	0.5054115
ny_ts.l6	-0.0030766	0.0721969	-0.0426138	0.9660371
ca_ts.l7	-0.1278711	0.0632379	-2.0220636	0.0440368
ny_ts.l7	0.1362068	0.0723773	1.8819009	0.0607953
ca_ts.l8	0.0494117	0.0645789	0.7651368	0.4447779
ny_ts.l8	-0.0726117	0.0722175	-1.0054581	0.3154682
ca_ts.l9	-0.0605163	0.0657608	-0.9202486	0.3581654
ny_ts.l9	0.2059487	0.0723155	2.8479197	0.0046975
ca_ts.l10	0.0136279	0.0588802	0.2314520	0.8171179
ny_ts.l10	-0.1652967	0.0654428	-2.5258177	0.0120458
const	-288.8184820	234.1537549	-1.2334565	0.2183490

```
# plot of VAR model
plot(y_model, names="ca_ts") # CA
```

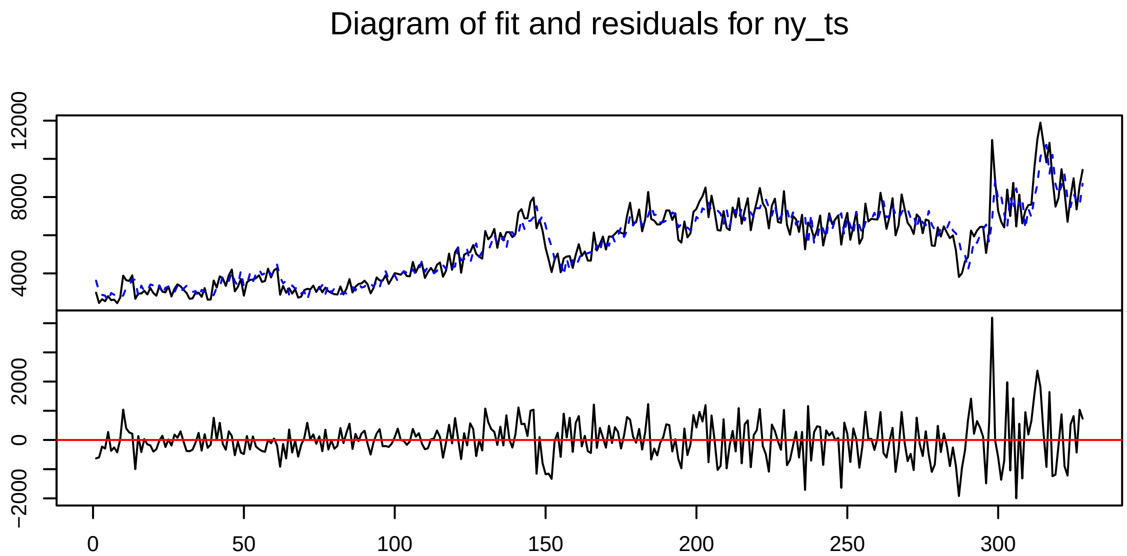
```
## Error in plot.new(): figure margins too large
```

Diagram of fit and residuals for ca_ts

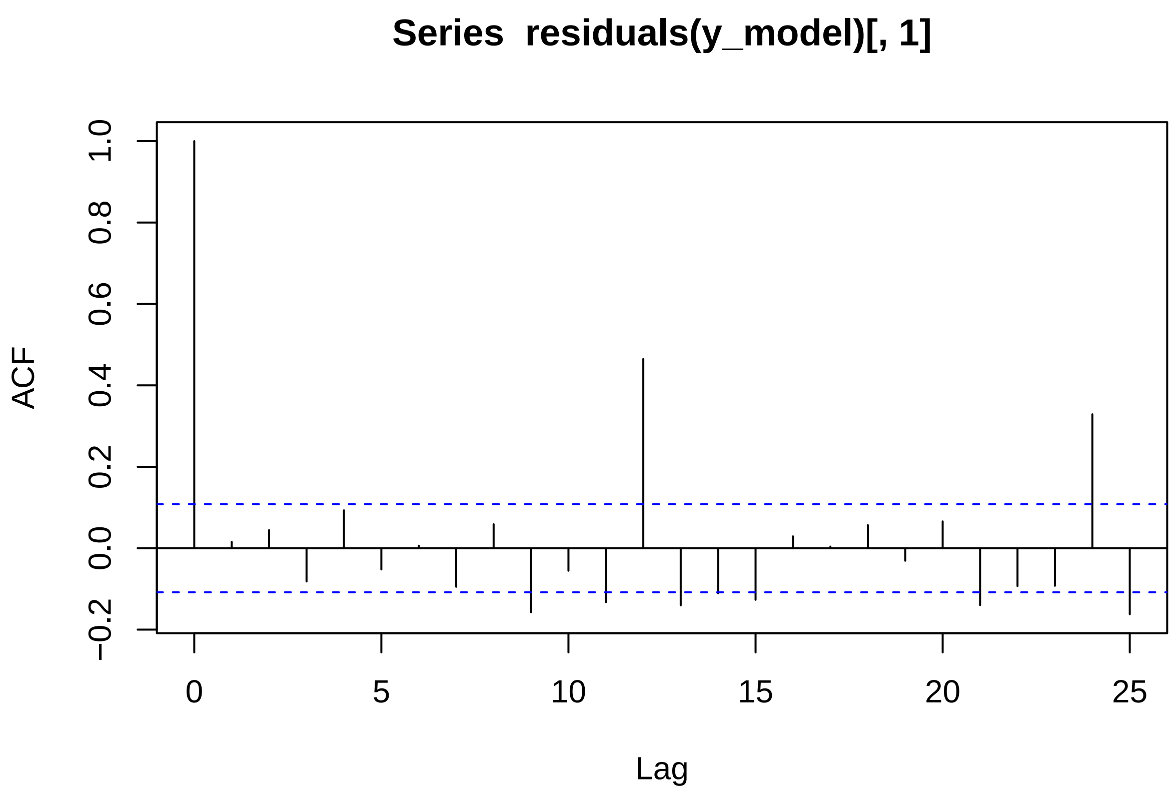



```
plot(y_model, names="ny_ts") # NY
```

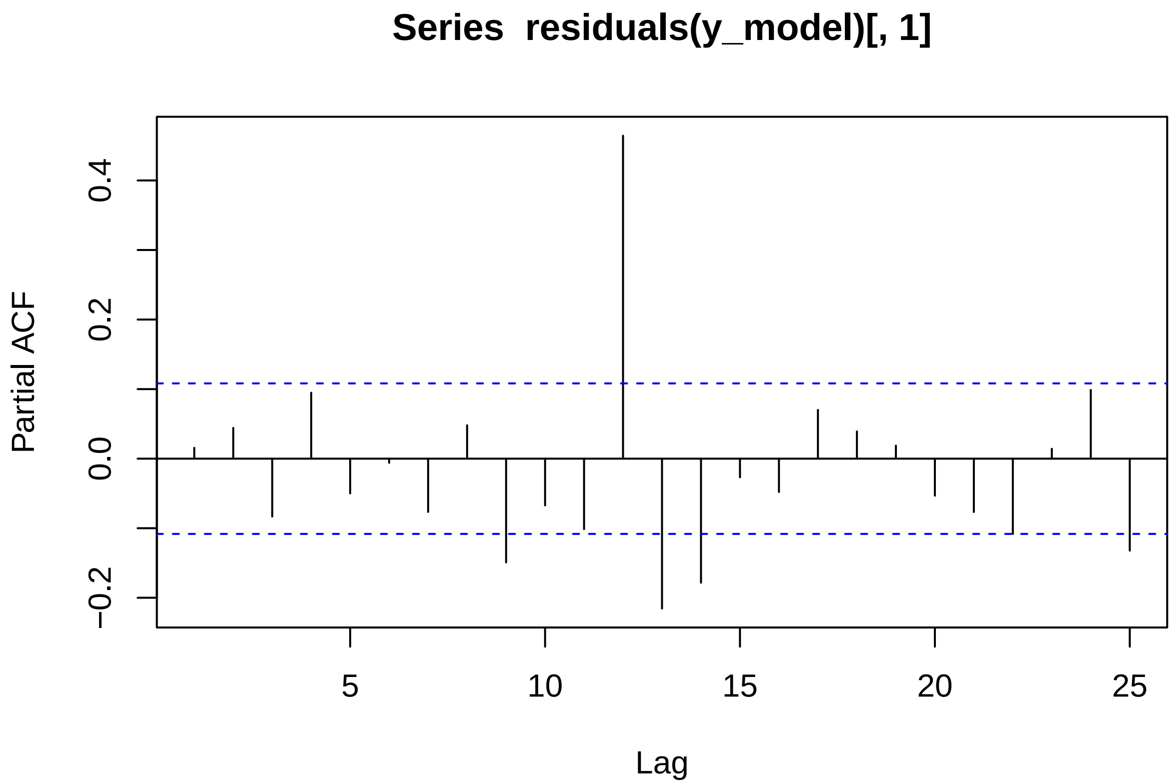
```
## Error in plot.new(): figure margins too large
```



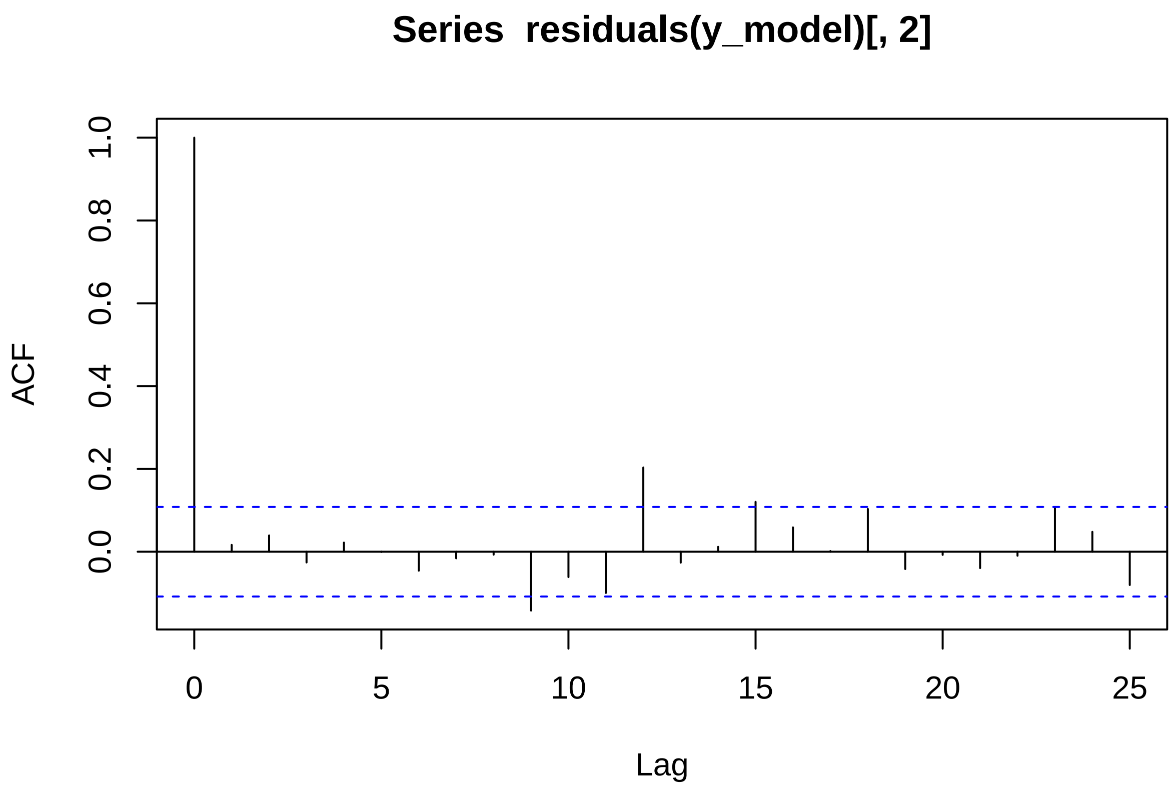

```r
acf(residuals(y_model)[,1]) # ACF for CA
```



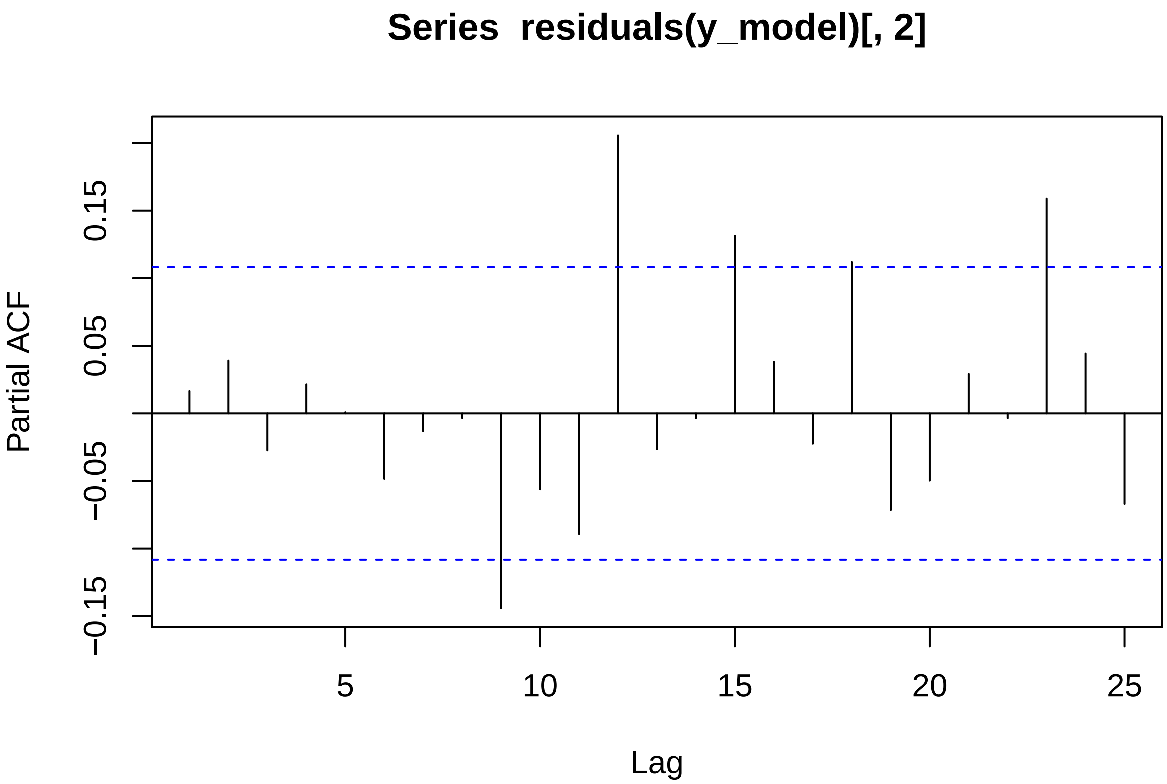

```r
pacf(residuals(y_model)[,1]) # PACF for CA
```




```
acf(residuals(y_model)[,2]) # ACF for NY
```




```
pacf(residuals(y_model)[,2])  # PACF for NY
```



In the VAR fitted model for California, we can see that our model fit is fairly accurate as the fitted values and the original data are closely aligned with one another. Similarly the VAR fitted model for New York also yields a fairly accurate fit. The observed values seemed to be slightly lagged in comparison to the fitted values, but they are still strongly associated. The residual plots for both states showed very similar results to the final ARIMA model we previously fitted, which had significant outliers and was non-stationary, suggesting we should fit the VAR model to differenced data in the future.

To further access the VAR model, we can look at the ACF and PACF plots. For California, The ACF plot shows a significant spike at lags 12 and 24 which suggests that seasonal behavior was not captured by the model. For New York, the ACF and PACF show a spike at lag 12 and there are some potentially significant spikes at other lags throughout the rest of the data. These spikes could indicate seasonal behavior not captured by the model. We suggest further adding a seasonal AR component to improve the model's ability to capture and represent the underlying seasonal patterns.

(m) Impulse Response Functions

```
kable(data.frame(irf(y_model)$irf$ca_ts, irf(y_model)$irf$ny_ts)) # IRF coefficients
```

ca_ts	ny_ts	ca_ts.1	ny_ts.1
778.6967	316.94969	0.00000	624.8635
397.6190	112.57910	167.98457	389.3530
255.6011	99.74743	125.74799	200.8483
397.1424	259.16018	68.53926	232.1708
271.4165	171.42861	78.10599	242.3992
396.7207	208.35536	47.60203	137.1902
347.3807	273.33854	139.68431	111.9265
368.8997	196.90546	117.88789	197.0603
280.5130	210.02884	117.92446	128.8176
399.0618	256.33995	190.31537	224.9727
375.7559	198.25533	91.11675	151.7284

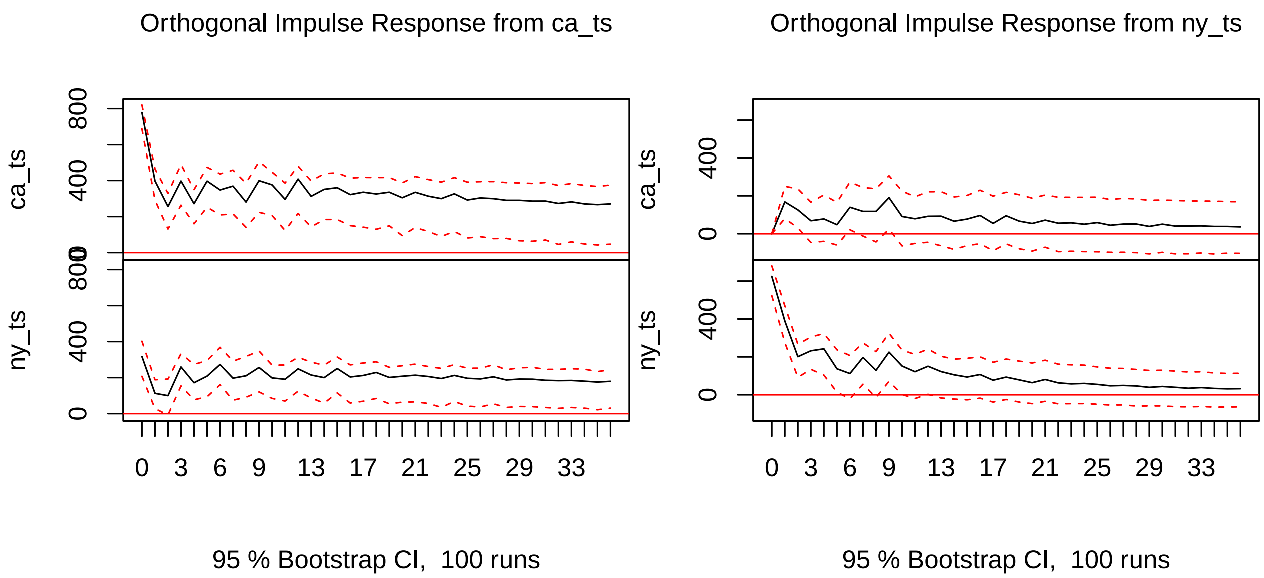
```
kable(data.frame(irf(y_model)$Lower$ca_ts, irf(y_model)$Lower$ny_ts)) # Lower CI Bound
```

ca_ts	ny_ts	ca_ts.1	ny_ts.1
694.6217	227.20093	0.000000	537.495310
291.3312	18.11810	80.397242	292.187407
134.7645	17.29695	7.414363	103.635698
276.8310	132.13489	-23.962937	123.804241
132.9533	70.90836	-38.503480	128.553823
256.7270	102.75428	-85.829610	16.833794
215.1166	157.94157	18.428784	8.936671
202.2084	51.98269	-12.280777	69.988413
116.2262	50.07857	-22.750320	11.568712
211.3707	123.56703	29.026457	83.839786
225.1632	92.19581	-69.264374	2.562664

```
kable(data.frame(irf(y_model)$Upper$ca_ts, irf(y_model)$Upper$ny_ts)) # Upper CI Bound
```

ca_ts	ny_ts	ca_ts.1	ny_ts.1
818.4944	392.4603	0.0000	676.1041
480.2373	192.9749	243.9171	449.5894
343.0161	171.6852	198.5659	257.8389
502.3748	341.7350	144.6721	320.4461
365.4230	259.2247	183.2792	329.0506
479.6031	277.3151	127.8919	227.3155
437.6576	353.8984	250.1977	195.7964
445.4827	254.4819	227.2829	278.4238
366.1476	282.3794	220.2277	229.6032
476.8043	316.9501	290.4981	308.0179
452.0427	251.8656	195.5635	245.9101

```r
plot(irf(y_model, n.ahead=36))
```



The first pair of plots show the response to a unit shock in California exports. We can see that it is highly persistent and turns insignificant after 3 months. We see an initial positive effect on both California and New York exports; however, the effect on California exports is much more significant (which makes sense). The impulse response slowly declines after the third month.

The second pair of plots shows the response to a unit shock in New York exports. We see the shock caused a significant initial positive effect on New York exports but a small negative effect on California exports. The impulse response declines towards zero after the second month in both states, but the decline is seemingly much slower in California.

(n) Granger-Causality Test

```r
grangertest(ca_ts ~ ny_ts, order = 10)
```

```
## Granger causality test
##
## Model 1: ca_ts ~ Lags(ca_ts, 1:10) + Lags(ny_ts, 1:10)
## Model 2: ca_ts ~ Lags(ca_ts, 1:10)
##   Res.Df  Df     F   Pr(>F)
## 1    307
## 2    317 -10 2.469 0.007458 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
grangertest(ny_ts ~ ca_ts, order = 10)
```

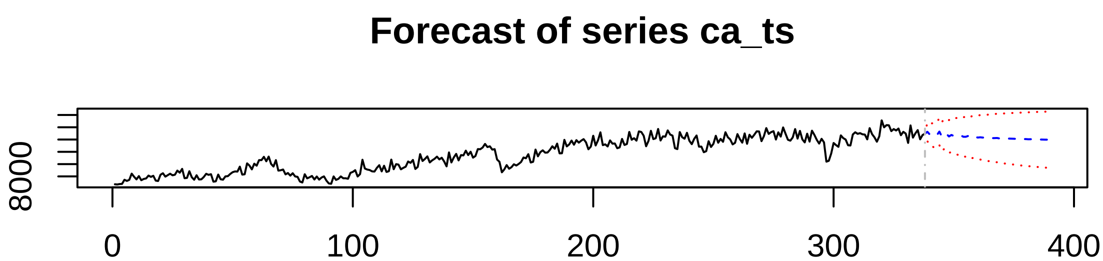
```
## Granger causality test
##
## Model 1: ny_ts ~ Lags(ny_ts, 1:10) + Lags(ca_ts, 1:10)
## Model 2: ny_ts ~ Lags(ny_ts, 1:10)
##   Res.Df  Df      F    Pr(>F)
## 1    307
## 2    317 -10 3.0886 0.0009205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

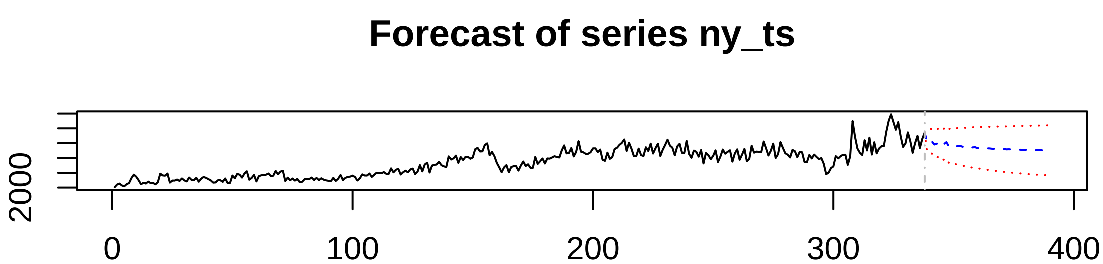
In the first Granger causality test, we are testing if New York exports Granger cause California exports. Since the p-value is 0.007458, we reject the null hypothesis at the 0.05 significance level. This suggests that New York exports can be used to predict California exports.

In the second Granger causality test, we are testing if California exports Granger cause New York exports. Since the p-value is 0.0009205, we reject the null hypothesis at the 0.05 significance level. This suggests the California exports can be used to predict New York exports. We observe that the second Granger test yields a more significant p-value compared to the first test. This implies that California exports have a more pronounced impact than New York exports in forecasting future export values.

(o) VAR Model Forecast

```
plot(predict(object=y_model, n.ahead=52))
```





Compared to the models we forecasted with in part (i), we see that these forecasts expect a more neutral or even slightly downward trend from both datasets, while the other forecasts predicted a more steady upward trend, especially for the California forecast. We also see that these trends do not have any seasonal component since they are very flat, while the other forecasts had a strong seasonal component in the predictions.

III Conclusions and Future Work

In our project, we fit four models to forecast exports in California and New York. We first fit a model that included trend, seasonality, and cyclical components. We used a linear trend on the data and using the ACF and PACF plots, we concluded the California dataset exhibited a seasonal AR(1) pattern and an AR(2) model was optimal to model cycles. For New York, we concluded the dataset exhibited a seasonal ARMA(1,1) pattern and an AR(3) model was optimal to model cycles. The second model we fit was an ARIMA model using auto.arima. The ARIMA model for California was an ARIMA(2,1,0)(2,0,0)[12] and for New York it was ARIMA(0,1,2)(0,0,1)[12]. In terms of MAPE, for California we concluded that the auto.arima model performs better as it has a lower MAPE compared to the manual model. And for New York, our model outperforms auto.arima since it has a lower MAPE. The third model we fit was a combined model using the previous two models. Comparing this model to the previous two models, we obtain an MAPE of 4.800544. We conclude that for California, the ARIMA model performs better and the manual model performs similarly to the combined forecast. However, for New York, the combined model performs better than the ARIMA and manual model. Lastly we fit a VAR model of order 10. It yielded a fairly accurate fit for both datasets having a very slight lag. We suggest adding a seasonal component after viewing the ACF/PACF plots and seeing that the forecast showed a flat trend. The Granger causality tests suggested that California exports cause New York exports and vice versa. However, California exports have a more significant effect. To improve our forecasts we can apply advanced techniques such as ets. Reevaluating and updating our model periodically would also be beneficial for future work forecasting exports in California and New York, in order to better understand California and New York state markets in addition to the US market as a whole.

IV References

U.S. Census Bureau, Exports of Goods for New York [EXPTOTNY], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/EXPTOTNY, November 16, 2023.

U.S. Census Bureau, Exports of Goods for California [EXPTOTCA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/EXPTOTCA, November 16, 2023.