HE3022 Report

Agarwal Gopal, Agarwal Heena, Manasi Murali, Sharma Uday

Contents

Introduction

Dataset

- Name: Retail Sales Index
- Coverage: January, 1985 to January, 2021
- Frequency: Monthly
- Source: Ministry of Trade and Industry - Department of Statistics (Singapore Department of Statistics, 2021)
- Base year: 2017

Importance of Retail Sales

Retail Sales is an important part of the economy of any country. It measures the sales of durable and non-durable goods and is used to gauge consumer demand over a certain period of time (Investopedia, n.d.). It indicates the health of the economy and whether it's heading towards contraction or expansion in the business cycle and is one of the most important economic indicators. This is also because consumer spending is a big part of the Gross Domestic Product (GDP) of a country.

Retail Sales is measured in the local currency, which, for our case, is Singapore Dollars (S$). For the month of January '21, the total retail sales value was estimated to be around S$3.8 billion for Singapore (ChannelNewsAsia, 2021).

Although forecasting this value would be of great economic importance, due to several factors such as inflation, this value should not be forecasted directly as it would lead to erroneous results and misleading conclusions. This is why we use the Retail Sales Index (referred to as RSI from now on).

What is RSI?

RSI measures the short-term performance of retail industries based on the sales records of retail establishments (Retail Sales Index, 2020).

The RSI is presented at both current prices and constant prices. The index at current prices measures the changes of sales values which can result from changes in both price and quantity. By removing the price effect, the index at constant prices measure the changes in the volume of economic activity. We study the index at constant prices, with 2017 as the base year.

What does RSI include?

RSI comprises of the following types of retailers in Singapore(SingStat, 2021):

1. Sell via both physical stores and online/e-commerce sites
2. Sell via physical stores only
3. Sell mainly via online/e-commerce sites

These include the following industries (Ministry of Trade & Industry [MTI], 2021):

- Motor Vehicles
- Computer & Telecomms Equipment
- Food & Alcohol
- Recreational Goods
- Wearing Apparel & Footwear
- Department Stores
- Optical Goods & Books
- Petrol Service Stations
- Supermarkets & Hypermarkets
- Mini-marts & Convenience Stores
- Watches & Jewellery

- Cosmetics, Toiletries & Medical Goods
- Furniture & Household Equipment
- Others

Recent Industry Trends

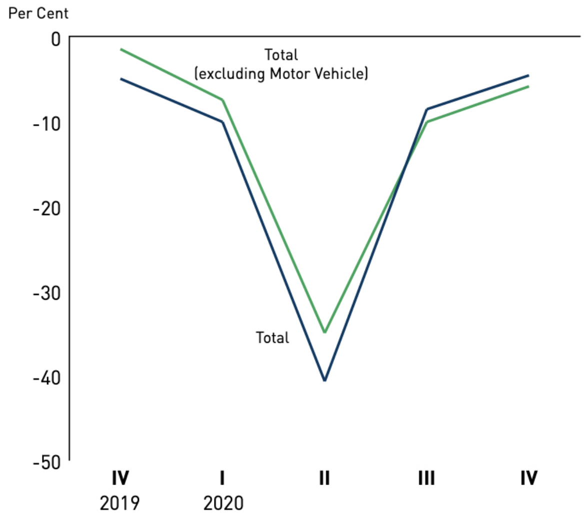


Figure 1. Trend of RSI in 2020 (MTI, 2021).

The Retail Trade sector of Singapore contracted by 4.7 per cent year-on-year in the fourth quarter of 2020, which was an improvement from the 8.6% decline in the previous quarter. For the whole of 2020, the sector shrank by 16%, an extremely poor performance as compared to the 2.4% contraction in 2019, marked by a sharp decline in the second quarter, which coincided with the Circuit Breaker. The primary reason to which this shrink can be attributed to is the outbreak of the global COVID-19 pandemic, which resulted in stricter health measures, including the Circuit Breaker and other border control measures.

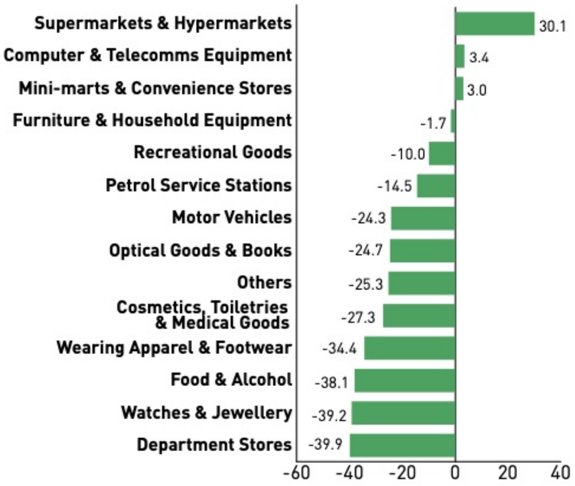


Figure 2. Sector-wise trends in RSI in Singapore for 2020 (MTI, 2021).

In terms of motor vehicle sales, the drop was in line with the reduction in COE supply.

However, in terms of non-motor vehicle sales, there was a large decline caused by slowed demand and public health measures. Although there was an increase in sales for Supermarkets & Hypermarkets (+30.1%) due to increased demand caused by the pandemic and lockdown, most of the other sectors took a hit, especially luxury items like watches & jewellery. This was amplified by the fact that Singapore had closed borders to tourists, who used to contribute a lot to the retail sales.

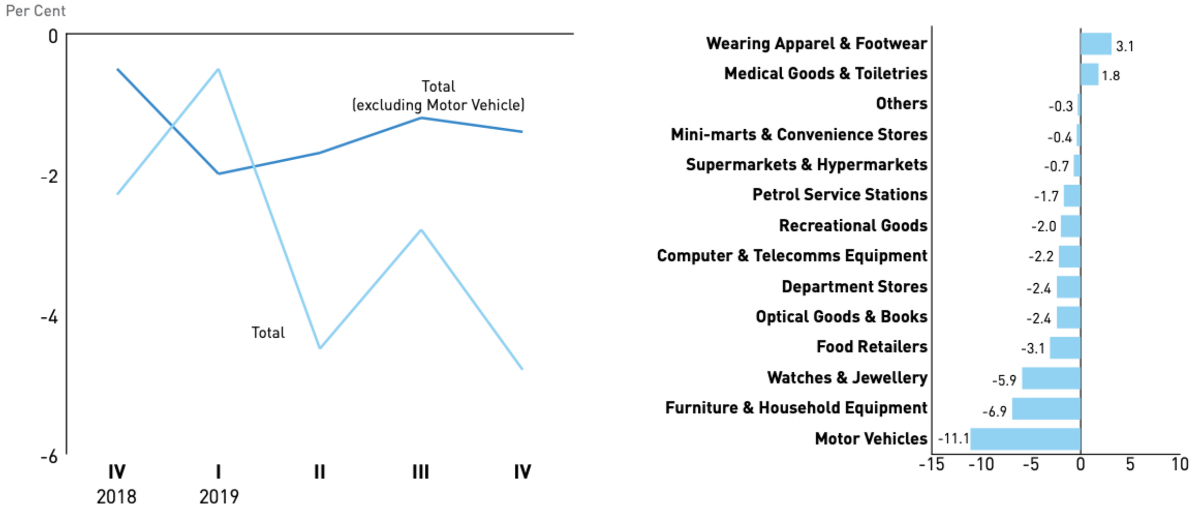


Figure 3. Trends for RSI in 2019, overall and sector-wise (MTI, 2020).

Studying the trends for 2019 shows that the RSI was already declining before the pandemic hit, amplified by dropping motor vehicle sales caused by a reduction in COE numbers. For non-motor vehicle sales, the drop was caused by a fall in sales volume for both discretionary and non-discretionary goods.

Outlook for 2021

- There is still uncertainty regarding the trajectory of the global economic recovery. Although the United States and Eurozone are heading towards herd immunity due to their strong vaccine drives, there is still instability in the region (MTI, 2021). There has been a recent surge of COVID infections in India, and it has been leading the world in number of cases (Fung, 2021). The net effect on the external demand for Singapore still seems uncertain.
- Singapore was faring well domestically in terms of controlling the number of cases and vaccinating its citizens, but a new cluster has recently developed, caused by the highly-contagious Indian variant of the virus (Teo, 2021). This has led to increased restrictions in the country and the possibility of another Circuit Breaker (Kit, 2021). The domestic outlook is uncertain.
- In terms of the RSI, the consumer-facing sector of retail trade is predicted to gain slowly due to an imminent global economic recovery over the course of the year, but the restriction on tourists along with capacity constraints due to the health measures are expected to offset some of this gain. It is not expected to reach pre-COVID levels this year (MTI, 2021).

Methodology

Reading the Dataset

The Excel file is loaded, and the data are converted to a timeseries object of monthly frequency, from January 1985 to January 2021.

Training and Testing Models

RSI data are split into a training set (80%) and a test set (20%). The training and test sets will be used to visualize the accuracy of the models.

However, as the test set includes the outlier values caused by the COVID-19 pandemic, using test set accuracy to compare different models will give skewed results. Hence, we use cross-validation root mean square errors (RMSE), calculated using the *tsCV* function, will be used to compare the different models, as this provides a more robust comparison.

At the end, the best models are picked using the cross-validation RMS errors and used for forecasting.

Exploratory Analysis

Time Series Plot

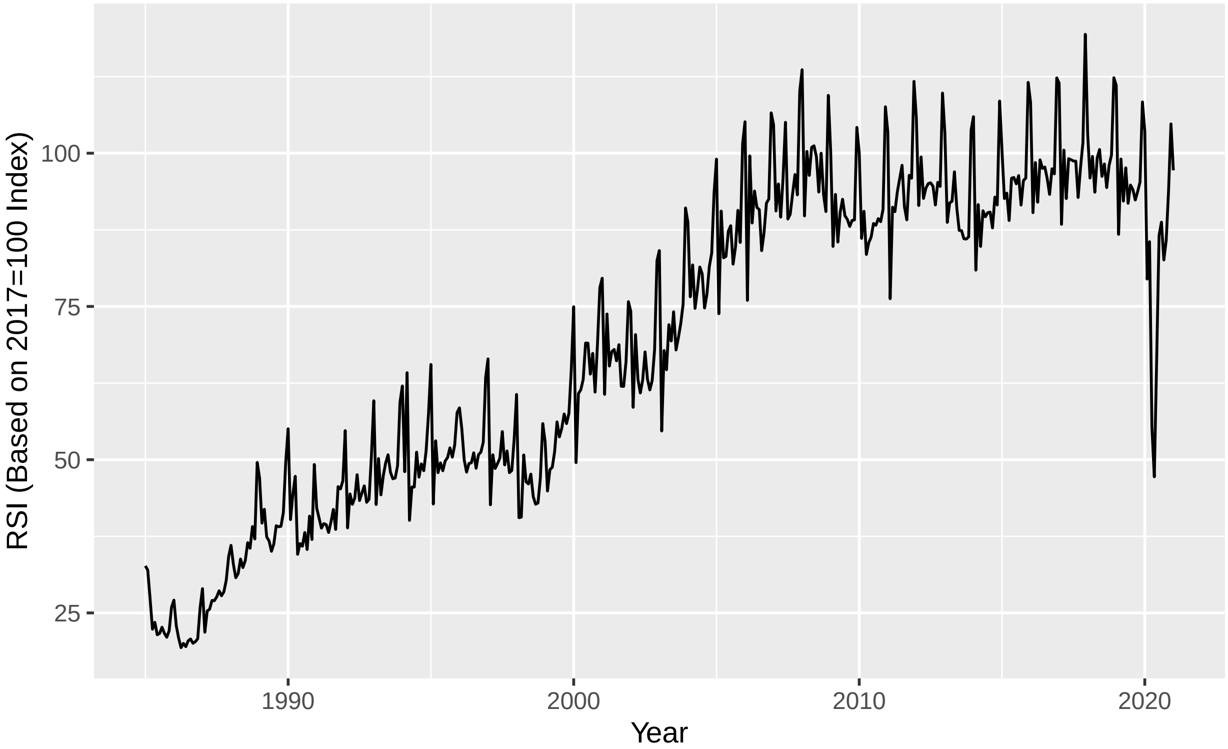


Figure 4. Time series plot of RSI.

- The RSI shows an overall upward trend from 1985-2021.
- Some of the slumps observed can be attributed to the following (Sng, 2020):
  - 1985 Recessionary Crisis
  - 1998 Asian Financial Crisis
  - 2001-03 Low Growth and High Unemployment Crisis (9/11, Bali Bombings, SARS Pandemic)
  - 2008 Global Great Recession
  - 2020 COVID-19 Pandemic
- The RSI also shows a seasonal pattern, that remains approximately the same size as the level of the series increases.
- There is no cyclic behaviour observed.

Seasonal Plot

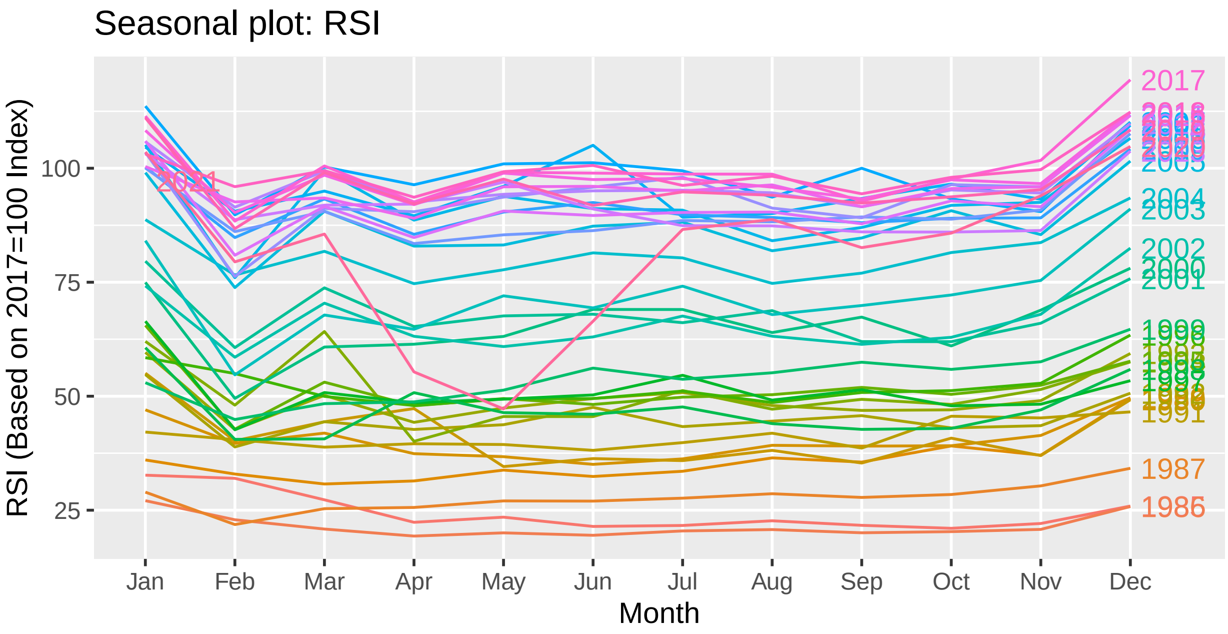


Figure 5. Seasonal plot of RSI.

- Most of the years show a similar seasonal pattern: there are increased sales in January and December, which could be attributed to the holiday season (Christmas), as well as a small spike in March which could be attributed to the Chinese New Year. The variation is small for the other months, but similar across the years.
- A sharp fall in April '20 marks the beginning of the Circuit Breaker in Singapore, which led to a lot of restrictions and hence a significant drop in the RSI.

Subseries Plot

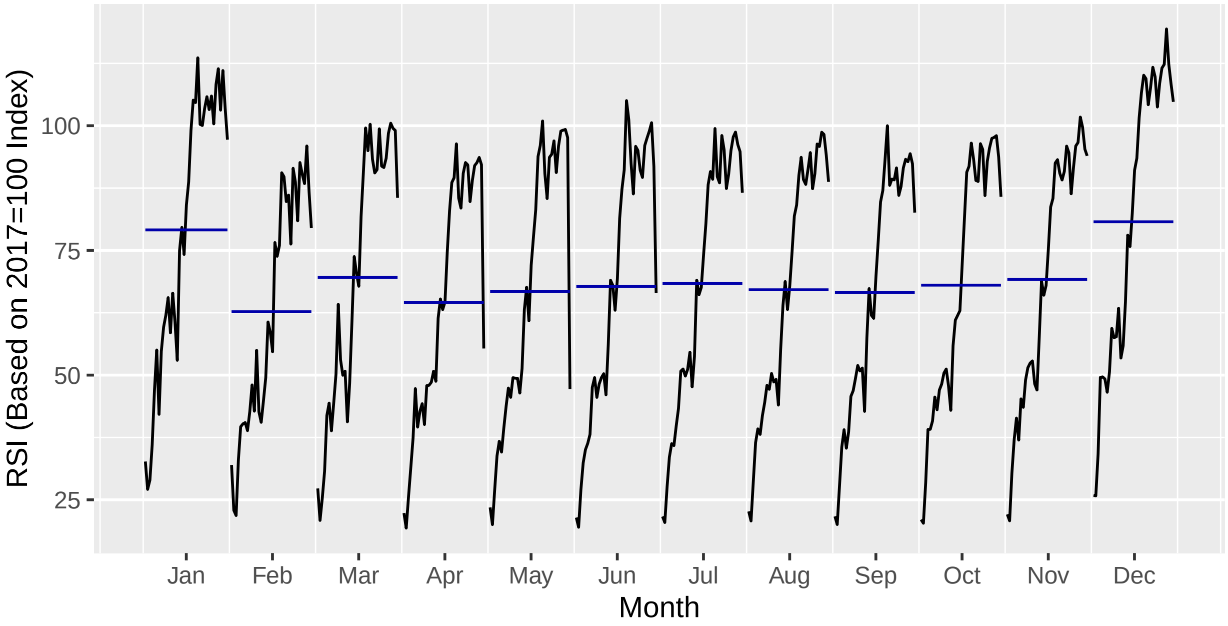


Figure 6. Subseries plot of RSI.

- Months like February with lower number of days have a lower RSI on average, and it fluctuates from April-November based on the number of days each month has.
- All months show an overall increase as the years increase, confirming the upward trend observed in the time series plot.

ACF Plot

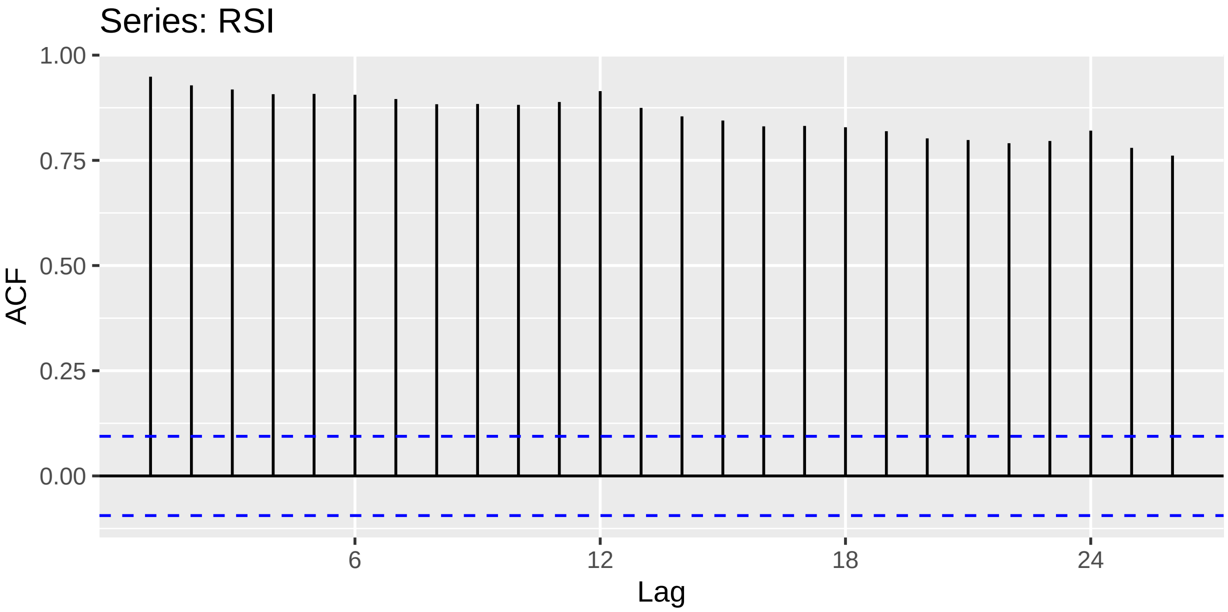


Figure 7. ACF plot of RSI.

- Autocorrelation exists for all lag values.
- The ACF plot shows a slowly decreasing trend, as well as a seasonal pattern of monthly frequency.

Lag Plot

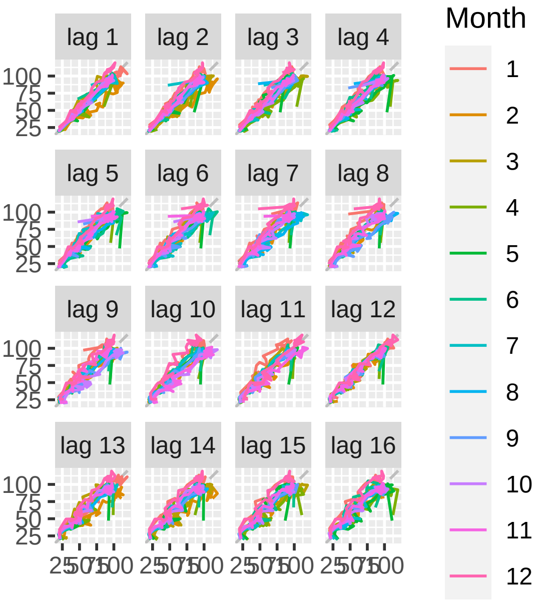


Figure 8. Lag plot of RSI.

- The lag plots show a bivariate scatter plot for the current and lagged values for each month.
- There exists a strong relationship for all lagged values, with lag 12 having the strongest correlation. This confirms the monthly seasonal pattern we observed above, as well as the ACF plot which showed that the RSI is serially correlated.

Benchmark Methods

Unsatisfactory Methods

The benchmark methods tested include Average, Naive, Seasonal Naive, and Random Walk with Drift methods. As these methods do not account for either the seasonality or the trend, their results will not be satisfactory (see Appendix).

STL-Random Walk with Drift

The STL method is used to decompose the time series into its trend, seasonal and remainder components. The method can be configured to perform multiplicative decomposition, however, as the seasonality of our time series approximately remains constant as the level of the series increases, additive decomposition is sufficient.

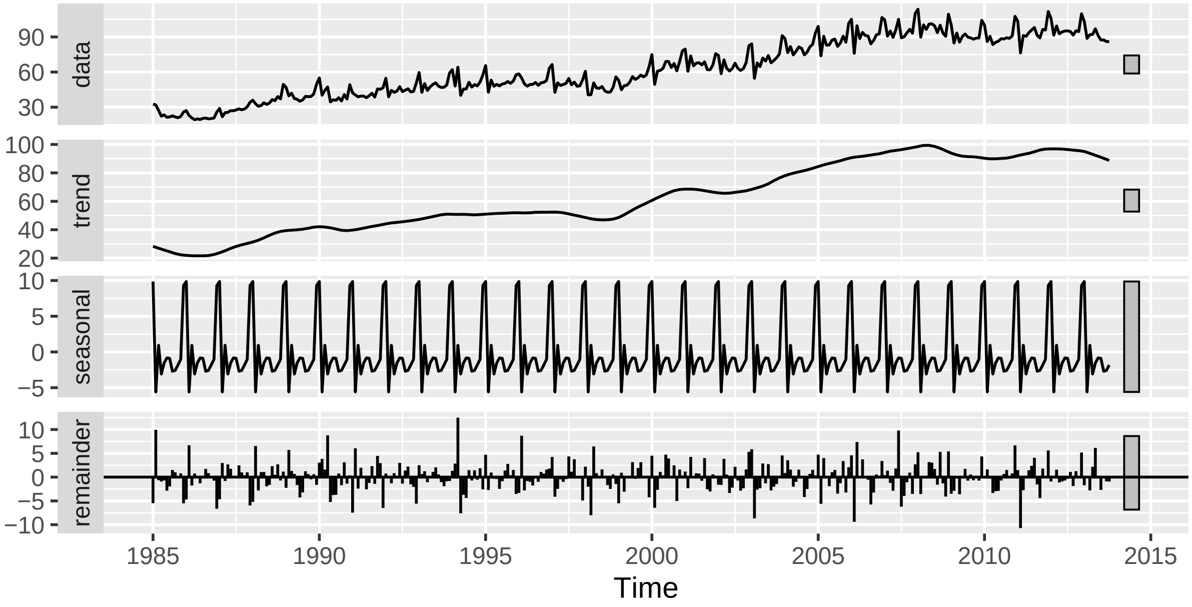


Figure 9. RSI after STL decomposition.

The STL-Naive method is unsatisfactory and will not lead to good forecasts as it doesn't account for a trend (see Appendix).

The training set is used to forecast using the STL-Random Walk with Drift method. The time series is decomposed into the three components, and the seasonally adjusted data is forecasted using the random walk with drift method. The forecasted time series is then re-seasonalized by adding the seasonal naive forecasts of the seasonal component.

The forecasted data is visualized along with the test set. The residuals are also checked.

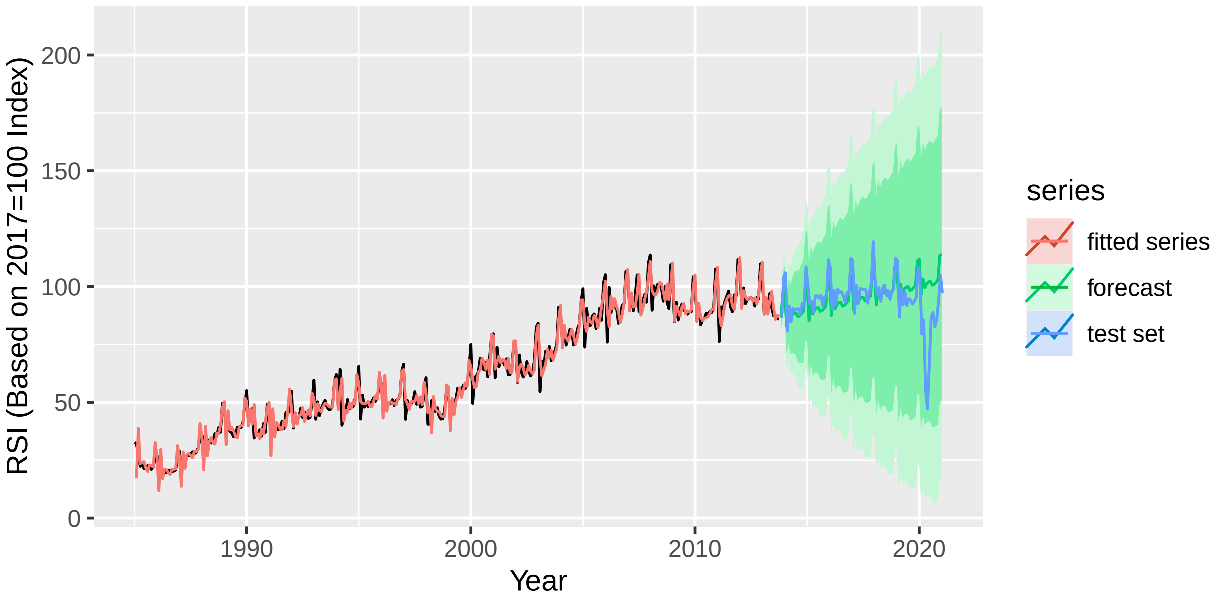
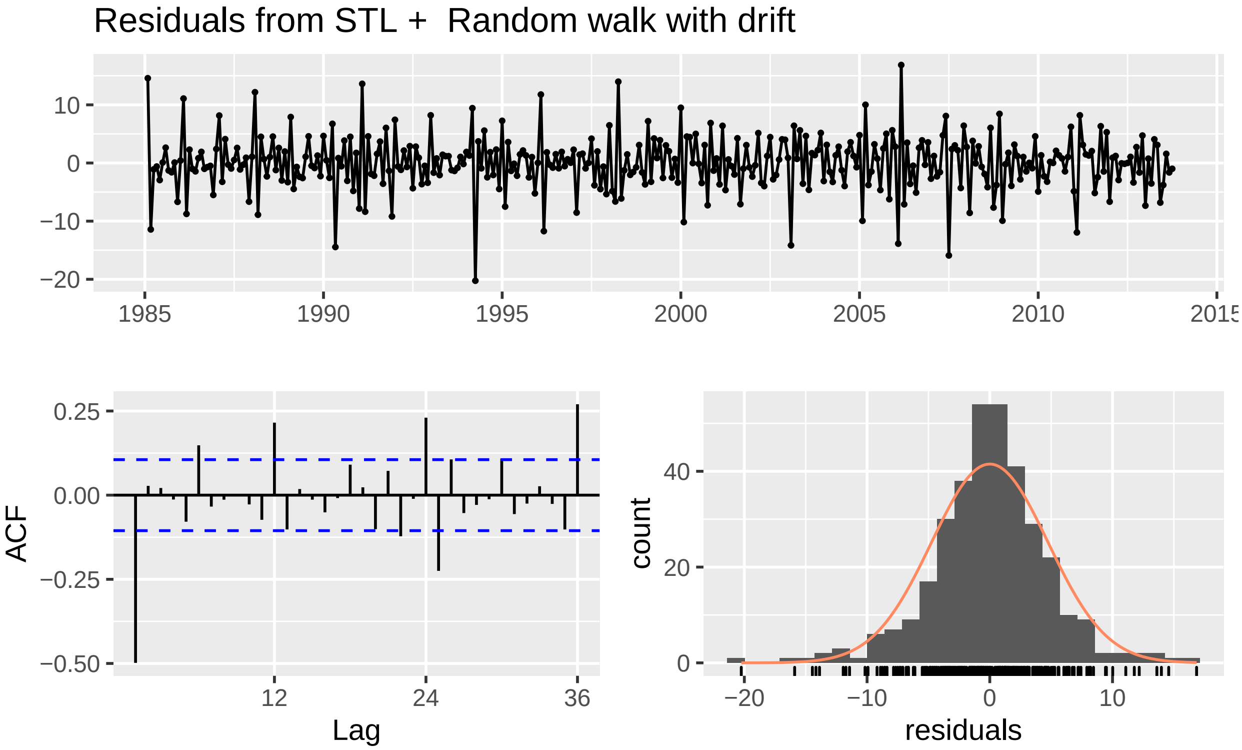


Figure 10. Forecasts from STL + Random Walk with drift.



```
##
##  Ljung-Box test
##
## data:  Residuals from STL +  Random walk with drift
## Q* = 155.31, df = 23, p-value < 2.2e-16
##
## Model df: 1.   Total lags used: 24
```

Figure 11. Residuals for forecasts from the STL + RW with Drift method.

- The residuals seem to be white noise as indicated by the time series plot. The model is unbiased.
- The low p-value (<0.05) in the Ljung-Box test allows us to reject the null hypothesis of no autocorrelation. Hence, the model shows serial correlation. This is confirmed by the ACF plot, which has a spike at the first lag. The model is inefficient, and the calculated prediction intervals are large.
- The residuals have constant variance, but are not normally distributed. There will be some difficulty in calculating the prediction intervals.

The tsCV RMS errors are calculated and stored for use in the later sections.

```
## STL + RW with Drift 5.24388
```

Exponential Smoothing Models

Unsatisfactory Methods

The ETS models tested include Simple Exponential Smoothing, Holt's linear trend, Holt's damped trend, Holt-Winters damped methods. As these methods do not account for the seasonality, their results will not be satisfactory (see Appendix).

Holt-Winters Seasonal Method

The Holt-Winters seasonal method has one forecast equation and three smoothing equations, for the level, trend, and seasonal component, respectively. In the additive method, the seasonal component is expressed in absolute terms and in the multiplicative method, in relative terms.

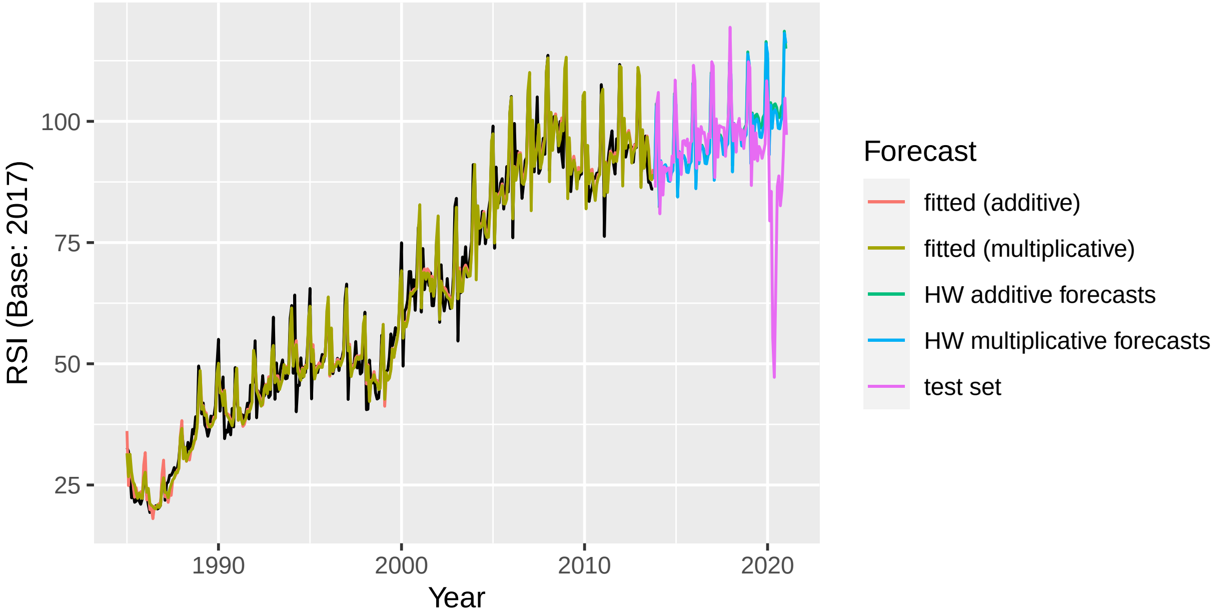


Figure 12. Forecasts from Holt−Winters seasonal methods.

The two models seem to have very similar performance, as the seasonal component of our time series approximately remains the same with an increase in level.

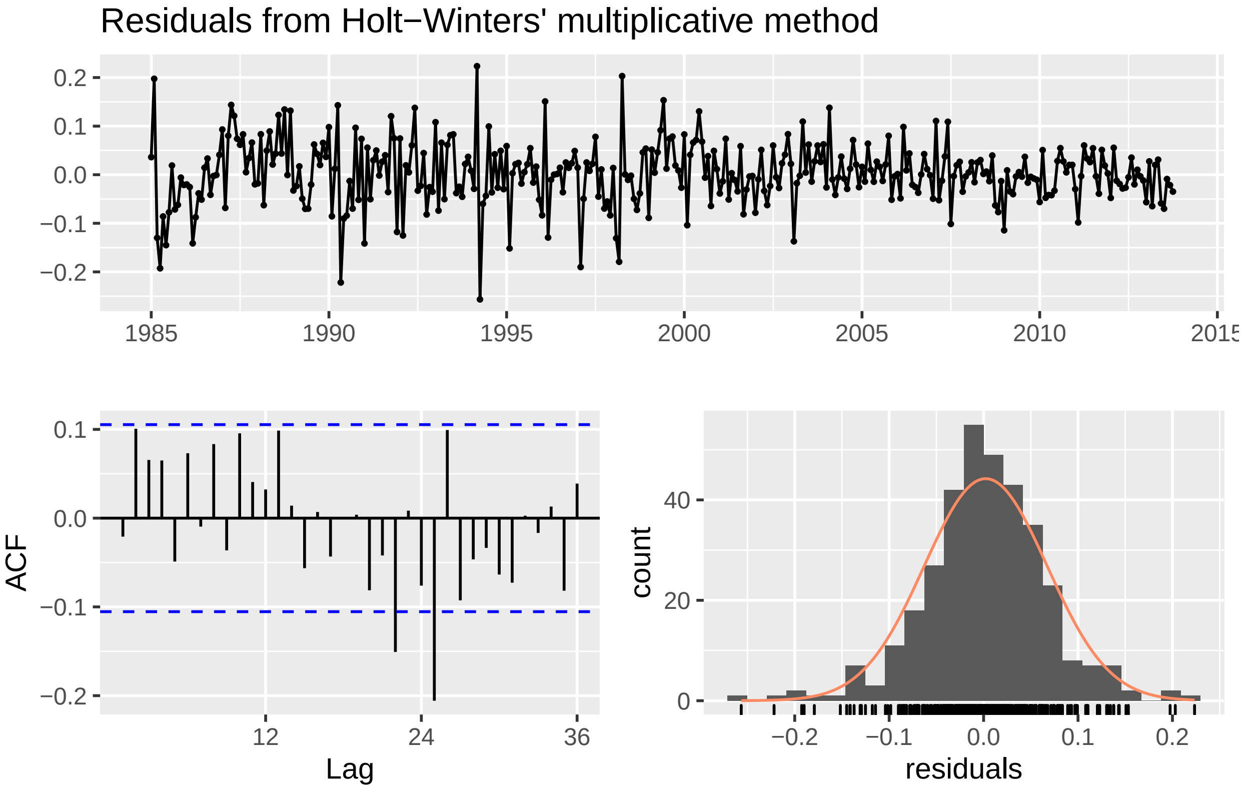
The tsCV RMS errors are calculated and stored for use in the later sections.

```
## Additive 4.554754
```

```
## Multiplicative 4.553761
```

The Holt-Winters multiplicative method has a slightly lower error than the additive method, which could be due to the fact that there is very small variation in the seasonal component over time that this model is able to capture.

We check its residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from Holt-Winters' multiplicative method
## Q* = 35.843, df = 8, p-value = 1.876e-05
##
## Model df: 16.   Total lags used: 24
```

Figure 13. Residuals for forecasts from the Holt-Winters seasonal methods.

- The residuals seem to have zero mean as indicated by the time series plot. The model is unbiased.
- The low p-value (<0.05) in the Ljung-Box test allows us to reject the null hypothesis of no autocorrelation. Hence, the model shows serial correlation. This is confirmed by the ACF plot, which has a spike at lags 22 and 25. The model is inefficient, and the calculated prediction intervals are large.
- The residuals are not normally distributed, and do not have constant variance. There will be a difficulty in calculating the prediction intervals.

Optimal ETS

The ETS model considers error, trend and seasonality components, as before. The *ets* function is used to pick the optimal ETS model.

```
## $method
## [1] "ETS(A,Ad,A)"
```

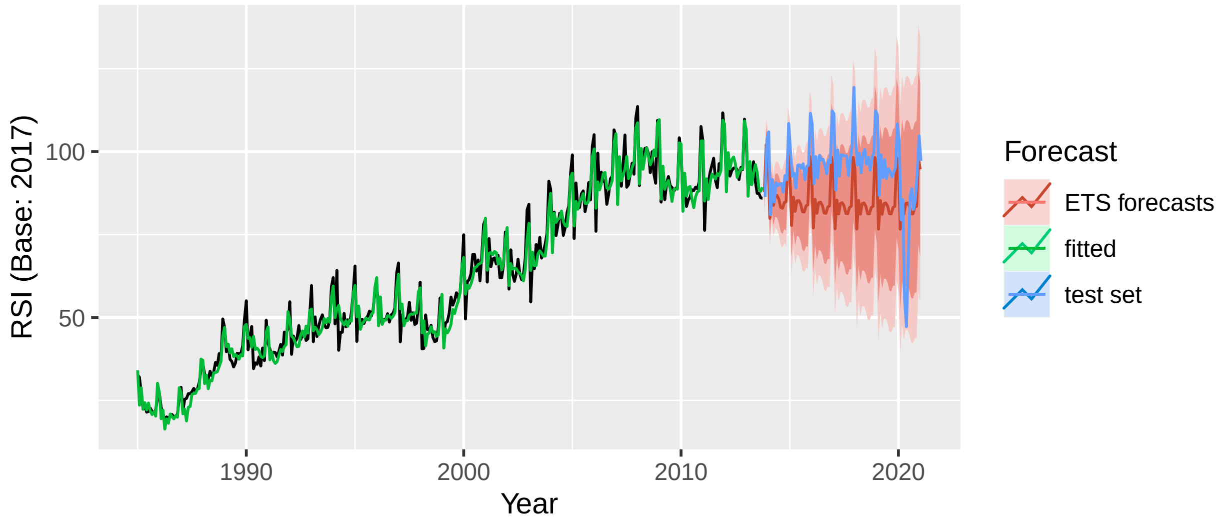
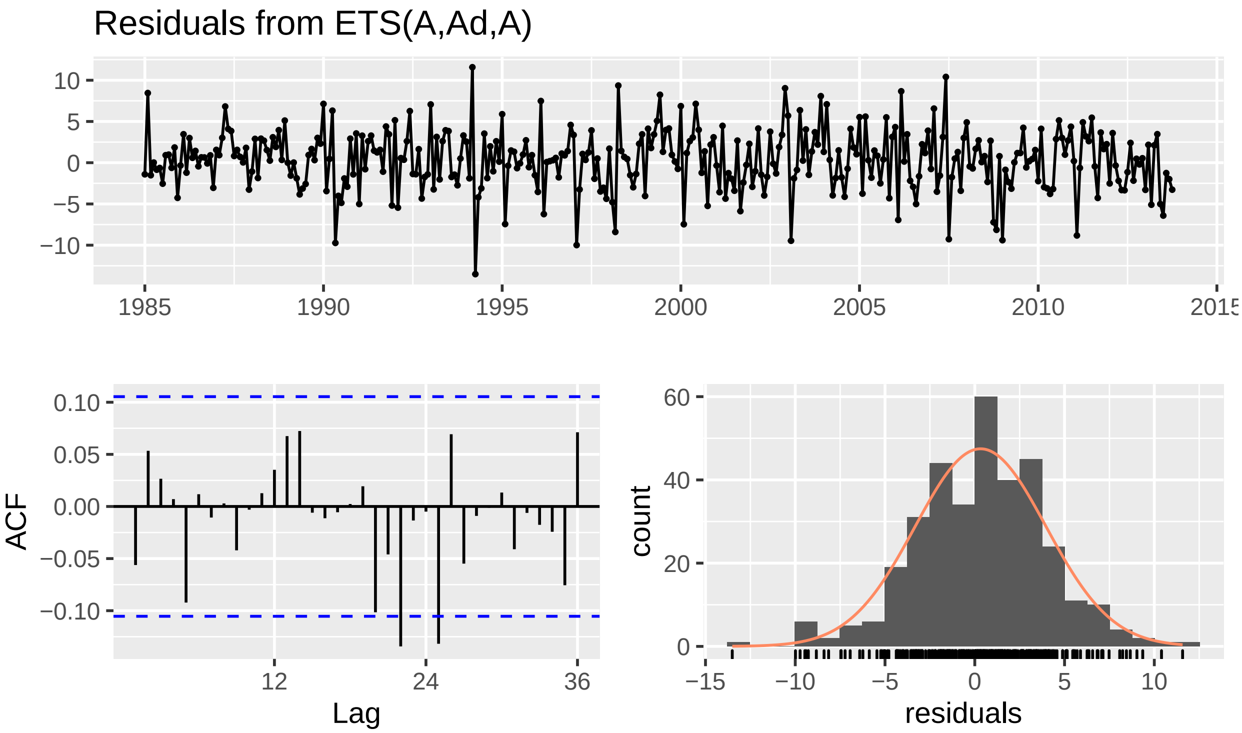


Figure 14. Forecasts from ETS(A,Ad,A).

The optimal ETS model is found to be ETS(A,Ad,A), which means that the error and seasonal components are additive, and the trend is additive damped.

We also check the residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(A,Ad,A)
## Q* = 21.722, df = 7, p-value = 0.002836
##
## Model df: 17.   Total lags used: 24
```

Figure 15. Residuals for forecasts from ETS(A,Ad,A).

The residual diagnostics are similar to those of the Holt-Winters multiplicative method:

- The model is unbiased.
- The model is inefficient as the model shows serial correlation at lags 22 and 25.
- The calculation of prediction intervals is difficult.

The tsCV RMS errors are calculated and stored.

```
## ETS(A,Ad,A) 4.5941
```

STL-ETS Method

The training set is used to forecast using the STL-ETS method.

The time series is decomposed into the three components, and the seasonally adjusted data is forecasted using the optimal ETS model. The forecasted time series is then re-seasonalized by adding the seasonal naive forecasts of the seasonal component.

We plot the RSI series along with the fitted series, as well as the forecasts with their prediction intervals.

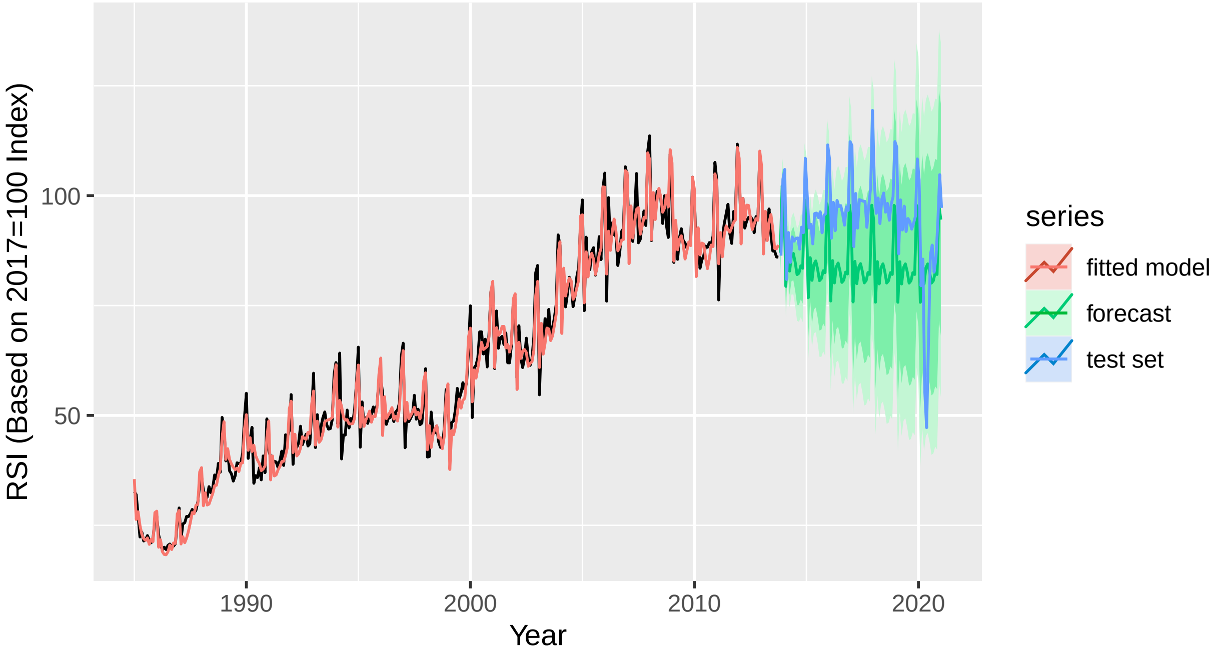
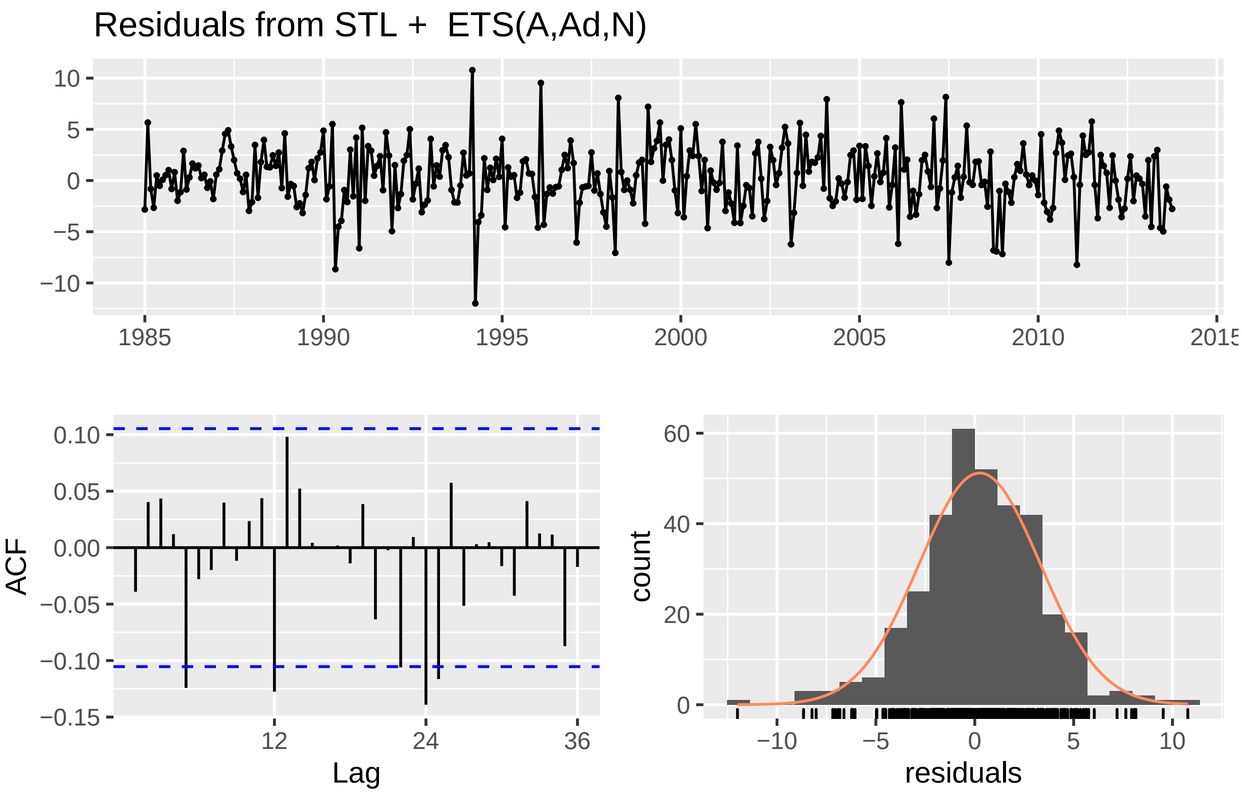


Figure 16. Forecasts from STL + ETS method.

The model seems to fare slightly better than just the ETS method.

We check the residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from STL +  ETS(A,Ad,N)
## Q* = 33.087, df = 19, p-value = 0.02349
##
## Model df: 5.   Total lags used: 24
```

Figure 17. Residuals for forecasts from STL + ETS method.

The residual diagnostics give similar conclusions as those of the ETS model.

The tsCV RMS error is calculated and stored.

```
## STL + ETS 4.531774
```

Comparing Exponential Smoothing Models

The tsCV RMS errors calculated for the different Exponential Smoothing models are compared.

```
##                              tsCV RMSE
## Holt-Winters Additive         4.554754
## Holt-Winters Multiplicative   4.553761
## ETS                           4.594100
## STL-ETS                       4.531774
```

Overall, the STL-ETS model has the best performance, followed by Holt-Winters Multiplicative model.

ARIMA Models

Differencing the Time Series

The KPSS test is performed.

```
## 
## ####################### 
## # KPSS Unit Root Test # 
## ####################### 
## 
## Test is of type: mu with 5 lags. 
## 
## Value of test-statistic is: 6.7404 
## 
## Critical value for a significance level of: 
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

The test statistic of 6.7 is larger than the 1% critical value of 0.739, implying that the null hypothesis of stationarity is rejected. Hence, RSI is not stationary. The data are differenced and the KPSS test is performed another time.

```
## 
## ####################### 
## # KPSS Unit Root Test # 
## ####################### 
## 
## Test is of type: mu with 5 lags. 
## 
## Value of test-statistic is: 0.0227 
## 
## Critical value for a significance level of: 
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

The test statistic of 0.0227 is much smaller than the 1% critical value, implying that the differenced data are stationary.

The function *ndiffs* is used to perform this process of sequentially carrying out KPSS tests to find the appropriate number of **seasonal** differences.

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

As *nsdiffs* returns 1, implying that one seasonal difference is needed, seasonal differencing is applied to the first differenced data we had obtained earlier, and the data are run through the *ndiffs* function again.

```
## [1] 0
```

The value of 0 indicates that the data are stationary and no further differencing is required.

Hence, first differencing and seasonal differencing are applied to the time series, and the PACF and ACF plots are checked.

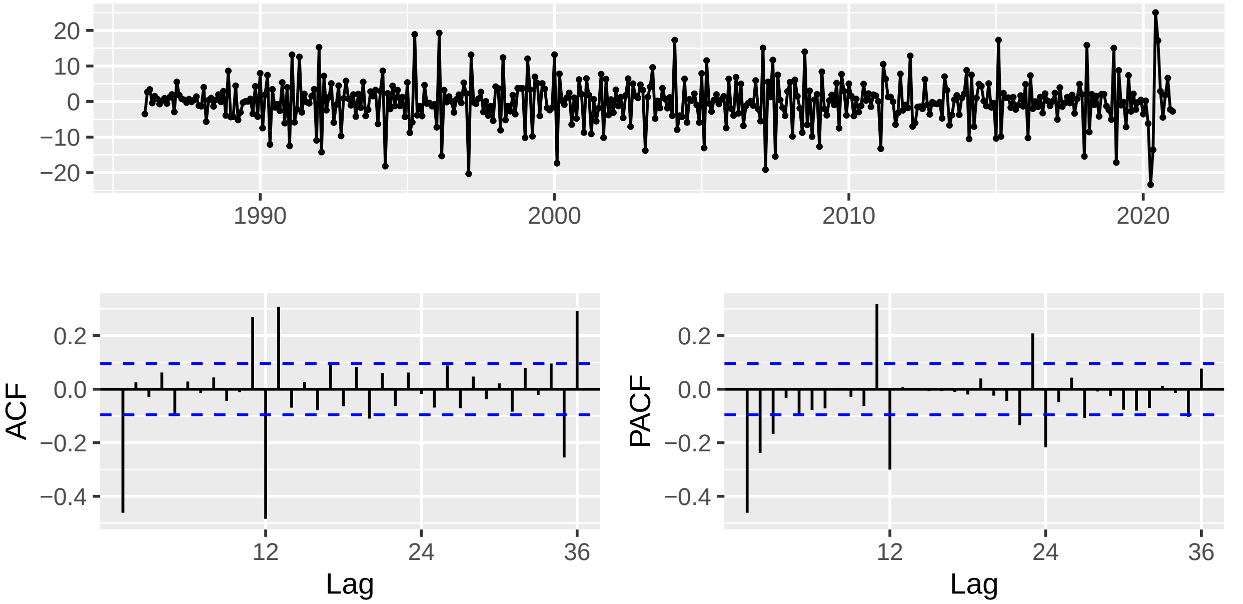


Figure 18. ACF and PACF plots for the first and seasonal differenced data.

For the non-seasonal part, the ACF shows significant spikes at the first lag, and the PACF shows significant spikes at the first, second, and third lags. This is suggestive of either an MA(1) component or an AR(3) component.

For the seasonal part, there are spikes in the PACF at lags 12 and 24, but nothing at seasonal lags in the ACF. This suggests a seasonal AR(2) component.

Fitting ARIMA Models

This initial analysis above suggests that some possible models for the data are ARIMA(3,1,0)(2,1,0)[12], ARIMA(2,1,0)(2,1,0)[12], ARIMA(1,1,0)(2,1,0)[12] or ARIMA(0,1,1)(2,1,0)[12]. We fit the first model, and display the ACF and PACF plots for the residuals.

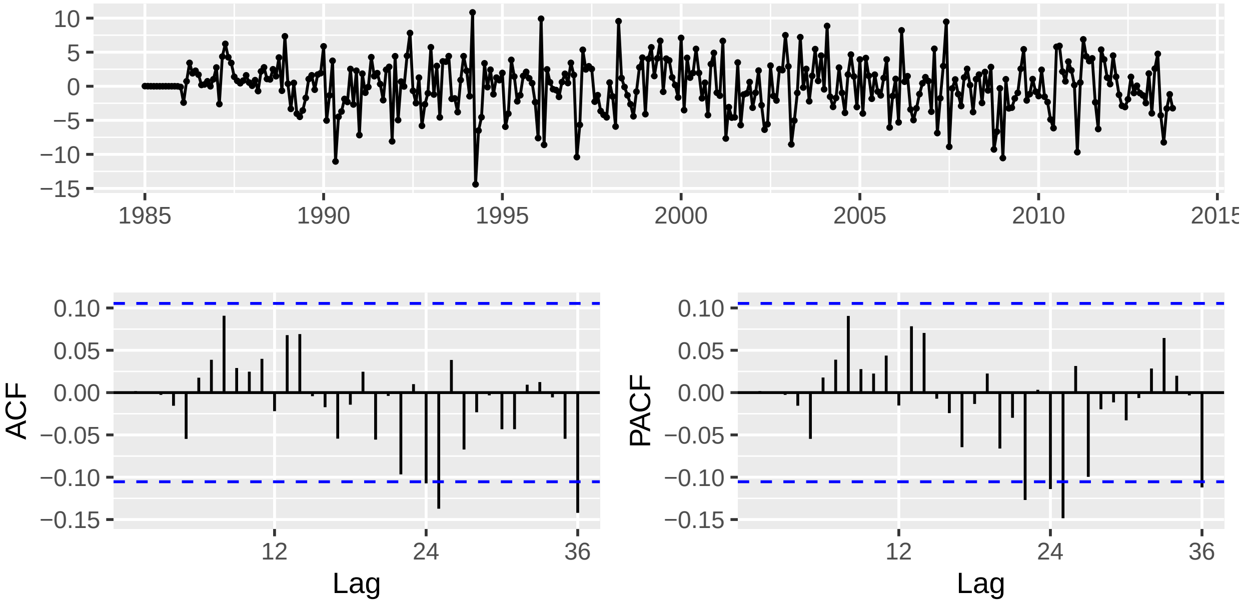


Figure 19. PACF and ACF plots for the ARIMA(3,1,0)(2,1,0)[12] model.

The model is not perfect as both the ACF and PACF plots have significant spikes. Alternatively, we use *auto.arima* to double-check the initial analysis and fit the ideal model.

```
## $method
## [1] "ARIMA(1,0,2)(0,1,1)[12] with drift"
```

In this case, *auto.arima* found an ARIMA(1,0,2)(0,1,1)[12] with drift model.

The discrepancy could be due to the fact that the full time series was used to check for appropriate differencing, but the model was only trained on the training set.

We check the ACF and PACF plots of the residuals.

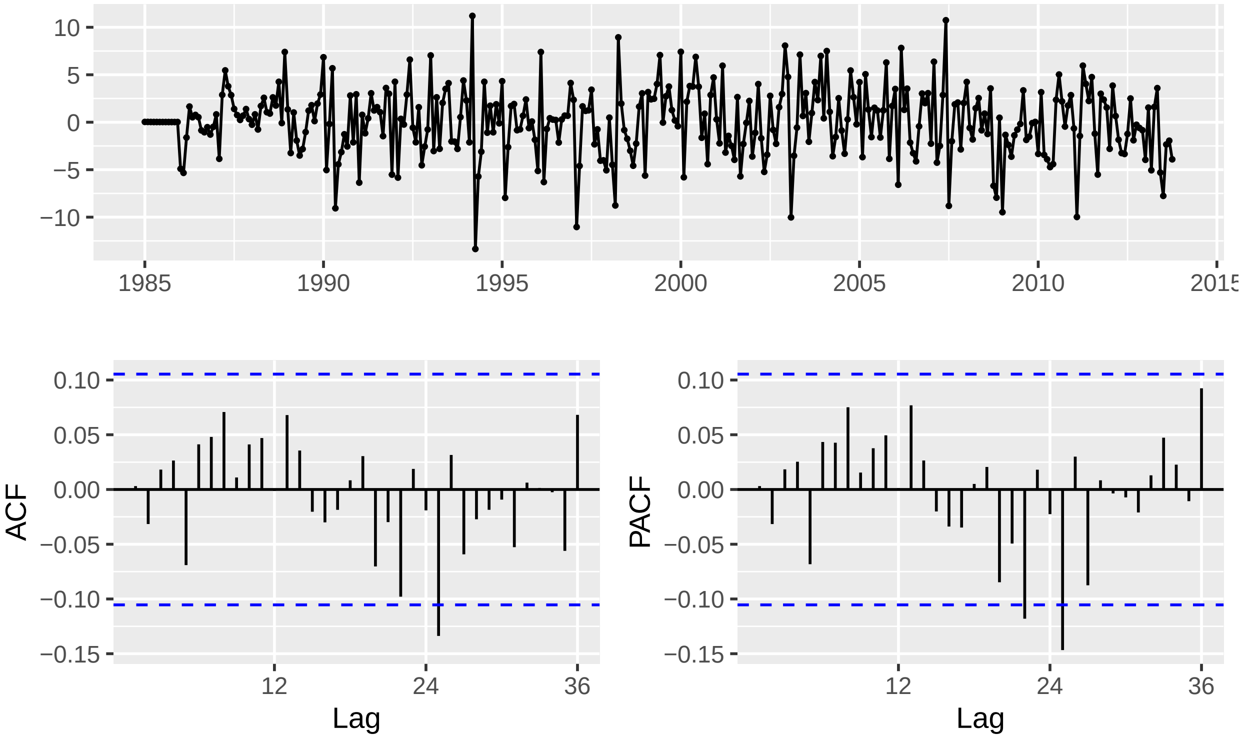


Figure 20. PACF and ACF plots for the ARIMA(1,0,2)(0,1,1)[12] model with drift.

Although the ACF and PACF plots still show significant spikes, the magnitude is smaller than that of the initial model fitted. This points towards the fact that the *auto.arima* model may be a better fit than the manually selected model. This can be further confirmed by plotting the RSI series along with the two ARIMA models' fitted values and forecasts.

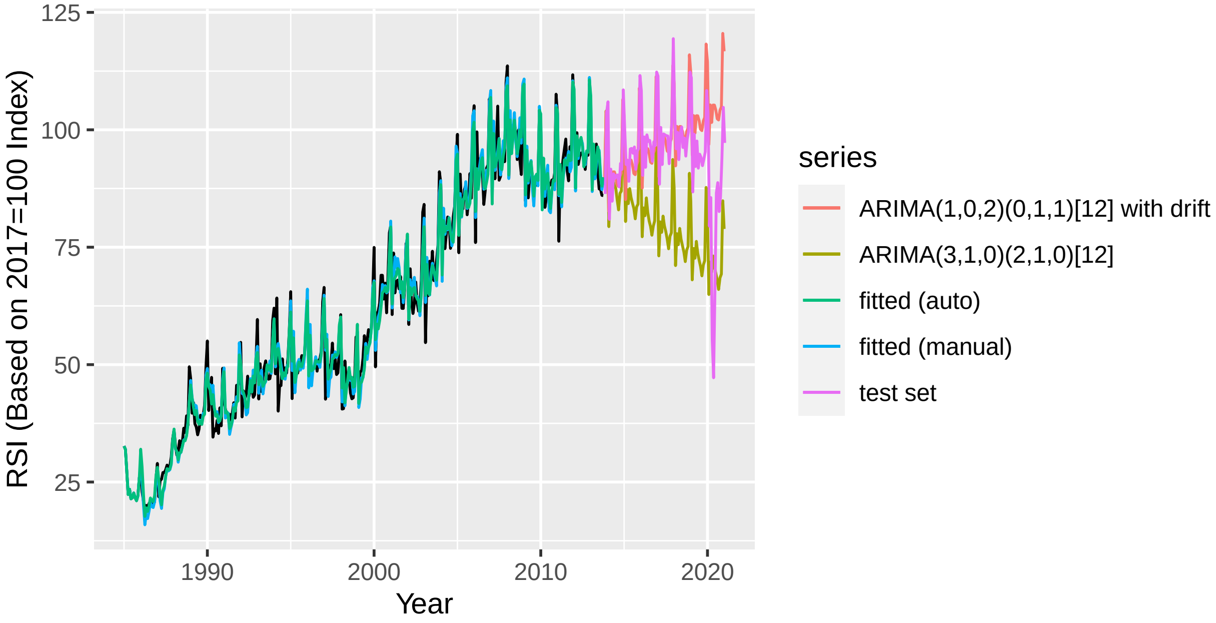


Figure 21. Forecasts from ARIMA models.

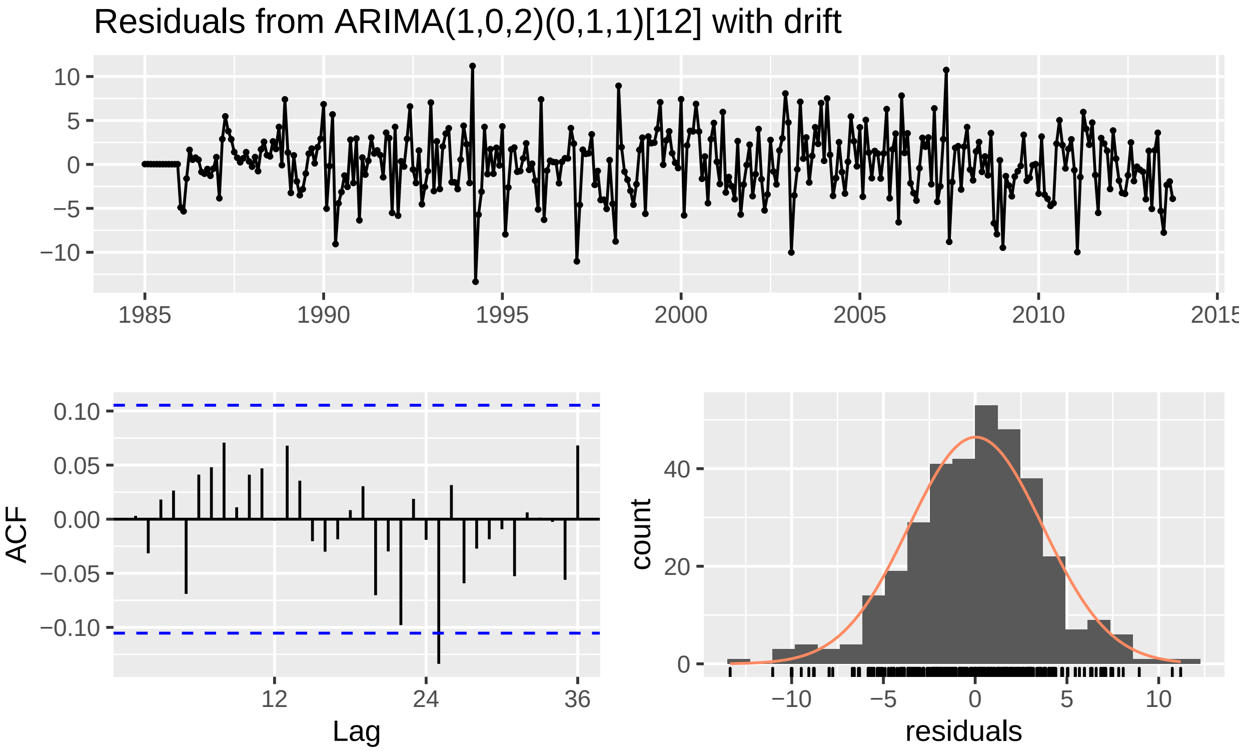
The *auto.arima* model seems to be a better fit. We cannot compare the AICC as the two models have different differencing (d). Hence, we compare the tsCV RMS errors and store them.

```
## ARIMA(1,0,2)(0,1,1)[12] with drift 4.510014
```

```
## ARIMA(3,1,0)(2,1,0)[12]             4.719721
```

The tsCV RMSE confirms that the *auto.arima* model is a better fit.

We now comment on its residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,2)(0,1,1)[12] with drift
## Q* = 29.085, df = 31, p-value = 0.5648
##
## Model df: 5.   Total lags used: 36
```

Figure 22. Residuals for forecasts from the ARIMA(1,0,2)(0,1,1)[12] model with drift.

- The residuals seem to have zero mean as indicated by the time series plot. The model is unbiased.
- As the p-value is high (>0.05), we are unable to reject the null of no autocorrelation. The model is efficient.
- The residuals are not normally distributed. There will be difficulty in calculating the prediction intervals.

Hence, the better model here is ARIMA(1,0,2)(0,1,1)[12] with drift, selected using *auto.arima*.

STL-ARIMA model

The STL-ARIMA model performs decomposition, forecasts the seasonally adjusted series using an appropriate ARIMA model, and re-seasonalizes the forecast by adding seasonal naive forecasts of the seasonal component.

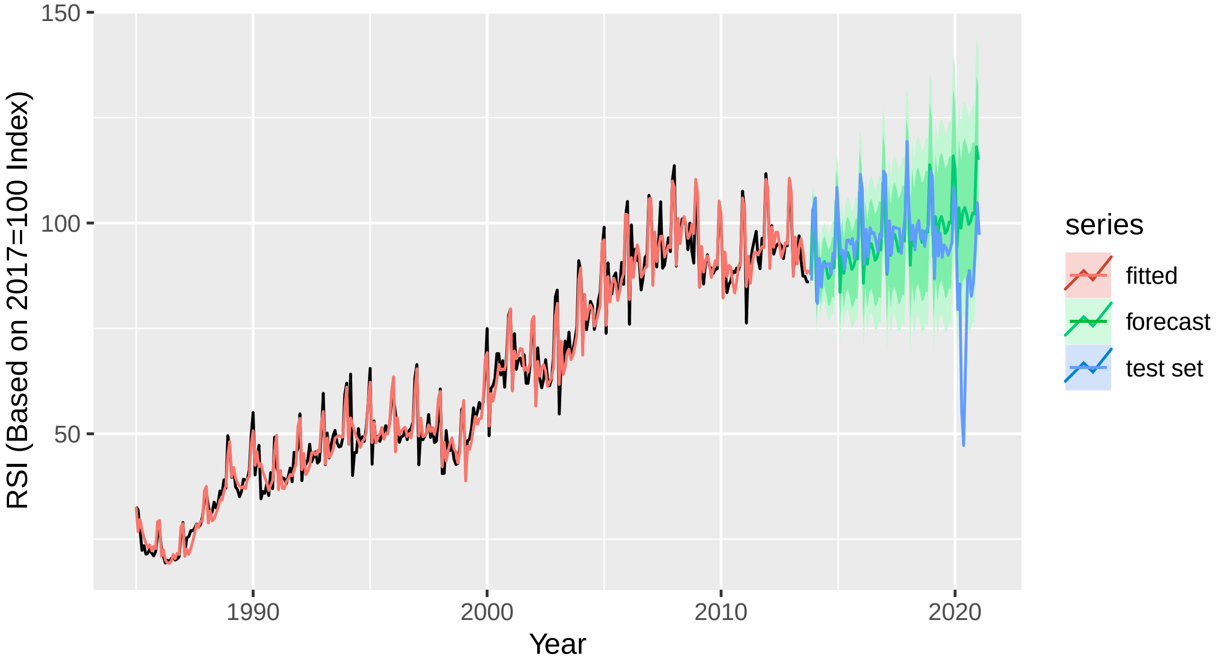
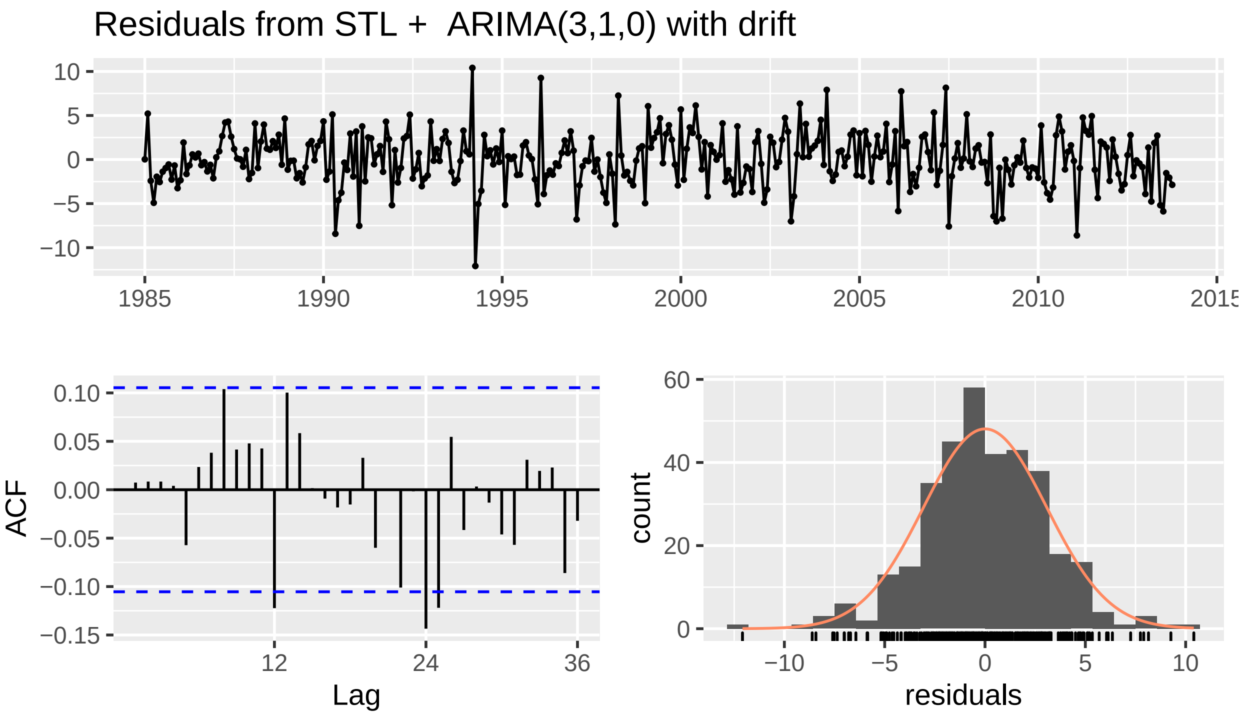


Figure 23. Forecasts from STL−ARIMA model.

At first glance, the model seems to fare more or less similar to the *auto.arima* model.

We check the residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from STL +  ARIMA(3,1,0) with drift
## Q* = 31.609, df = 20, p-value = 0.04765
##
## Model df: 4.   Total lags used: 24
```

Figure 24. Residuals for forecasts from STL-ARIMA model.

- The residuals have zero mean and hence the model is unbiased.
- The model is inefficient as the residuals show serial correlation.
- The residuals are not normally distributed and hence the calculation of prediction intervals is difficult.

The tsCV RMS errors are calculated and stored.

```
## STL-ARIMA 4.581719
```

Comparing ARIMA Models

The tsCV RMS errors are compared for the different ARIMA and STL-ARIMA models considered above.

```
##                           tsCV RMSE
## ARIMA(3,1,0)(2,1,0)[12]    4.719721
## Auto ARIMA                 4.510014
## STL-ARIMA                  4.581719
```

On comparing the error values, we see that auto ARIMA model has the lowest error and is the best model out of the three.

Regression

Linear Regression

Model

We perform linear regression, with the RSI as the dependent variable. The predictor variables are:

- Per Capita Gross National Income (Per Capita GNI, 2019). This has a direct impact on the purchasing power of the consumers, and hence may affect the RSI. The dataset has yearly frequency, and we use cubic spline interpolation to convert it to monthly frequency. The time series is available from 1985 to 2018.
- Number of Residents (Singapore Residents, 2020). As we are considering the *per capita* GNI, we should also consider the number of residents in Singapore. The dataset has an annual frequency and consists of the number of Singapore residents decomposed by age group, ethnic group, and sex. We first use Pivot tables on Excel to aggregate the data for each year across age group, ethnic group and sex, and then use cubic spline interpolation to convert it to monthly frequency. The time series is available from 1985 to 2019.
- Average CPF Contribution Rates (CPF Contribution Rates, 2017). The CPF contribution rates directly affect the take-home income the residents have, given the same salary/wage. The dataset has CPF contribution rates for the different "effective from" dates, decomposed by age group and contributing party (employee/employer). We first use Pivot tables on Excel to aggregate the data for each "effective from" date by taking the average contribution rate across age groups and contributing party. Then, we transform the data into a time series with an annual frequency. Finally, we use cubic spline interpolation to convert it to monthly frequency. The time series is available from 1985 to 2021.
- Total Certificate of Entitlement (COE) Quota (COE Trends, 2021). The RSI also includes Motor Vehicle Sales, which is directly affected by the CPF quota for that period. The dataset we use has monthly data on COE quota decomposed by bid number (there are two bids each month) and category of vehicle (A,B,C,D or E). For the sake of simplicity, we use Pivot tables on Excel to aggregate the monthly data by the total number of quotas across both bids and all vehicle classes. Some additional formatting of the date string had to be done using Excel functions to allow it to be sorted in ascending order of date and be passed to R. As the current system of two open bidding exercises per month was introduced in 2002, our dataset is from April 2002 to March 2021.
- Month. We add dummy variables for each month to account for the seasonality present in the RSI.

We select a time period common to all the predictors, which is April 2002 to January 2018. We take windows for all variables.

Performing the regression,

```
##
## Call:
## tslm(formula = RSI.reg ~ GNI.reg + Pop.reg + CPF.reg + COE.reg +
##     season, lambda = "auto")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1036.53  -275.45   -42.13   240.08  1315.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.675e+03  1.022e+03   1.639   0.1031
## GNI.reg      3.201e-02  1.300e-02   2.462   0.0148 *
## Pop.reg      2.051e-03  4.913e-04   4.175 4.70e-05 ***
## CPF.reg     -6.656e+02  9.646e+01  -6.900 9.29e-11 ***
## COE.reg      5.409e-02  1.104e-02   4.898 2.21e-06 ***
## season2     -1.551e+03  1.466e+02 -10.580  < 2e-16 ***
## season3     -8.244e+02  1.466e+02  -5.623 7.35e-08 ***
## season4     -1.252e+03  1.446e+02  -8.660 3.14e-15 ***
## season5     -9.475e+02  1.444e+02  -6.561 5.89e-10 ***
## season6     -9.055e+02  1.444e+02  -6.272 2.74e-09 ***
## season7     -9.955e+02  1.443e+02  -6.897 9.42e-11 ***
## season8     -1.201e+03  1.444e+02  -8.323 2.43e-14 ***
## season9     -1.244e+03  1.443e+02  -8.619 4.03e-15 ***
## season10    -1.056e+03  1.442e+02  -7.321 8.73e-12 ***
## season11    -1.065e+03  1.442e+02  -7.382 6.16e-12 ***
## season12     2.959e+02  1.442e+02   2.052   0.0417 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 407.9 on 174 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.7869
## F-statistic: 47.52 on 15 and 174 DF,  p-value: < 2.2e-16
```

We also visualize the fitted series along with the RSI.

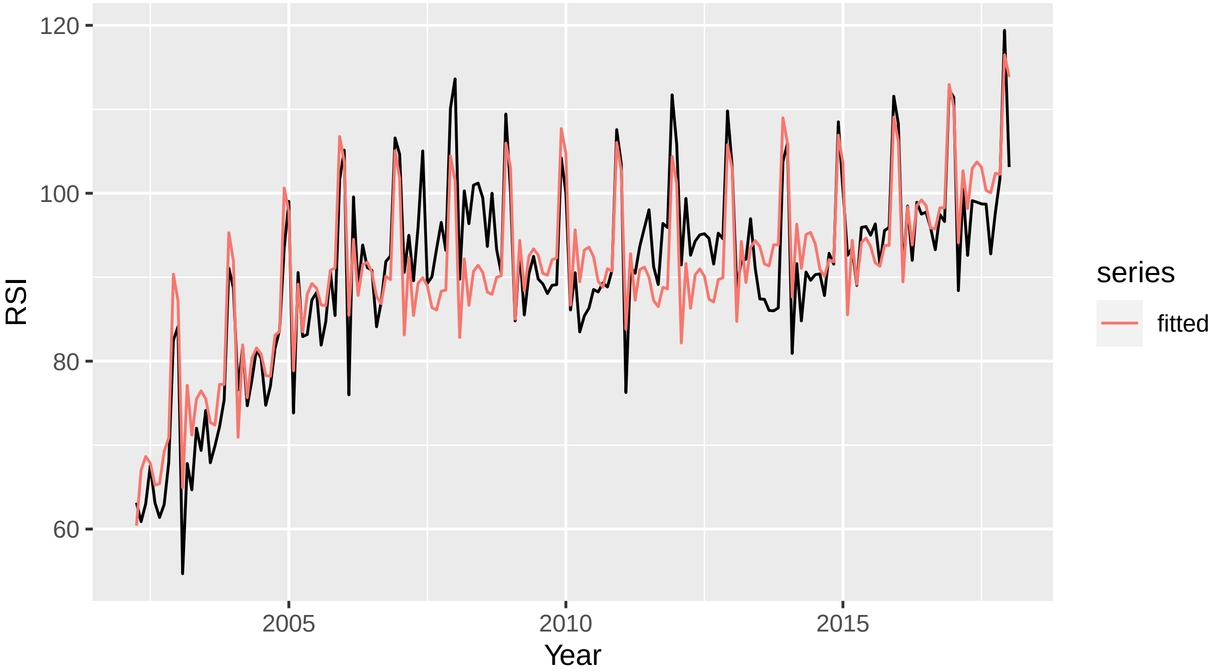


Figure 25. Fitted model for linear regression.

The model seems to be a good fit and has an R-squared value of 0.7869.

Residual and Other Diagnostics

We check the correlations of the different variable pairs.

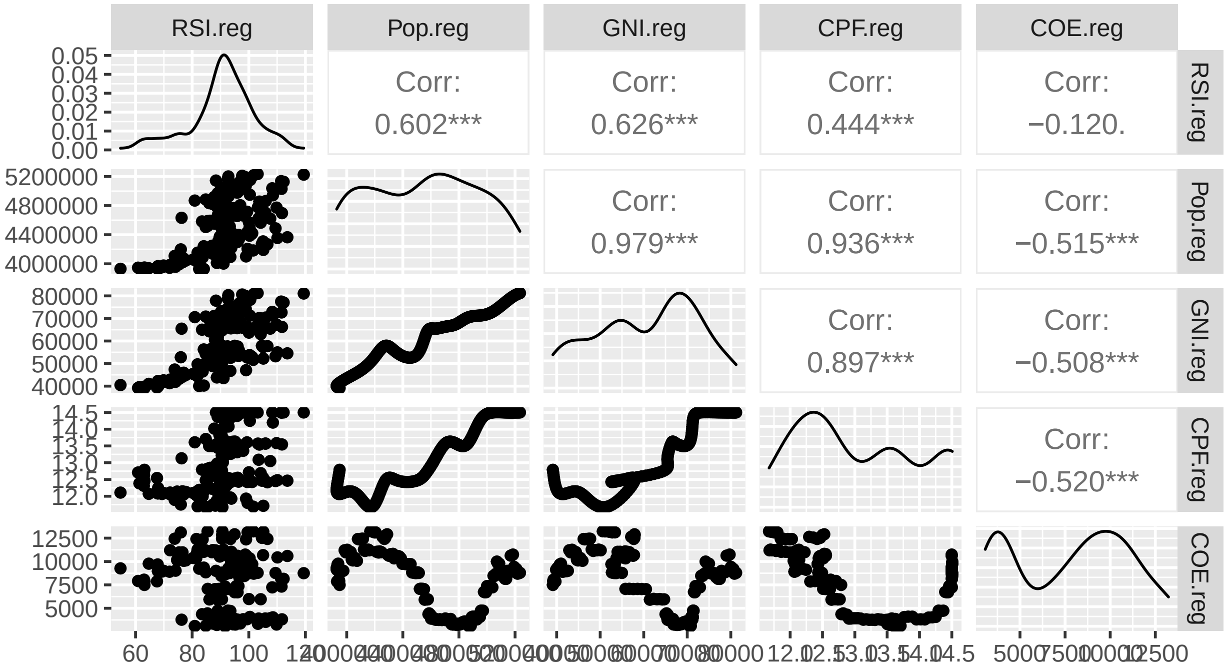
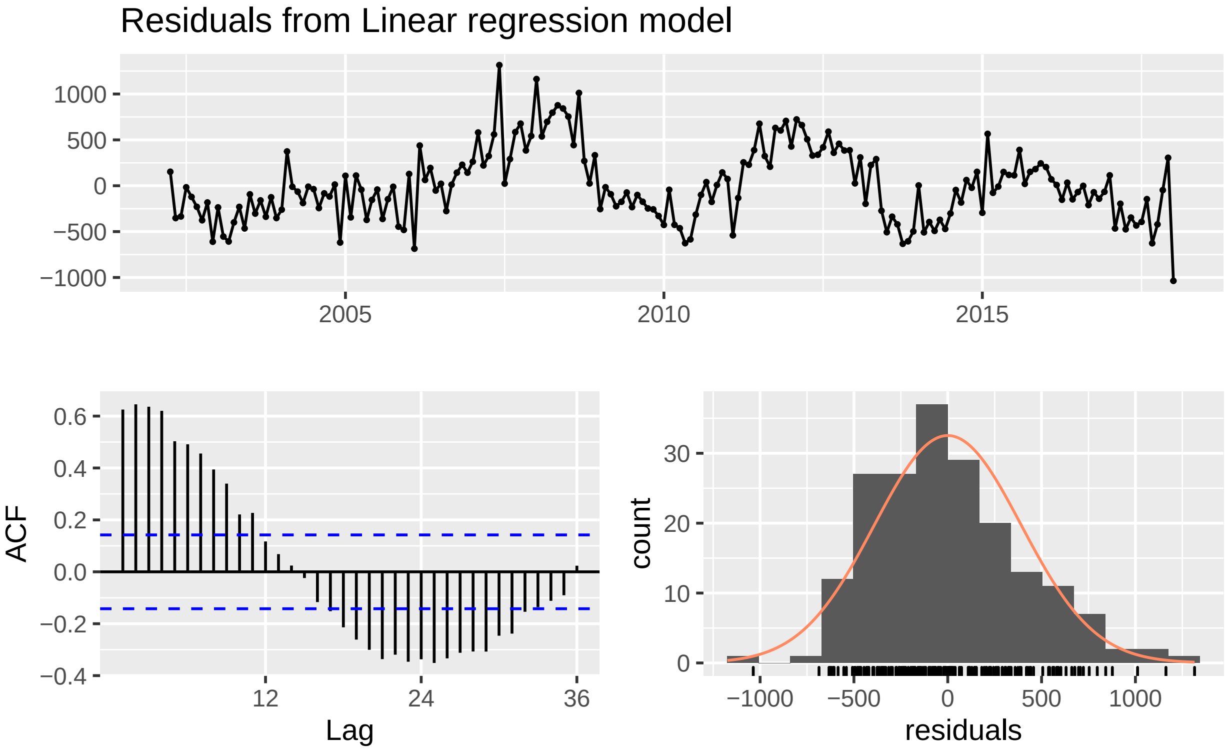


Figure 26. Correlation coefficients between all variable pairs.

- We observe that the dependent variable is not very highly correlated with any of the predictor variables. This means that a single predictor is not enough to properly predict the RSI, and a combination of predictors is required.
- Some of the predictors are highly correlated with each other. This multicollinearity is not a problem here as our focus is on forecasting the dependent variable.

Next we perform the residual diagnostics.



```
##
##  Breusch-Godfrey test for serial correlation of order up to 24
##
## data:  Residuals from Linear regression model
## LM test = 119.92, df = 24, p-value = 1.004e-14
```

Figure 27. Residuals for linear regression.

- As the mean of the residuals is not very clear from the time plot, we use the mean() function to check.

```
## [1] 1.268576e-14
```

The mean seems to be close to 0. This means that the forecasts are unbiased.

- From the Breusch-Godrey test, as the p-value is smaller than 0.05, we reject the null of no autocorrelation in favour of the alternate hypothesis of serial correlation, at the 5% significance level. That is, the residuals show autocorrelation. Hence, the model is inefficient and some other information can be included to make the forecasts better. Additionally, the prediction intervals will be bigger.
- The residuals are not normally distributed. This makes the calculation of prediction intervals more difficult.

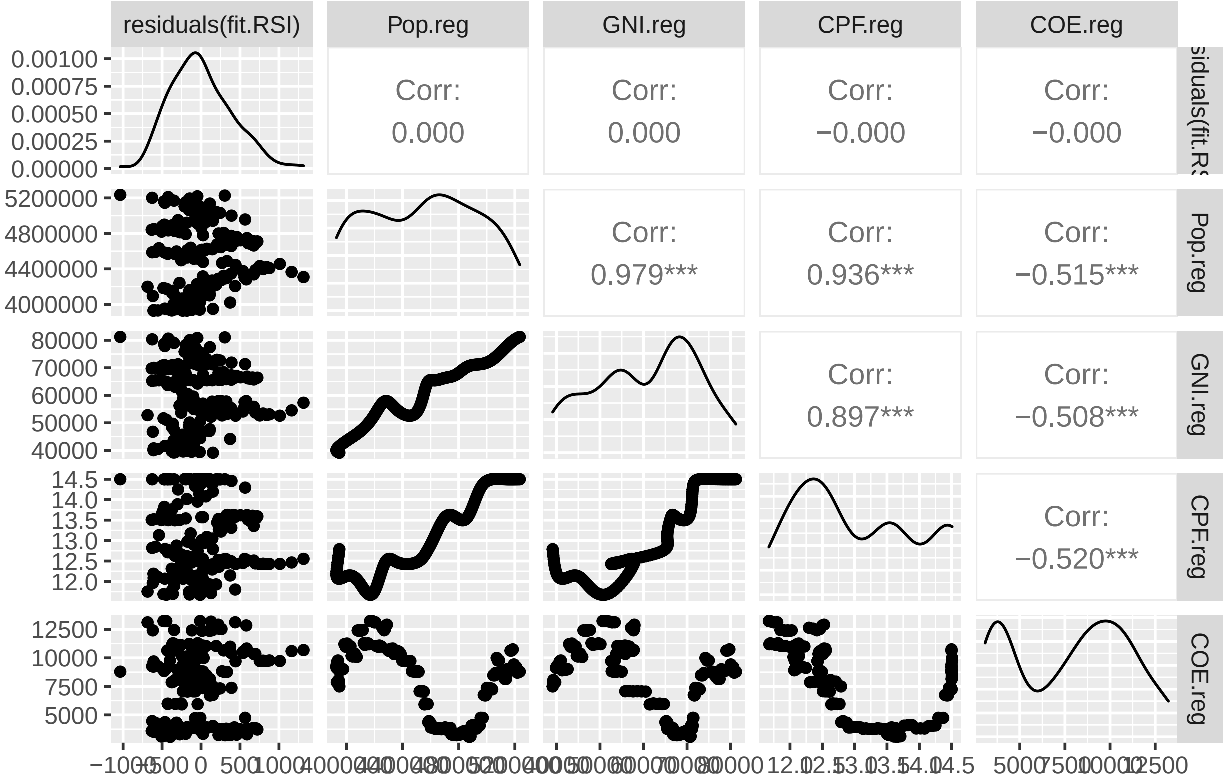


Figure 28. Correlation coefficients for residuals and different regressors.

- From the first row, we infer that the correlation between the residuals and the predictor variables is also zero. This means that the model doesn't suffer from the problem of endogeneity. To fix the problem of autocorrelation, we look at dynamic regression.

Dynamic Regression

Regression with ARIMA Errors

In this model, the errors from the regression are allowed to contain autocorrelation, and are assumed to follow an ARIMA model.

The *auto.arima* function selects the best ARIMA model for the error terms, and differences the variables if needed.

```
## Series: RSI.reg 
## Regression with ARIMA(0,1,1)(2,1,1)[12] errors 
## 
## Coefficients:
##           ma1     sar1     sar2     sma1   xreg
##       -0.6538  -0.1540  -0.0844  -0.7001  0e+00
## s.e.   0.0492   0.1558   0.1307   0.1534  1e-04
## 
## sigma^2 estimated as 12.53:  log likelihood=-478.55
## AIC=969.11   AICc=969.6   BIC=988.16
## 
## Training set error measures:
##                       ME    RMSE      MAE        MPE    MAPE      MASE
## Training set -0.2682378 3.36799 2.402054 -0.3341326 2.64122 0.4754063
##                     ACF1
## Training set -0.07625269
```

We also visualize the fitted series along with the RSI.

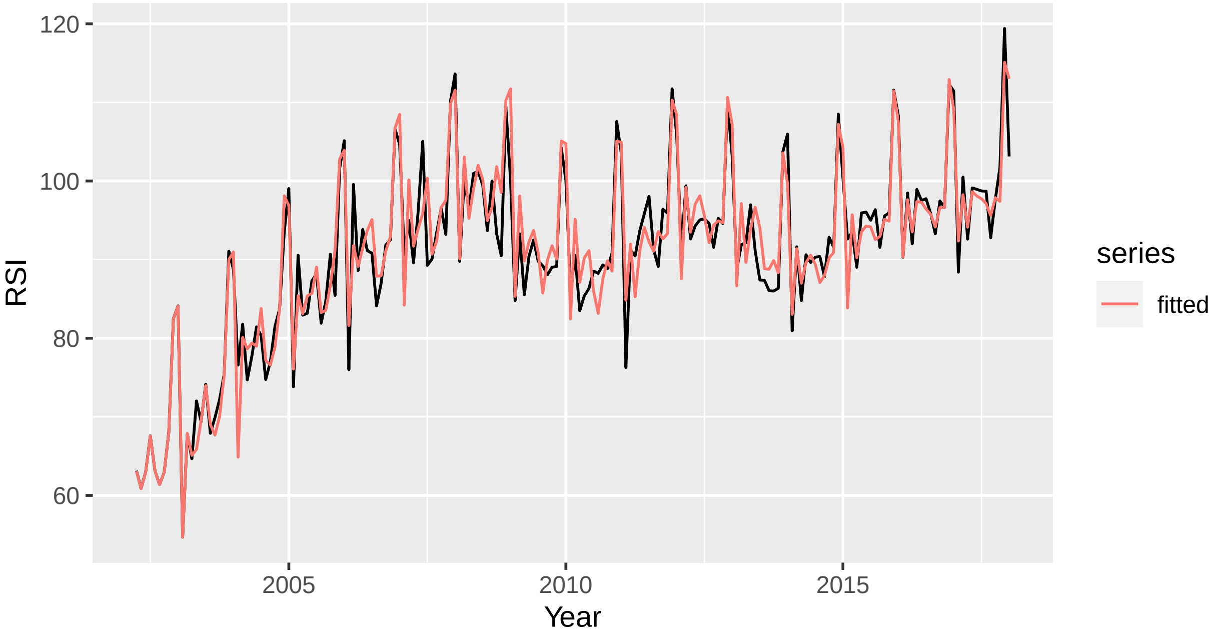
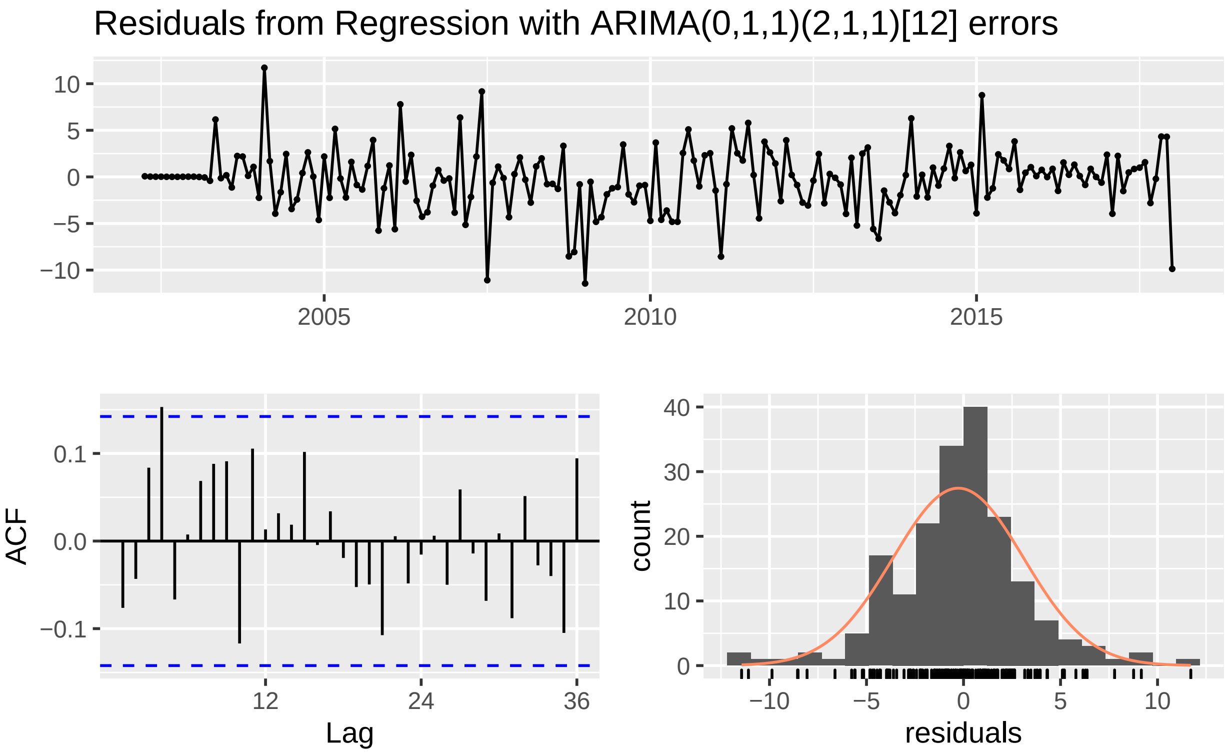


Figure 29. Fitted model for regression with ARIMA(0,1,1)(2,1,1)[12] errors.

The function uses an ARIMA(0,1,1)(2,1,1)[12] model for the errors. We check the residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,1)(2,1,1)[12] errors
## Q* = 24.519, df = 19, p-value = 0.177
##
## Model df: 5.   Total lags used: 24
```

Figure 30. Residuals for regression with ARIMA(0,1,1)(2,1,1)[12] errors.

- The Breusch-Godfrey test has a larger p-value > 0.1 this time, and we are unable to reject the null of no autocorrelation. This means that the residuals are not autocorrelated and the model is efficient. This is an improvement from the linear regression model.
- Additionally, the errors seem to be white noise.

Regression with ARIMA Errors and Lagged Predictors

After successful bidding for the COE, the bidder gets a temporary COE that is valid from 3-6 months depending on the category of the vehicle (Certificate of Entitlement (COE), n.d.). These can be used to purchase a vehicle, the sales of which are included in the RSI. Hence, we include lagged COE values for up to 6 months.

The fit the model with different number of lags and compare the AICC to pick the best model. Some observations in the beginning are skipped when we take lagged values.

```
##         Lag 0  Lag 0-1  Lag 0-2  Lag 0-3 Lag 0-4  Lag 0-5 Lag 0-6
## AICC 1159.663 1154.261 1150.746 1147.397 1144.48 1140.249 1137.04
```

It can be observed that the model with all 6 lagged values has the lowest AICC. Hence, we pick that one, and perform regression with *auto.arima* errors.

```
## Series: window(RSI.reg, start = c(2002, 11), end = c(2018, 1))
## Regression with ARIMA(1,0,0)(1,0,0)[12] errors
```

```
##
## Coefficients:
##          ar1    sar1  intercept    Lag0     Lag1     Lag2    Lag3    Lag4    Lag5
##       0.5530  0.8562    77.8086  0.0016  -2e-04  -3e-04  1e-04  3e-04  2e-04
## s.e.  0.0643  0.0398     6.2804  0.0010   1e-03   1e-03  1e-03  1e-03  9e-04
##         Lag6
##       -4e-04
## s.e.   9e-04
##
## sigma^2 estimated as 24.89:  log likelihood=-556.75
## AIC=1135.5   AICc=1137.04   BIC=1170.8
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE     MASE       ACF1
## Training set 0.7950207 4.850609 3.663779 0.5985904 4.093691 0.735982 -0.2663566
```

The function picks an ARIMA(1,0,0)(1,0,0)[12] model for the errors.

We also visualize the fitted series along with the RSI.

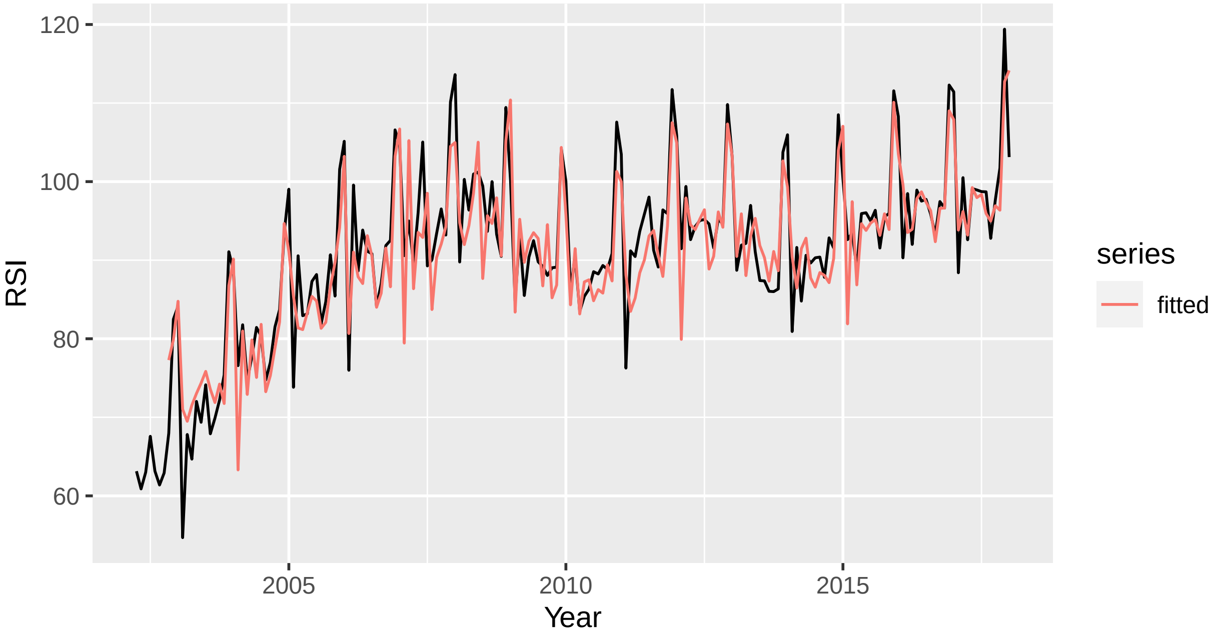
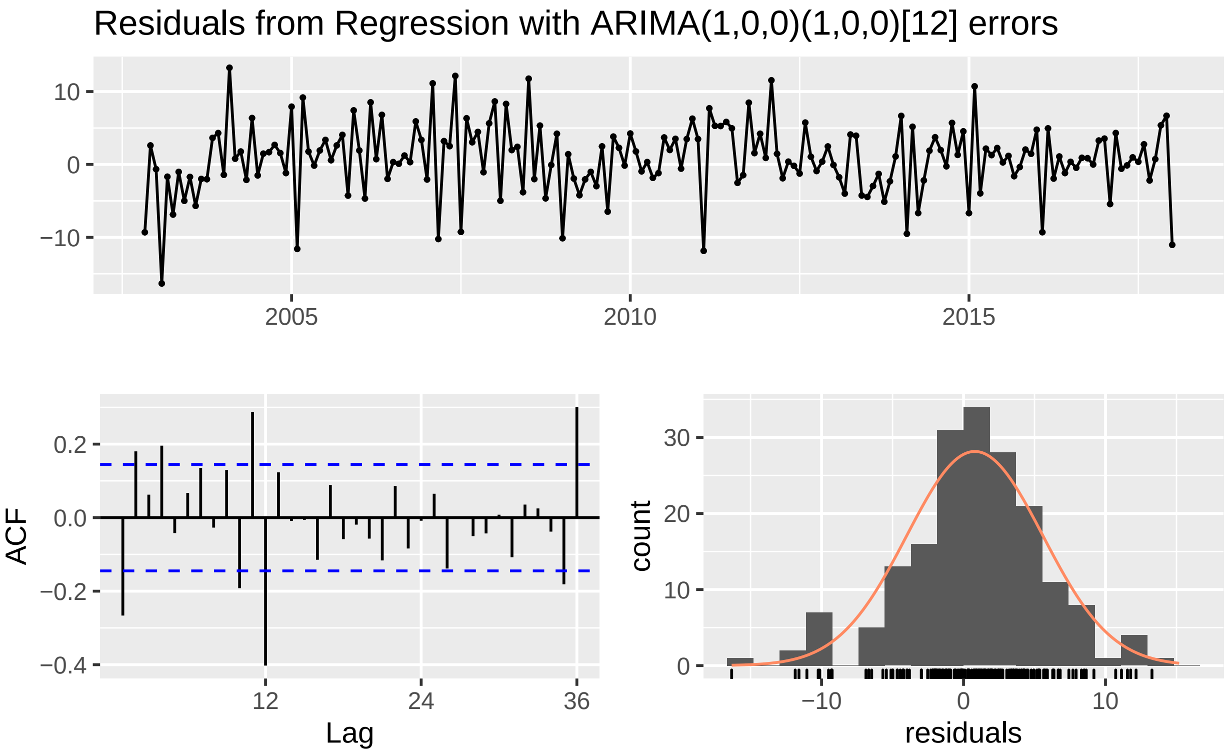


Figure 31. Fitted model for regression with ARIMA errors and lagged predictors.

At first glance, the models seems to be a worse fit than the model with just ARIMA errors.



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,0,0)(1,0,0)[12] errors
## Q* = 105.65, df = 14, p-value = 3.331e-16
##
## Model df: 10.   Total lags used: 24
```

Figure 32. Residuals for regression with ARIMA(1,0,0)(1,0,0)[12] errors and lagged predictors.

- The residuals seem to be white noise and have zero mean. Hence, the model is unbiased.
- The Breusch-Godfrey test has a small p-value < 0.05, and we cannot reject the null of no autocorrelation. Hence, the residuals show autocorrelation and the model is not efficient, unlike the model without lagged predictors.
- The residuals are still not normally distributed but the distribution is closer to normal than in the previous case.
- Overall, the residual diagnostics also point to the fact that this model is slightly inferior to the dynamic regression model with just ARIMA errors.

Comparing the Models

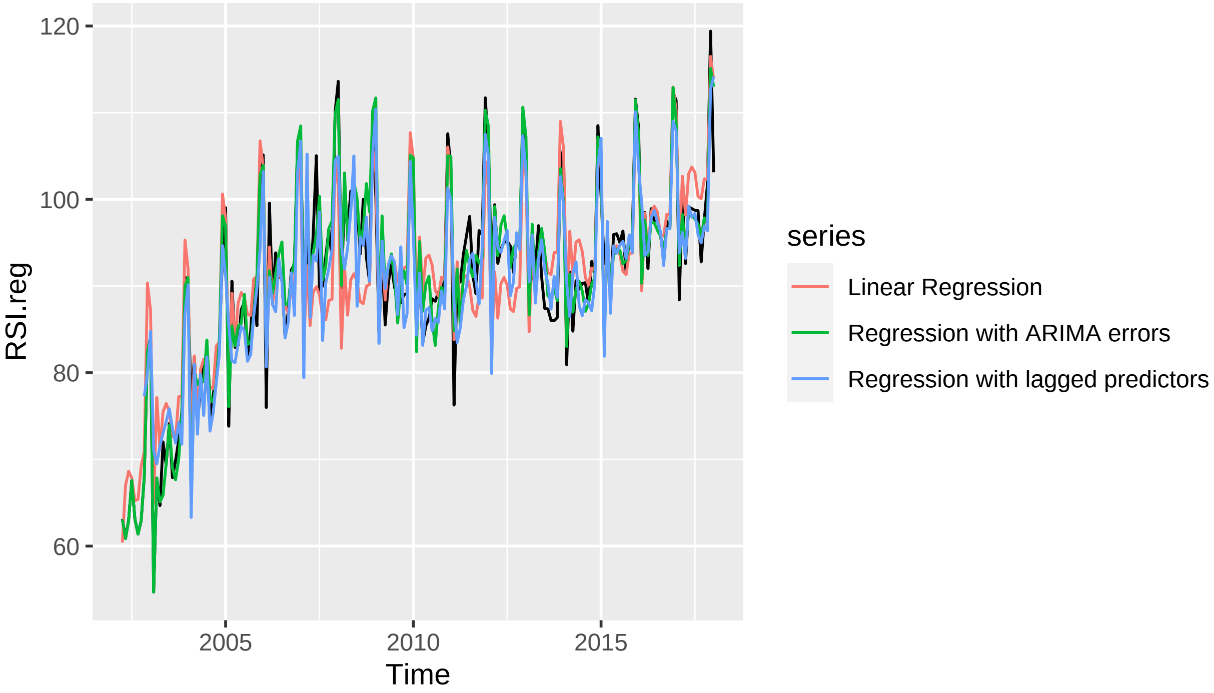


Figure 33. Comparing different regression models.

```
##                                       RMSE      MAE     MAPE      MASE
## Linear Regression                 4.787422 3.808618 4.298456 0.7537885
## Regression with ARIMA errors      3.367990 2.402054 2.641220 0.4754063
## Regression with lagged predictors 4.850609 3.663779 4.093691 0.7359820
```

The model with the lowest errors is the dynamic regression model with ARIMA errors. Hence, it is the best fit and is used for forecasting.

Forecasting

We perform forecasting using the best-fitted model.

Ex-Ante

In order to forecast using ex-ante forecasting, we first must forecast values for the predictor variables. We use the automatically selected ARIMA models to do so, using *auto.arima*, upto January 2021.

The RSI is then regressed on the forecasted values of the predictor variables to get the predictions using the dynamic regression model with ARIMA errors defined above.

The training and test set accuracy is calculated.

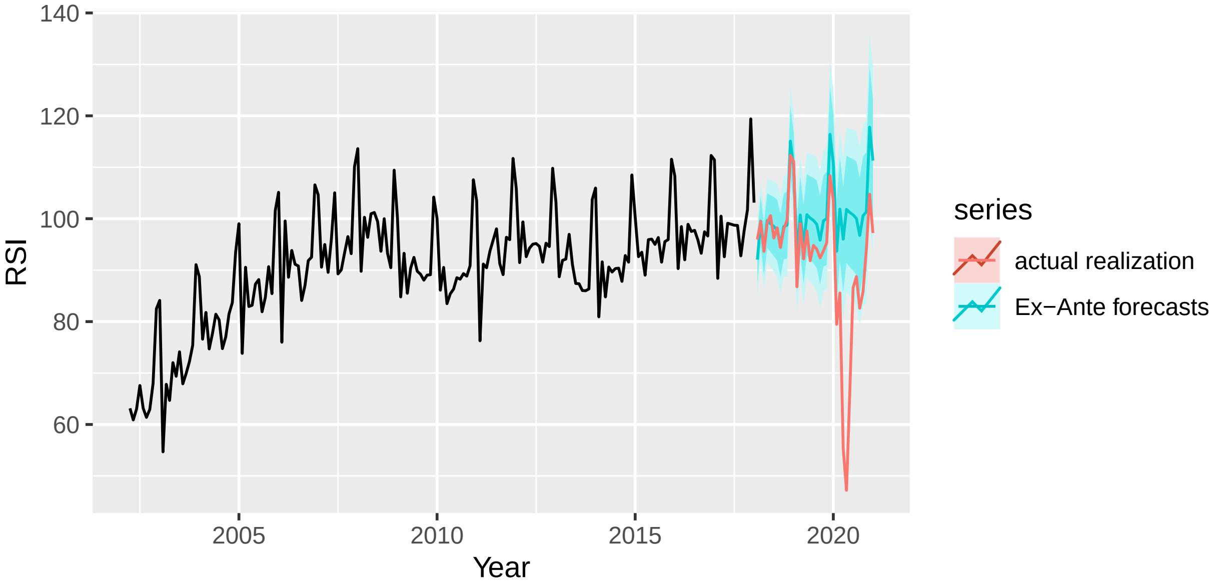


Figure 34. Ex−Ante forecasts using regression with ARIMA errors

```
##                   RMSE      MAE     MAPE      MASE
## Training set  3.36799 2.402054  2.64122 0.4754063
## Test set     14.79175 9.042227 12.64653 1.7896062
```

Ex-Post

For Ex-Post forecasting, we assume knowledge of the predictor variables but not the dependent variable. Hence, we take a 80-20 split of the variables, and use 80% of the data to fit the regression model, and the remaining 20% to test it.

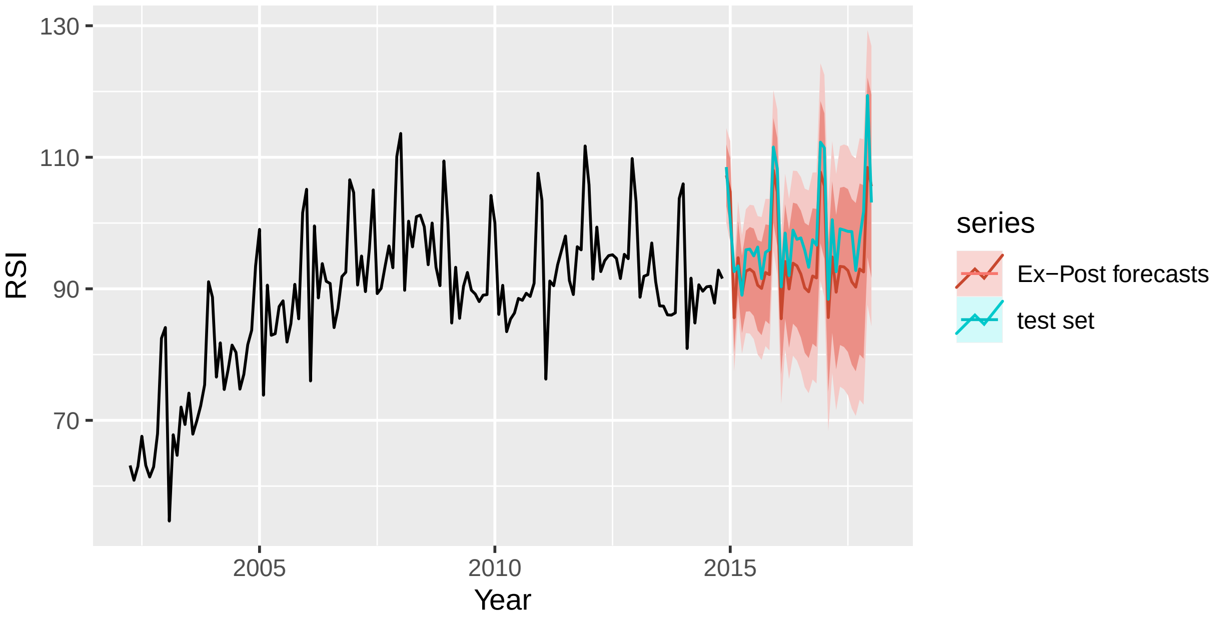


Figure 35. Ex−Post forecasts using regression with ARIMA errors

```
##                  RMSE      MAE     MAPE      MASE
## Training set 3.430977 2.479636 2.764416 0.4510432
## Test set     4.872822 4.378136 4.402383 0.7963785
```

As expected, the ex-post forecasting has a lower error on the test set than the ex-ante forecasting, as the true realizations of the predictors are more accurate than the predictions.

The tsCV RMS errors are calculated based on ex-post forecasting rather than ex-ante forecasting, as we will not have knowledge of the future realizations of the predictor variables while making predictions for the future values of the RSI. Hence, the ex-post forecasting method is comparable to the other models covered above.

```
## Ex-Post Forecast 4.849818
```

Comparing Models

Here are the tsCV RMS errors for all of the models considered above.

```
##                                       tsCV RMSE
## STl-Random Walk wtih Drift             5.243880
## Holt-Winters Additive                  4.554754
## Holt-Winters Multiplicative            4.553761
## ETS                                    4.594100
## STL-ETS                                4.531774
## ARIMA(3,1,0)(2,1,0)[12]                4.719721
## Auto ARIMA                             4.510014
## STL-ARIMA                              4.581719
## Dynamic Regression with ARIMA errors   4.849818
```

The top three models are:

- Auto ARIMA
- STL-ETS
- Holt-Winters Multiplicative

We combine these three models and test the accuracy.

```
## Combination Forecast NaN
```

As we get an NaN value for the tsCV RMS error, we compare the train-test errors for the three models to that of the combination model.

```
##                                 RMSE       MAE     MAPE      MASE
## Combination Model           12.24997  8.744623 10.15706 0.3630873
## Auto ARIMA                  13.64725  9.118890 10.97716 1.7318509
## STL-ETS                     15.29506 12.952221 13.93535 2.4598734
## Holt-Winters Multiplicative 12.86251  8.671510 10.36476 1.6468850
```

Here we see that the combination forecasted better than any of the 3 models. Hence, the combination model will be used for the final forecasting.

Forecasts

Forecasting

To decide the weights for each component of the combination model, the forecasts are first visualized, for upto 5 years in the future.

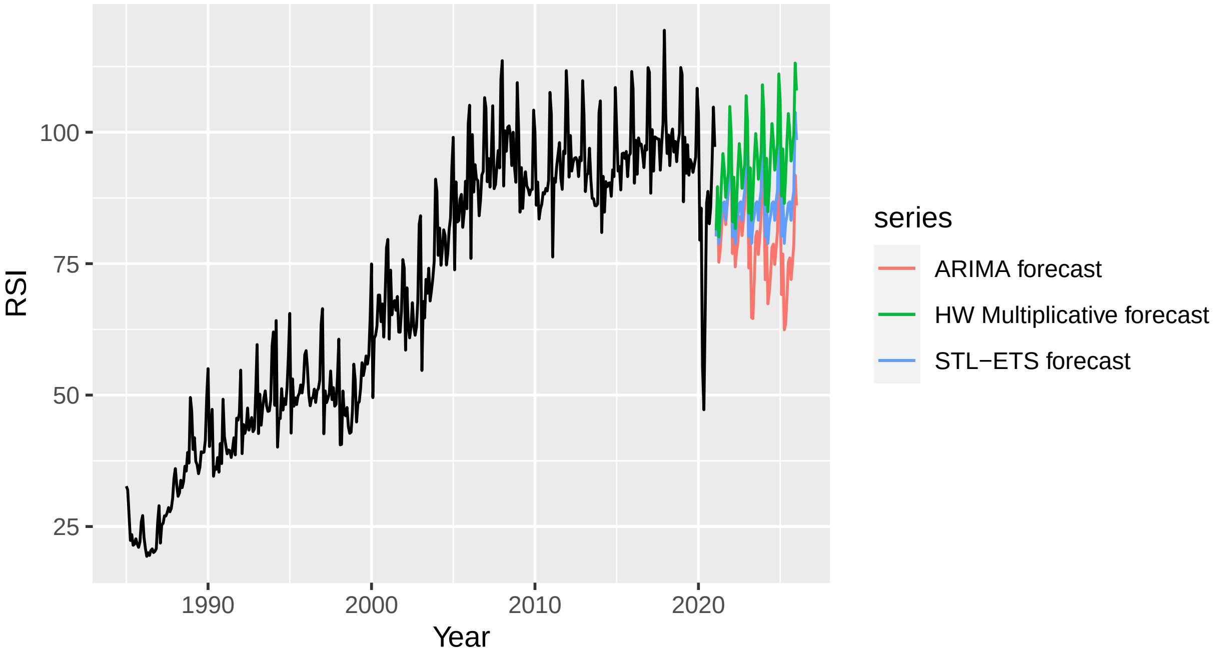


Figure 36. Forecasts using top three models

The ARIMA and STL-ETS forecasts put more emphasis on the recent dip in RSI due to the COVID-19 pandemic, and forecast a decreasing or constant trend into the future. However, we know that the dip was an outlier due to a black swan event, and the economy is expected to recover in the coming years. (Ting, 2021) Hence, we place a lower weight on the ARIMA forecasts, at 20%, a medium weight of 30% on the STL-ETS forecasts, and a higher weight of 50% on the HW-Multiplicative forecasts.

The forecasts for the next five years from the combination method with the above-mentioned weights are shown below.

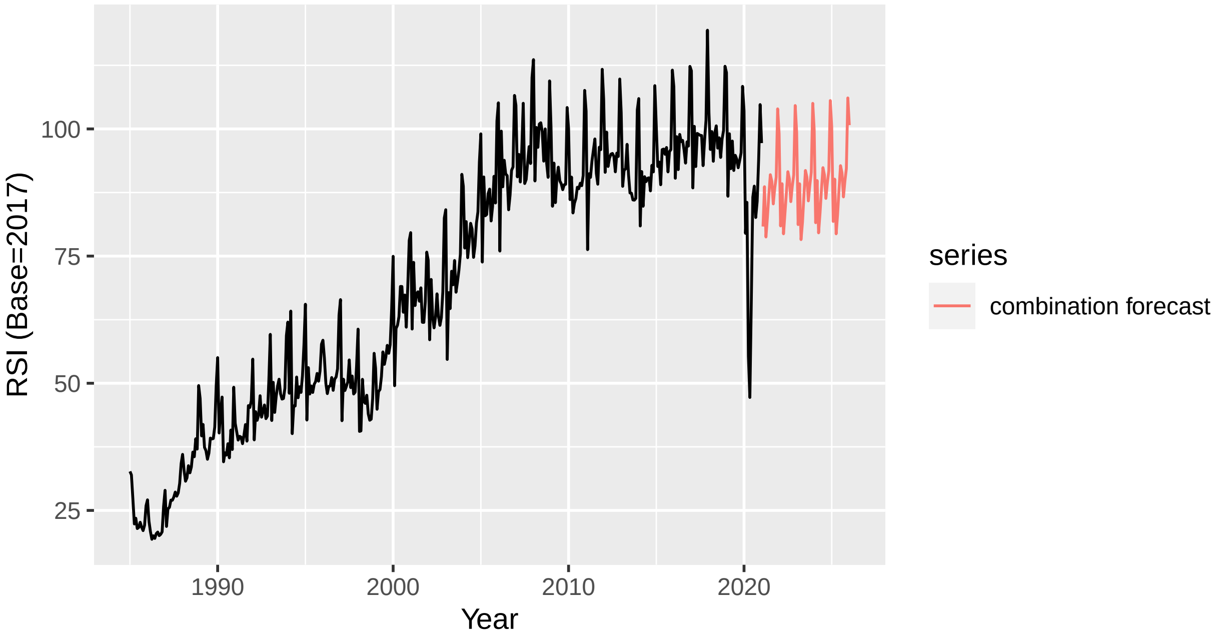


Figure 37. Forecasts using combination method

However, the predictions lack a confidence interval. They are generated using the following method:

- 250 time series similar to RSI are generated using the bootstrap method.
- Each bootstrapped time series is used to make forecasts using each of the top three models, and these forecasts are used to make the combination forecast with the desired weights.
- The predictions are combined and different percentiles are taken to construct the confidence intervals, and the previously calculated values are taken for the point forecast (rather than taking the mean of the simulated series).

To make the process quicker, the model picked by *auto.arima* was found and hardcoded into the simulation to fit all the bootstrapped series.

```
## $method
## [1] "ARIMA(0,1,1)(2,1,2)[12]"
```

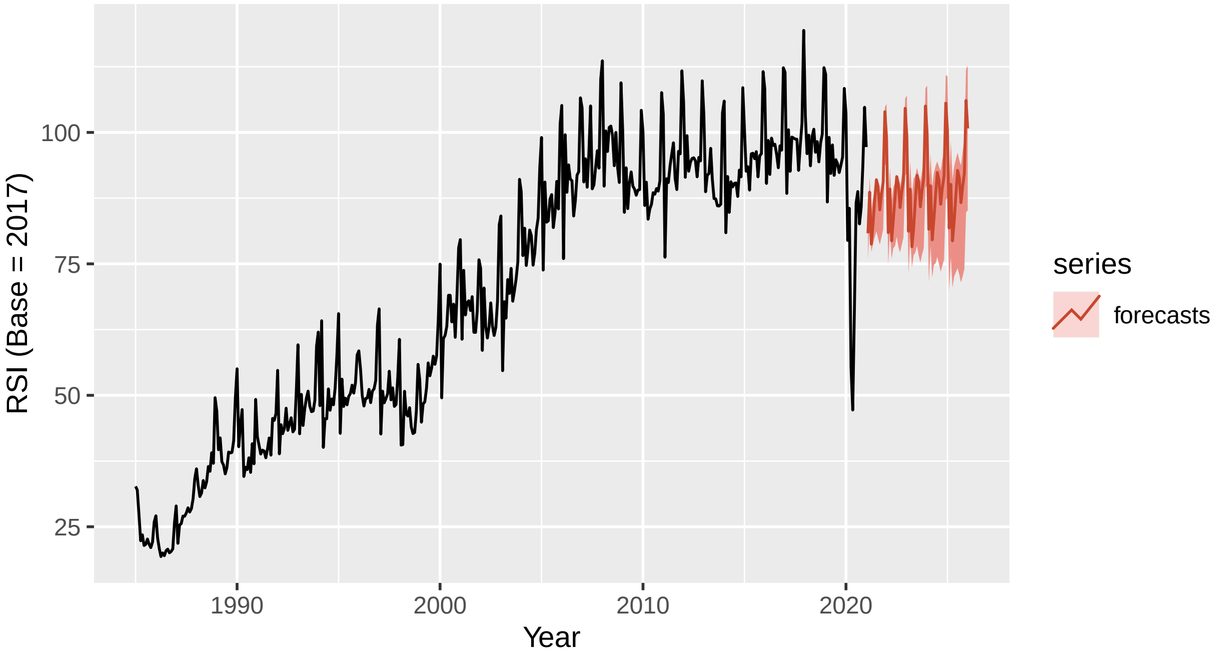


Figure 38. Forecasts using combination method (with prediction intervals).

These are the final forecasts.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2021	NA	80.81538	88.63918	78.75561	82.65957	87.18606	91.00484	89.72295	85.25794	88.31844	90.52412	103.9474
2022	99.06734	80.92870	89.22772	79.37424	83.41444	87.57382	91.59749	90.18693	85.70434	88.69807	90.97240	104.5957
2023	99.31123	81.19130	89.21774	78.23691	81.70836	87.05598	91.83160	90.53435	85.85244	88.93688	91.46632	105.0272
2024	99.81614	81.56255	89.85241	79.56303	83.54257	88.16939	92.38713	90.96664	86.33816	89.41805	91.76570	105.5769
2025	100.30997	81.80995	90.12710	79.37549	83.18521	88.21526	92.77900	91.36194	86.63187	89.74294	92.20929	106.0887
2026	100.74699	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Analysing Forecasts

We look at various plots for the total time series, which is constructed by appending the dataset and the forecasts.

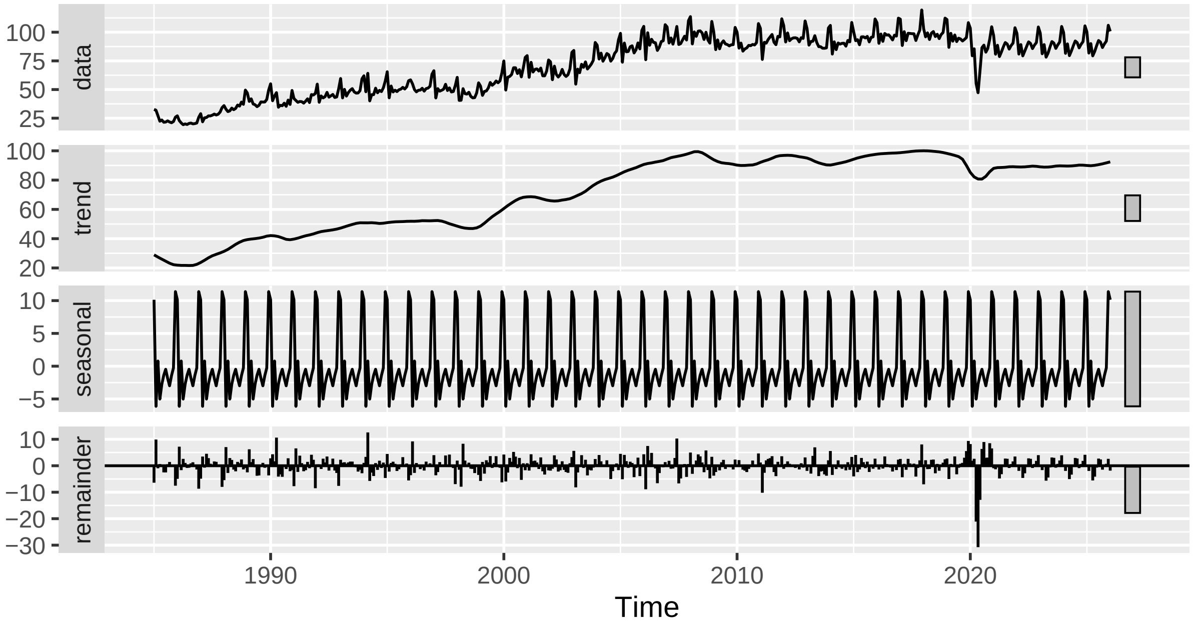


Figure 39. Decomposition plot for predicted values of RSI.

The decomposition plot shows a gradual recovery from the COVID-19 crisis. The trend is seen to be gradually increasing for the forecasts. The seasonal component for the forecasts is more or less the same as that of the dataset.

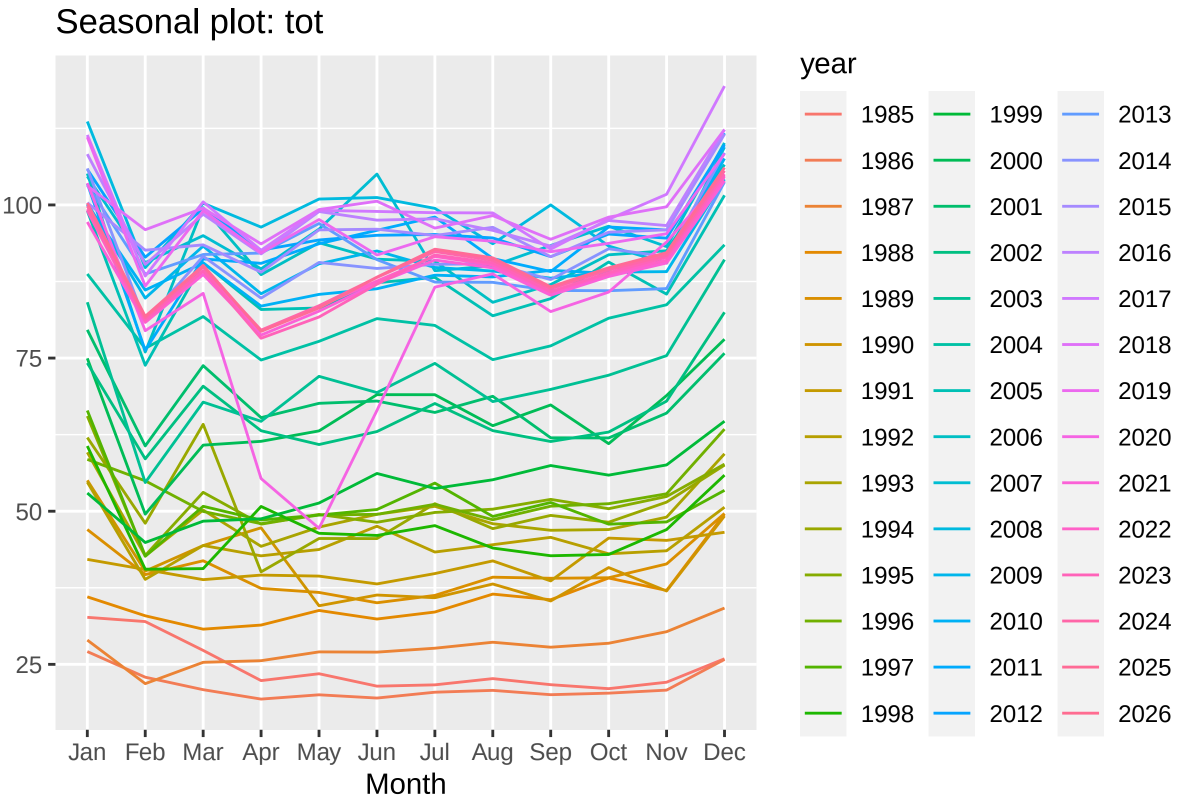


Figure 40. Seasonal plot for predicted values of RSI.

The forecasts continue the seasonal pattern observed in the dataset: there is a spike in the RSI during December and January, and a slump in the middle of the year. The drop at April is slightly exaggerated for the forecasts as the model is skewed due to outlier values caused by the COVID-19 pandemic.

Conclusion

The best forecasting models were found to be ARIMA(0,1,1)(2,1,2)[12], STL-ETS and Holt-Winters Multiplicative method. The three were combined with 20%, 30% and 50% weights, respectively, to predict the values for the RSI five years into the future, till Janurary 2026.

The forecasts show a slow but steady recovery in RSI, which suffered a large drop in 2020 due to the COVID-19 pandemic. The model predicts that it will be back to the pre-COVID values by the end of 2021. This is in line with the 2021 outlook given in the Economic Survey of Singapore, and other forecasts by the Monetary Authority of Singapore (Ting, 2021). They also match the trend and seasonality observed in the dataset.

Some limitations to our predictive model are:

- The forecasts may be skewed due to the outlier values in 2021 caused by the COVID-19 pandemic.
- The prediction intervals contructed by the bootstrap method are not very accurate.
- If appropriate forecasting methods are available for predictor models, the dynamic regression model with ARIMA errors may be considerd to be added to get even more accurate forecasts for RSI.

References

Certificate of Entitlement (COE). (n.d.). Retrieved from https://onemotoring.lta.gov.sg/content/onemotoring/home/buying/upfront-vehicle-costs/certificate-of-entitlement--coe-.html

COE Trends. (n.d.). Retrieved from https://coe.sgcharts.com

ChannelNewsAsia. (2021, March 5). *Singapore retail sales fall 6.1% in January, decline in most industries.* Retrieved from https://www.channelnewsasia.com/news/singapore/singapore-retail-sales-index-jan-2021-fall-6-1-percent-14340458

CPF Contribution Rates. (2017, May 15). Retrieved from https://data.gov.sg/dataset/contribution-rates-allocation-rates-and-applicable-wage-ceiling?resource_id=65db3d22-9b16-43a3-8d4b-a2133043a78b

Fung, M. (2021, May 3). *India leads surge in Covid-19 infections across the globe.* Retrieved from https://www.straitstimes.com/asia/india-leads-surge-in-covid-19-infections-across-the-globe

Investopedia. (n.d.). Retail Sales Definition. Retrieved from https://www.investopedia.com/terms/r/retail-sales.asp

Kit, T. S. (2021, May 4). *Possibility of circuit breaker 'not ruled out' as COVID-19 taskforce announces tighter measures.* Retrieved from https://www.channelnewsasia.com/news/singapore/possibility-of-circuit-breaker-not-ruled-out-covid-19-singapore-14742990

Ministry of Trade & Industry. (2020). Economic Survey of Singapore 2019. Retrieved from https://www.mti.gov.sg/-/media/MTI/Resources/Economic-Survey-of-Singapore/2019/Economic-Survey-of-Singapore-2019/FullReport_AES2019.pdf

Ministry of Trade & Industry. (2021). Economic Survey of Singapore 2020. Retrieved from https://www.mti.gov.sg/-/media/MTI/Resources/Economic-Survey-of-Singapore/2020/Economic-Survey-of-Singapore-2020/FullReport_AES2020.pdf

Per Capita GNI. (2019, September 5). Retrieved from https://data.gov.sg/dataset/per-capita-gni-and-per-capita-gdp-at-current-market-prices-annual

Retail Sales Index. (2020, February 21). Retrieved from https://data.gov.sg/dataset/retail-sales-index-2017-100-at-constant-prices-ssic-2015-monthly-sa?resource_id=b5c95339-7a1c-4f71-874f-70dec941e91b

Singapore Department of Statistics. (2021). Retail Sales Index. Retrieved from https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=16924

Singapore Residents. (2020, July 17). Retrieved from https://data.gov.sg/dataset/resident-population-by-ethnicity-gender-and-age-group

SingStat. (2021). Retail Sales Index and Food & Beverage Index. Retrieved from https://www.singstat.gov.sg/-/media/files/news/mrsjan2021.pdf

Sng, H. Y. (2020). Lecture on Crisis Management and Wage Policy in Singapore. Personal Collection of H. Y. Sng, Nanyang Technological University, Singapore.

Teo, J. (2021, May 4). *Five patients in TTSH Covid-19 cluster found to have virus variant from India.* Retrieved from https://www.straitstimes.com/singapore/health/ttsh-covid-19-cluster-five-patients-have-india-variant-of-virus

Ting, C. Y. (2021, April 28). *S'pore economy to grow faster than 6% in 2021, but recovery will be more uneven across sectors: MAS.* Retrieved from https://www.straitstimes.com/business/economy/singapore-economy-to-grow-faster-than-6-in-2021-but-recovery-will-be-more-uneven

Appendix A

Benchmark: Unsatisfactory Methods

Average Method:

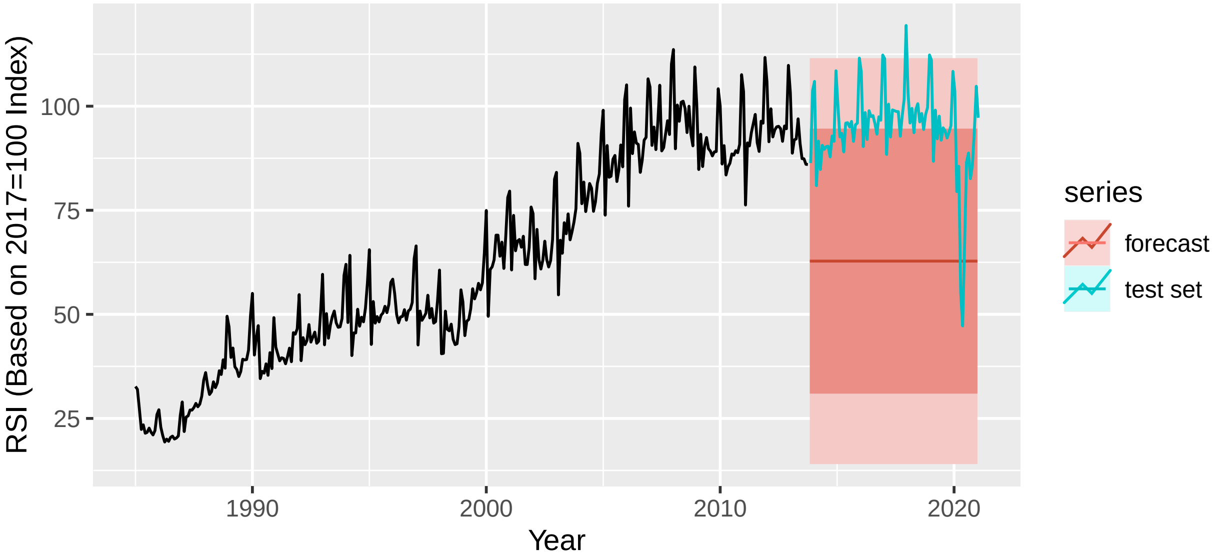


Figure A1. Forecasts from average method

- The values are forecasted by setting them equal to the average value of the test set.
- This is insufficient as we observe an upward trend.

Naive Method

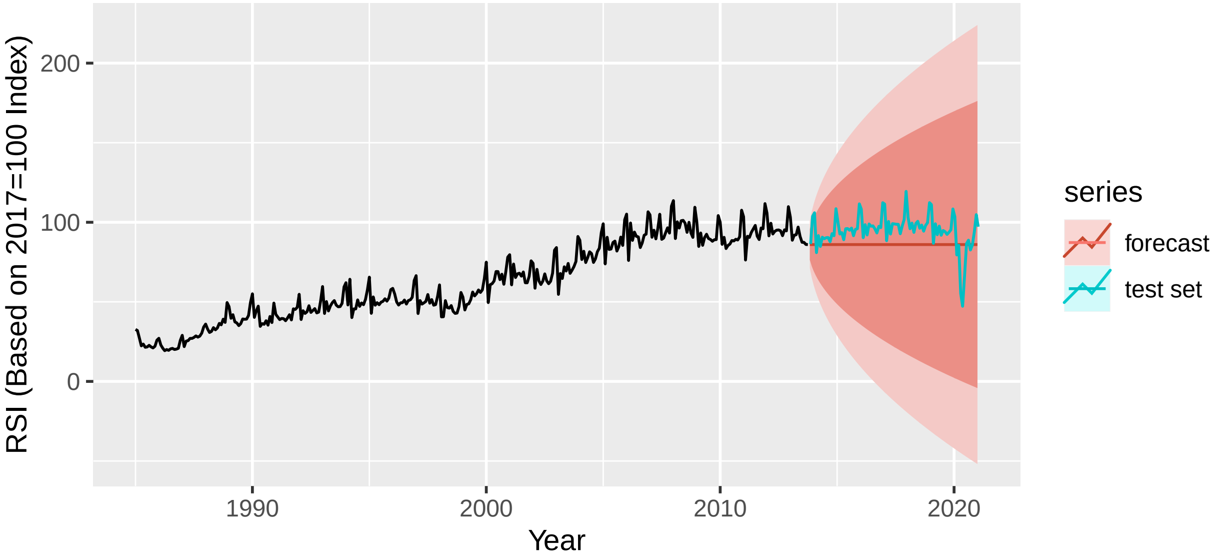


Figure A2. Forecasts from naive method

- In this method, the forecast is set to the last observed value from the train set.
- This method is insufficient because the value doesn't remain constant with time.

Seasonal Naive Method

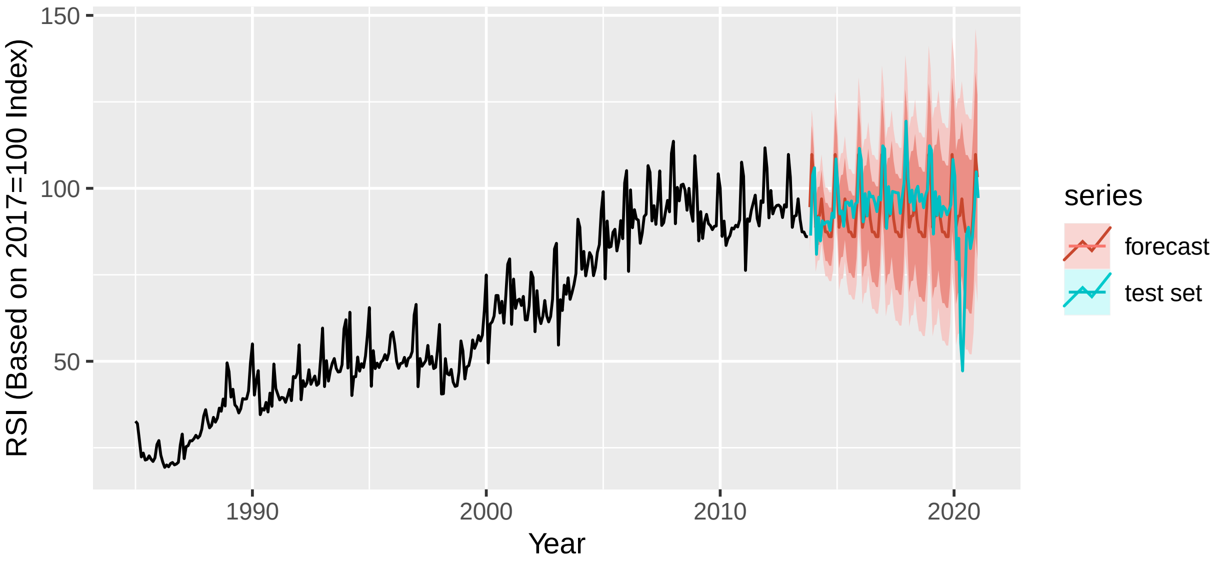


Figure A3. Forecasts from seasonal naive method

- This method predicts the last observed value of the same season of the last year.
- It doesn't seem to work well because it just considers data of the last year for forecasting and hence this method is insufficient.

Random Walk with Drift Method

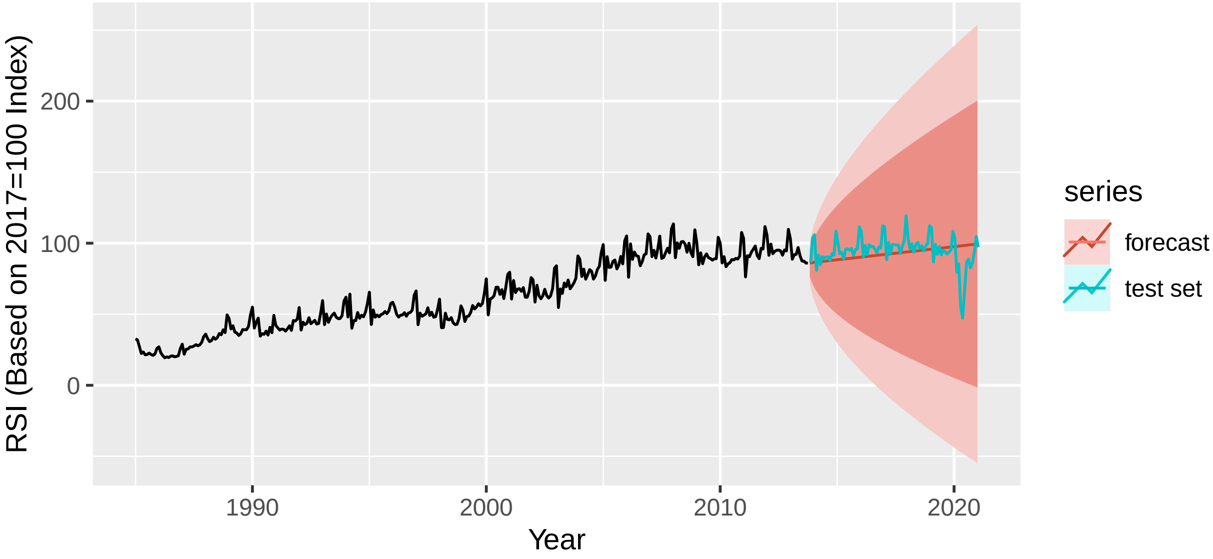


Figure A4. Forecasts from random walk with drift

- In this method, the forecast is set to the previous value plus the average trend over time.
- It is insufficient as it assumes the trend to be constant over time.

STL-Naive method

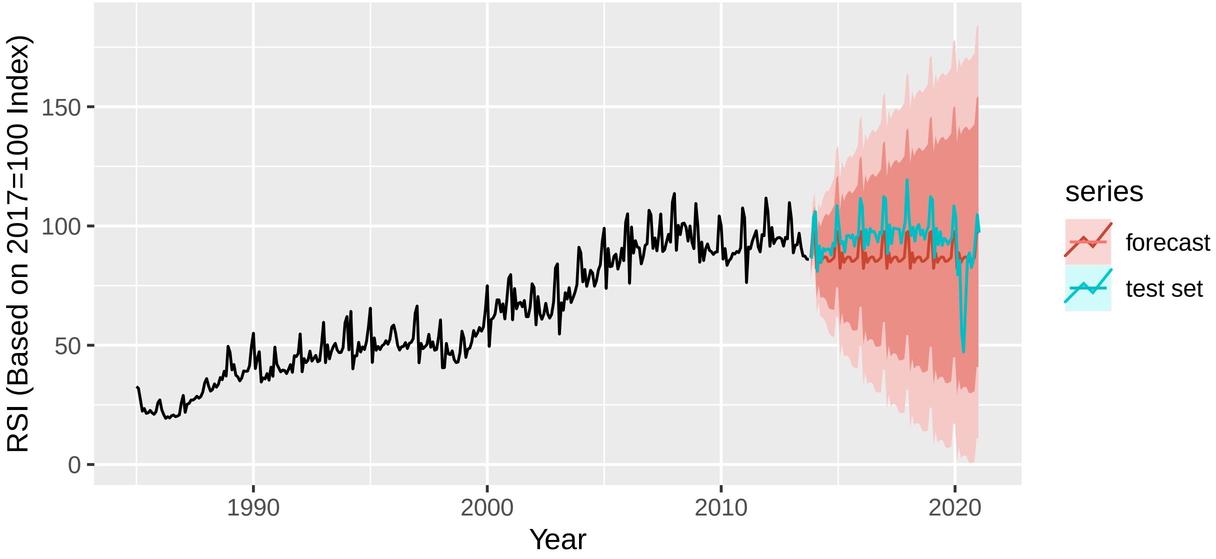


Figure A5. Forecasts from STL−Naive method

- Seasonally adjusted data is forecasted using the Naive method, and then is re-seasonalized by adding the seasonal component.
- It is insufficient as it doesn't consider a trend.

Exponential Smoothing: Unsatisfactory Models

Simple Exponential Smoothing

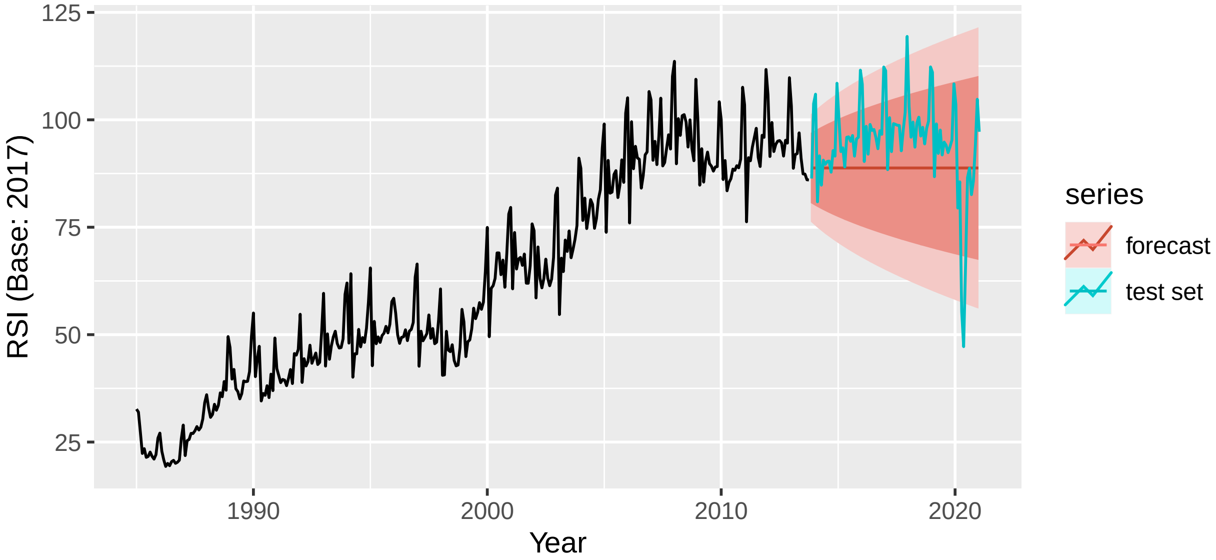


Figure A6. Forecasts from simple exponential smoothing method

- This method is suitable for forecasting data with no clear trend or seasonal pattern.
- Here it is insufficient as there is a clear upward trend and seasonality.

Holt’s Method

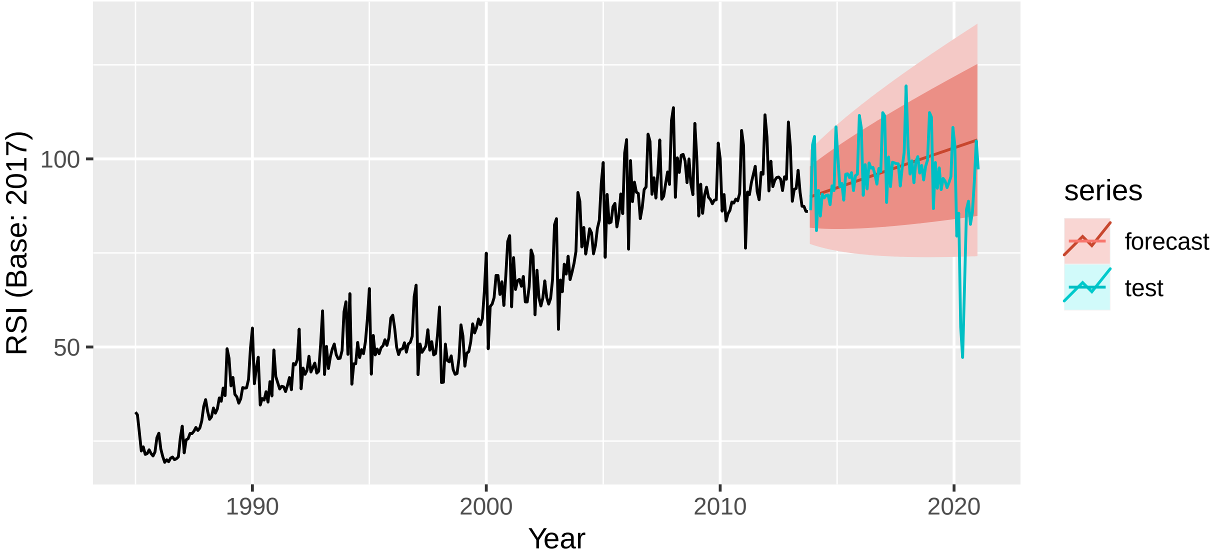


Figure A7. Forecasts from Holt's method

- This method is an extended simple exponential smoothing which helps with forecasting of data with a trend.
- Usually, the forecasts generated by Holt’s linear method display a constant trend (increasing in this case) indefinitely into the future. Because of this, this method tends to over-forecast. Hence, it is insufficient.

Damped Holt’s Method

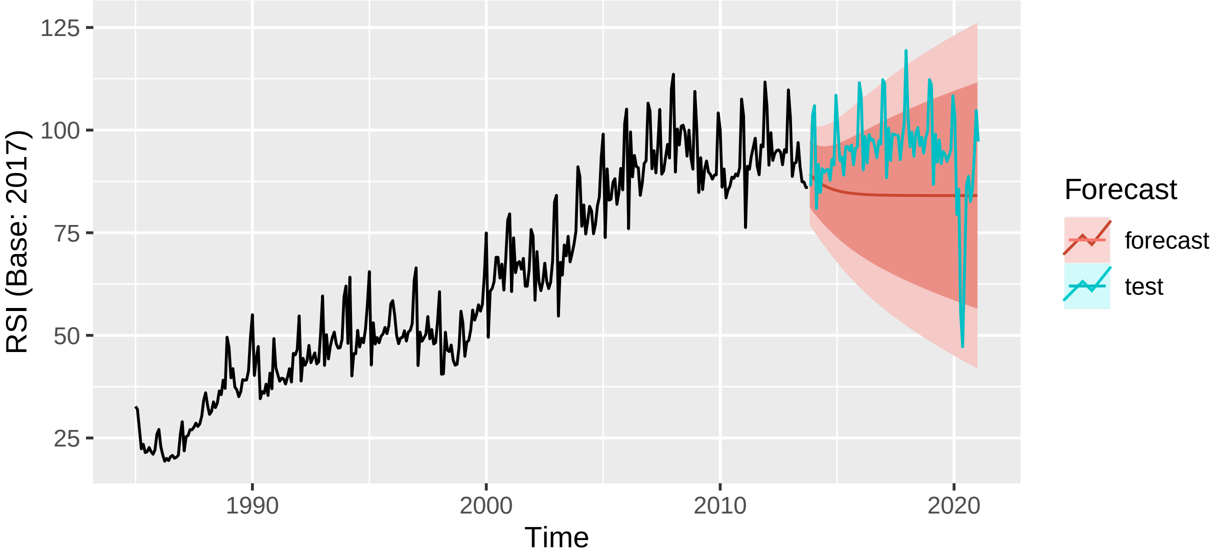


Figure A8. Forecasts from damped Holt's method

- In conjunction with the smoothing parameters in Holt’s method, this method also includes a damping parameter.
- As it can be seen, this method underestimates the RSI values when compared with the test set and hence, this method is insufficient.

Holt-Winters Multiplicative Damped Method

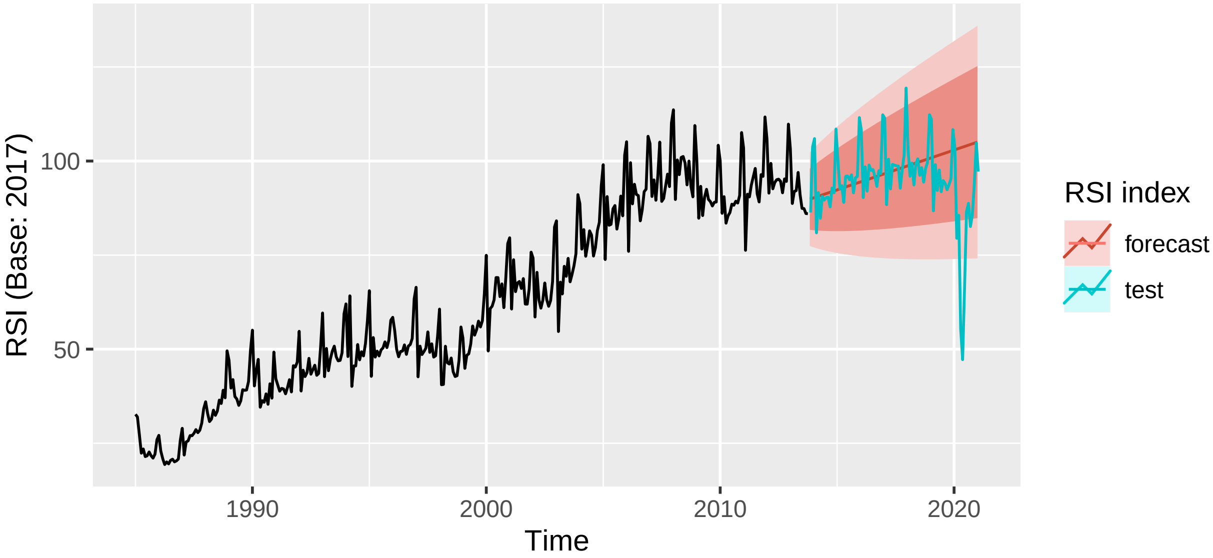


Figure A9. Forecasts from HW Multiplicative damped method

- With the multiplicative method, the seasonal component is expressed in relative terms (percentages), and the series is seasonally adjusted by dividing through by the seasonal component.
- As it can be seen, the model has identified the seasonal pattern but since the forecasts are not a close match to the test data, this model is insufficient.