Totalnsa: Total number of new vehicles sold in the U.S. in a given month

Massimiliano Pastorino and Caterina Penna

The purpose of this study is to analyze the time series of the total number of new vehicle sales in the United States. The time series was retrieved from the website of the Federal Reserve Bank of St. Louis.

Here is the url:
https://fred.stlouisfed.org/series/TOTALNSA

The data taken into consideration range from January 1976 to December 2021. In the first part of the project, we will analyze the main characteristics of the time series, and we will consider some common methods for extract the three components of a time series (a trend-cycle component, a seasonal component and a residual component), applying the method of classical decomposition and the method of STL decomposition.

In the second part of the project, we want to identify the process that "best" represents the series. More specifically, we will ask which type of process has realizations that most closely resemble the observed series.

1 Characteristics of Time Series

As we can see from Figure 1, data shows some clear seasonality patterns. Upward trend start to 1992 to 2007, with a sharp drop in 2008-2009 (Great Financial Crisis). Sharp recovery until 2020, where vehicle sales declined due to the COVID pandemic.

The data shows clear cycles, the fluctuations are not of a fixed frequency then they are cyclic. In economic time series, these fluctuations are usually due to economic conditions.

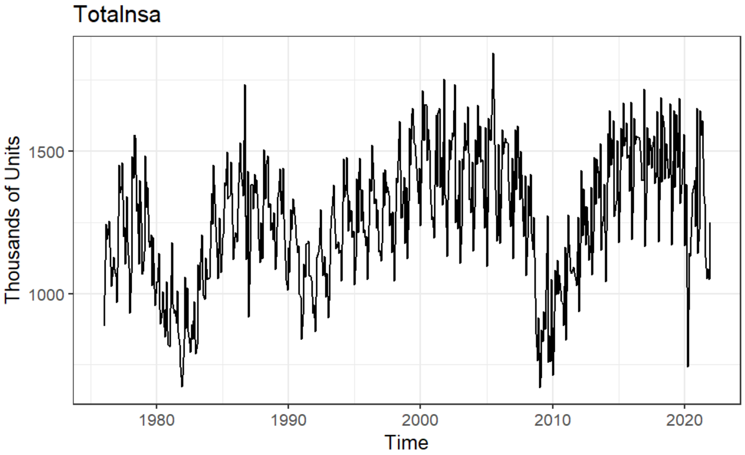


Figure 1: **Time plots**

The real data often contain missing values and observations that are very different from the majority of the observations in the time series (outliers). As we can see from Figure 2, totalnsa time series has no missing values, while from Table 1 we can see the presence of anomalous values classified as potential outliers.

The unit of measurement of the totalnsa time series is Thousands of Units. As mentioned earlier, data are characterized by periods of expansion alternating with periods of recession.

During periods of recession, the number of vehicles sold in the US in a given month is very low, while during periods of expansion, the number of vehicles sold in the US in a given month is very high.

Compared to most values assumed by the time series, the values identified as potential outliers are simply unusual values and not values derived for example from a transcription error.

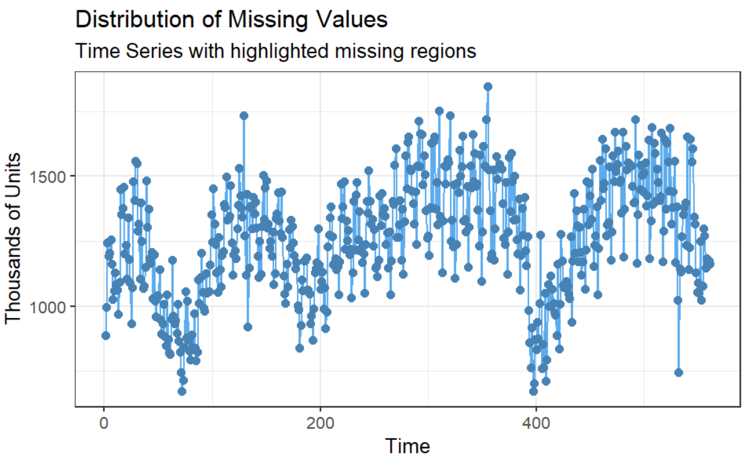


Figure 2: **Distribution of Missing Values**

Table 1: **Possible outliers**

DATE	**TOTALNSA**
1980-05-01	893.800
1986-09-01	1733.600
2001-10-01	1753.200
2009-03-01	872.848
2020-03-01	1023.317
2020-04-01	743.276
2020-05-01	1142.705

As we mentioned earlier, the data exhibit seasonality. A plot that emphasises the seasonal patterns is where the data for each season are collected together in separate mini time plots. Figure 3 shows the Seasonal subseries plots.

The horizontal lines indicate the means for each month. This form of plot enables the underlying seasonal pattern to be seen clearly, and also shows the changes in seasonality over time. It is especially useful in identifying changes within particular seasons.

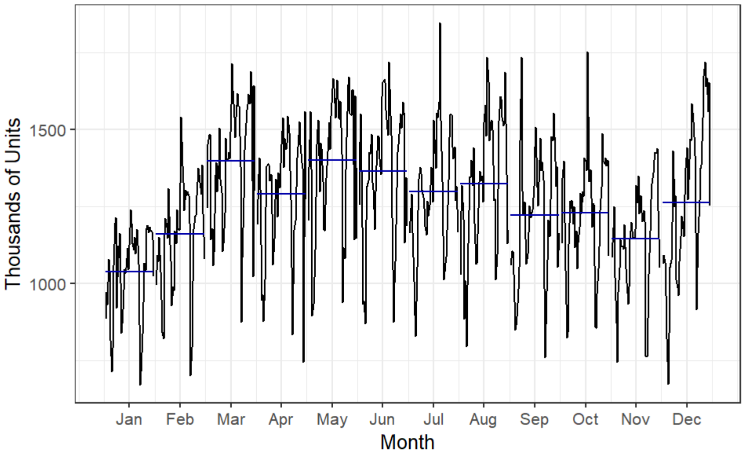


Figure 3: **Seasonal subseries plots**

1.1 Trend and seasonality in ACF plots

Just as correlation measures the extent of a linear relationship between two variables, autocorrelation measures the linear relationship between lagged values of a time series. Figure 4 shows the ACF (Autocorrelation function).

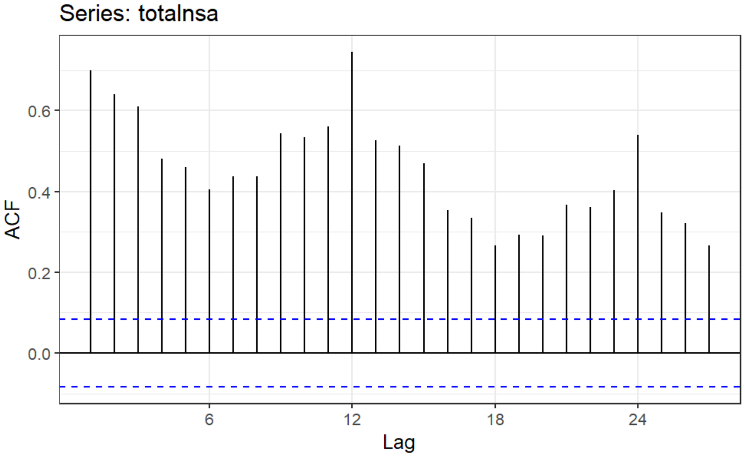


Figure 4: **ACF plot**

When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags. When data are both trended and seasonal, you see a combination of these effects.
In Figure 4:

- $\hat{p}(12)$ is higher than for the other lags. This is due to the seasonal pattern in the data.
- $\hat{p}(6)$ is smaller than for the other lags.
- $\hat{p}(0)$ is always equal to 1.

The dashed blue lines indicate whether the autocorrelations are significantly different from zero. Time series that show no autocorrelation are called white noise. For white noise series, we expect each autocorrelation to be close to zero. Of course, they will not be exactly equal to zero as there is some random variation.

If time series is white noise, it can be shown that the sample autocorrelations $\hat{p}(h)$, with h>0, are approximately $N(0, \frac{1}{T})$ for sufficiently large T. For a white noise series, we expect 95% of the spikes in the ACF to lie within ± $\frac{1.96}{\sqrt{T}}$ where T is the length of the time series.

If one or more large spikes are outside these bounds, or if substantially more than 5% of spikes are outside these bounds, then the series is probably not white noise. In this case, totalnsa time series is not white noise.

The correlogram is also useful for identifying non-stationary time series:

- The ACF for a stationary time series decreases to zero very rapidly.
- The ACF for a non-stationary time series decreases slowly.
- For non-stationary data, the value of $\hat{p}(1)$ is often large and positive.

In this case, totalnsa time series is not stationary.

In addition to the autocorrelation function, we also analyze the partial autocorrelation function. Partial autocorrelations measure the linear dependence of one variable after removing the effect of other variable(s) that affect to both variables. Figure 5 shows the PACF (Partial autocorrelations functions).

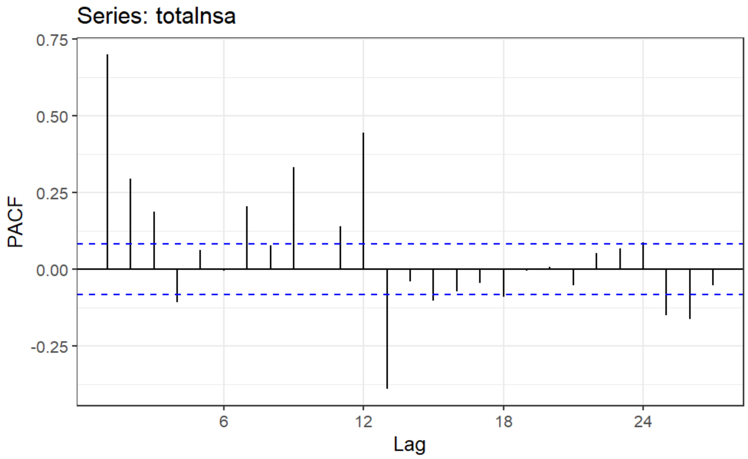


Figure 5: **PACF plot**

PACF suggests that the time series is not stationary.

2 Time series decomposition

Let us consider some common methods for extracting the components of a time series. Let's talk about the classical decomposition method.

The first step in a classical decomposition is to use a moving average method to estimate the trend-cycle. The estimate of the trend-cycle at time t is obtained by averaging values of the time series within k periods of t. Observations that are nearby in time are also likely to be close in value.

We call the an m-MA, meaning a moving average of order m. The order of the moving average determines the smoothness of the trend-cycle estimate. In general, a larger order means a smoother curve. We can see this from Figure 6.

5-MA is the average of the observations in the five year window centred on the corresponding year. In this case, there are no values for either the first two years or the last two years, because we do not have two observations on either side.

The second graph, shows a 2×12-MA (Moving averages of moving averages), in this case we have a we might take a moving average of order 12, and then apply another moving average of order 2 to the results. One reason for doing this is to make an even-order moving average symmetric. In a moving average of order m=2k+1, the middle observation, and k observations on either side, are averaged. But if m was even, it would no longer be symmetric.

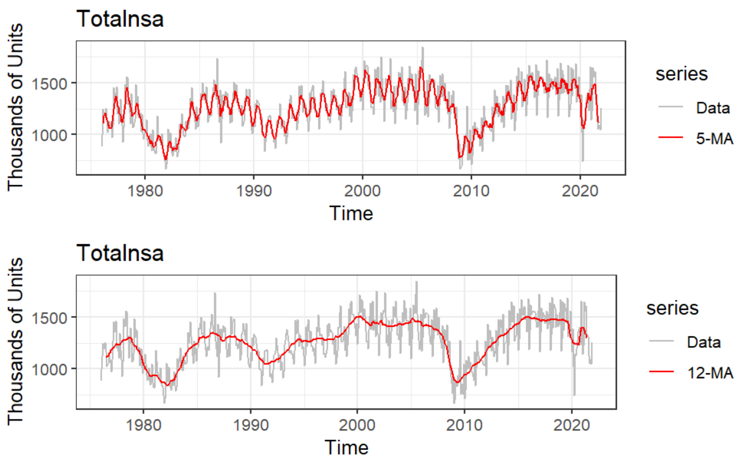


Figure 6: **Different moving averages applied to Totalnsa**

2.1 A classical additive decomposition

In classical decomposition, we assume that the seasonal component is constant from year to year. There are two forms of classical decomposition: an additive decomposition and a multiplicative decomposition.

The additive model is appropriate when the magnitude of seasonal fluctuations does not vary with the level of the series. In this case we will use additive decomposition.

In an additive model the first component we need to estimate is the trend/cyclic component.

If m is an even number, we calculate $\hat{T}_t$ using a 2m$-$Ma, if m is an odd number we calculate $\hat{T}_t$ using an m$-$MA.

In the second step calculate the detrended series: $y_t - \hat{T}_t$.

In the third step estimate the seasonal component for each season, $\hat{S}_t$ simply average the detrended values for that season.

In the final step the remainder component is calculated by subtracting the estimated seasonal and trend-cycle components: $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$.

From Figure 7, we see that the seasonal component exhibits periodic and regular fluctuations with constant amplitude. Seasonality is constant from year to year. This reflects the assumption on seasonality that underlies the classical decomposition, which assumes that seasonality is constant from year to year.

The residual component has zero mean. Oscillations fluctuate around zero.

We have variations that are more or less constant.

From the remainder component, we can see that between 2019 and 2020, we have values significantly less than 0, which suggests that the cyclical trend component has not captured the dynamics of the series correctly. In the residual component we have left some information behind as the cyclical trend component estimated in Step 1 has "smoothed" the data too much.

In fact, from the data we see an obvious downward and the trend component in this case is too smooth. We have a poor estimate of the trend in the period between 2019 and 2020.

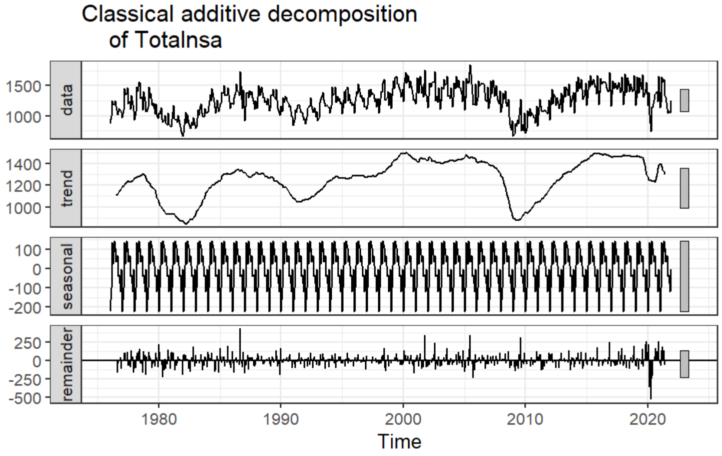


Figure 7: **A classical additive decomposition of Totalnsa**

Classical additive decomposition has the limitation of estimating a cyclic trend component which in most cases is too smooth when we have sudden upward or downward variations in the data. Moreover, classical decomposition methods are not robust to unusual values or shocks in the time series.

2.2 STL decomposition(Seasonal and Trend decomposition using Loess)

In the STL decomposition the seasonal component is allowed to change over time, and the rate of change can be controlled by the user.

The smoothness of the trend-cycle can also be controlled by the user.

It can be robust to outliers, so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component.

On the other hand, STL has some disadvantages. In particular, it does not handle trading day or calendar variation automatically, and it only provides facilities for additive decompositions.

The parameters t.window and s.window control how rapidly the trend-cycle and seasonal components can change. Smaller values allow for more rapid changes.

t.window is the number of consecutive observations to be used when estimating the trend-cycle while s.window is the number of consecutive years to be used in estimating each value in the seasonal component. s.window must specify as there is no default.

With s.window = “periodic” we assume that the seasonal component is no longer variable but constant over time.

Figure 8 show different values of t.windows with s.windows equal to periodic. With t.window equal to 5 we have a less smooth trend, while with a value of t.window equal 15 we have a very smooth trend.

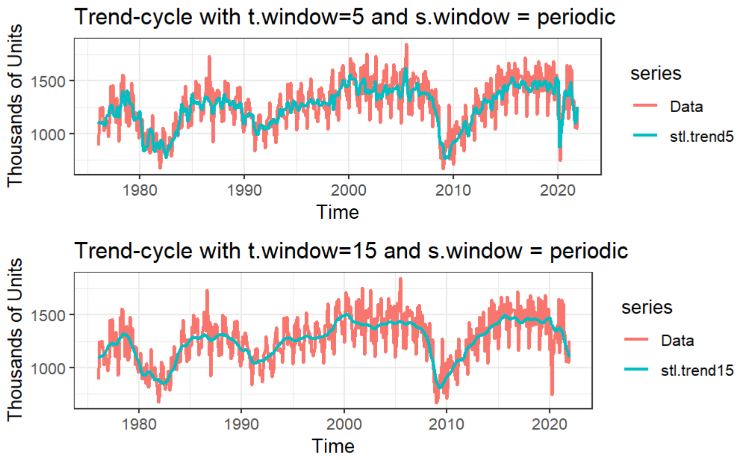


Figure 8: **Trend-cycle with t.windows(5) and t.windows(15)**

In order to have a good balance between overfitting the seasonality and the possibility of this changing slowly over time, we use the function mstl() in which the parameter t.window is chosen automatically while s.window is equal to 13. Overfitted model to data can lead to very wrong predictions. Figure 9 shows the automated STL decomposition, while figure 10 shows the trend-cycle component.

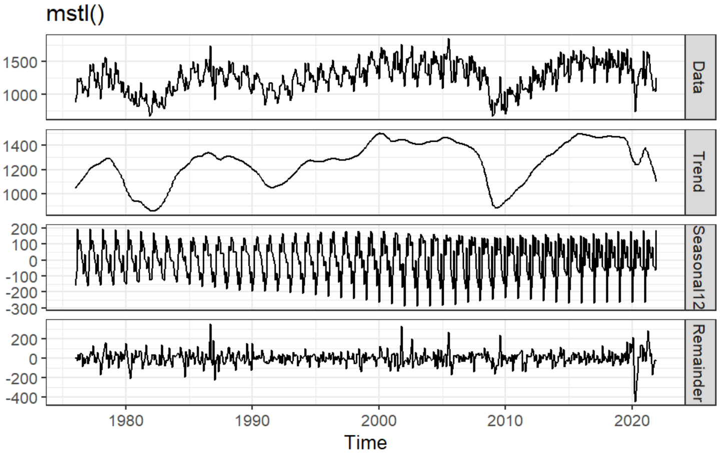


Figure 9: **STL decomposition with mstl() function**

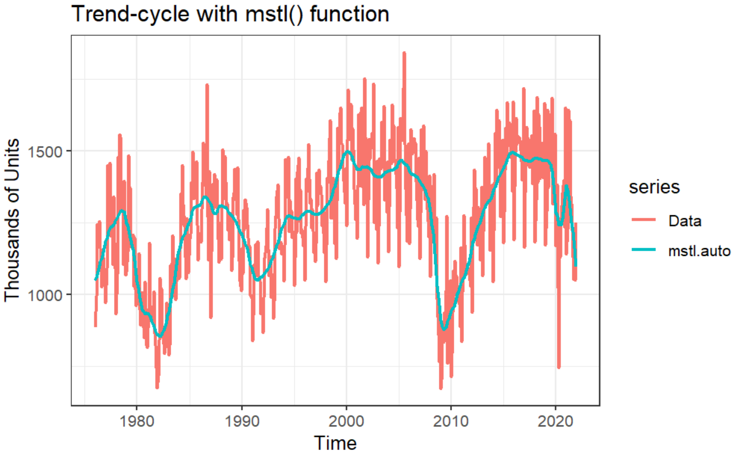


Figure 10: **Trend-cycle with mstl() function**

3 ARIMA Models

3.1 Preliminary analysis

In order to construct an ARIMA model, we must first determine whether our time series can be considered a realization of a stationary process. If it is not, we must transform the time series in order to get stationarity.

From the plot of the time series values we can obtain useful indications concerning the stationarity of the process. If the observed values of the time series seem to fluctuate with constant variation around a constant mean, then it is reasonable to suppose that the process is stationary, otherwise, it is nonstationary.

The time series totalnsa is not stationary, we have seen this from the ACF and PACF plot (Figure 4 and Figure 5). We have two types of non-stationarity:

- eteroschedasticity (non constant variance)
- non stationarity in mean (trend)

To achieve omoschedasticity a Box-Cox transformation can be used. The problem is to find λ^* such that the series is as omoschedastic as possible. In this case, the lambda value that makes the series as homoschedastic as possible is 0.5.

A power transformation was applied to the original data:

$$y_t = \begin{cases} \frac{x_t^{0.5}-1}{0.5} \end{cases} \tag{1}$$

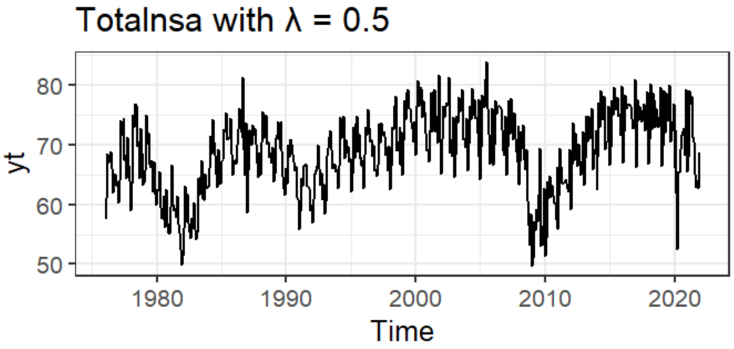


Figure 11: **Totalnsa transformation with $\lambda = 0.5$**

Now let's deal with non - stationarity in mean.

As we can see from Figure 12 (Acf and Pacf), the trend and seasonality make the totalnsa timeseries non - stationary in mean.

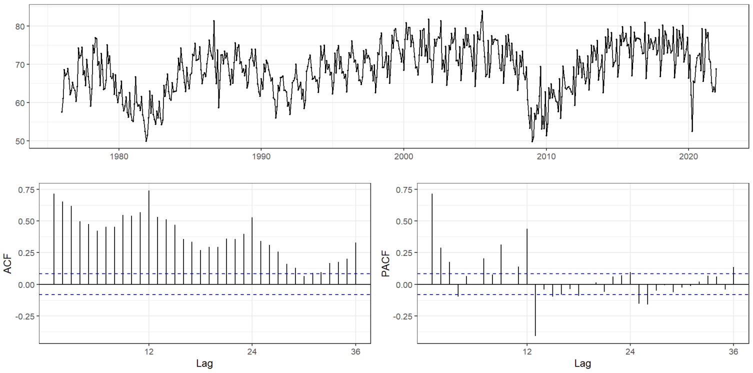


Figure 12: **Acf and Pacf of the Totalnsa with $\lambda = 0.5$**

When the series exhibits a seasonality of period s we can remove seasonality from the time series. We can use the seasonal difference operator:

$$\Delta_s X_t = (1 - L^s) X_t = X_t - X_{t-s} \tag{2}$$

In this case, we can use the seasonal difference operator for monthly data, in which there are 12 periods in a season:

$$\Delta_{12} X_t = (1 - L^{12}) X_t = X_t - X_{12} \tag{3}$$

Seasonal differencing usually removes the gross features of seasonality from a series, as well as most of the trend.

As we can see from Figure 13, the ACF shows that the global autocorrelations for the first lags decay very slowly to 0, this tells us that there is still some trend in the data that needs to be removed.

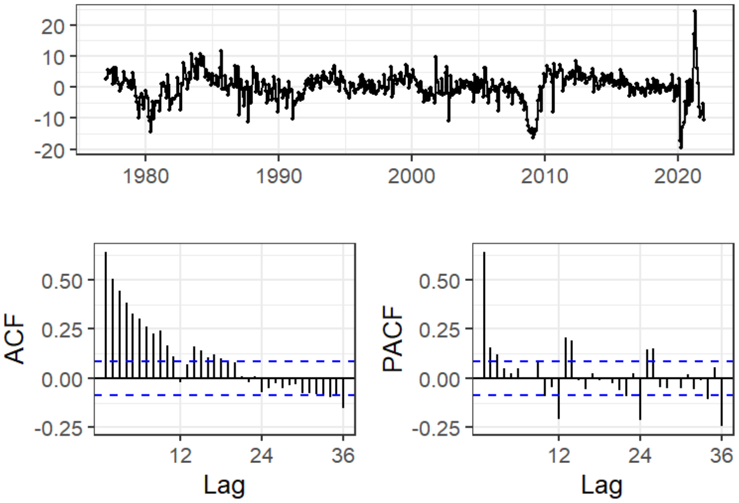


Figure 13: **Acf and Pacf of the seasonal differencing with s = 12**

In this case non-seasonal and seasonal differences can be combined to stationarize the series:

$$\Delta^d \Delta_s^D X_t = (1-L)^d (1-L^s)^D X_t \tag{4}$$

In general, it could possible to consider generic values of d and D:

- D should never be more than 1
- d+D should never be more than 2

As we can see from Figure 14, the time series obtained combined non seasonal and seasonal differences exhibit the typical pattern of a stationary time series. In general, a stationary time series will not have predictable long-term patterns.

Graphically, stationary series are approximately horizontal (although cyclical behavior is possible), with constant variance.

The ACF shows that the global autocorrelations decay very quickly to 0.

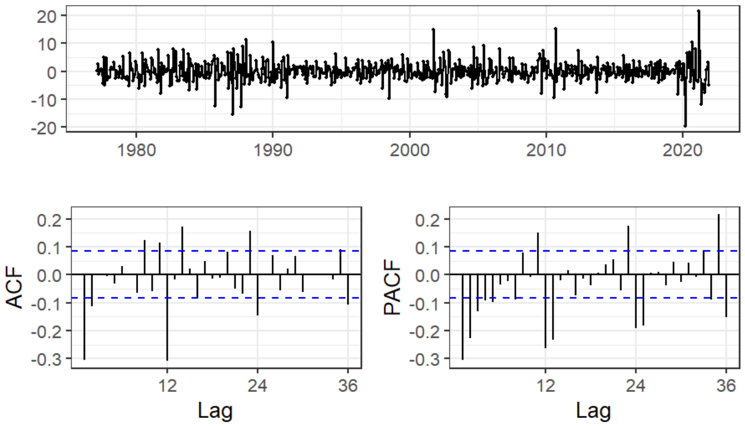


Figure 14: **Acf and Pacf of the combined non seasonal and seasonal differences**

Another empirical criterion for the choice of d is to determine the minimum value for which:

$$Var(\Delta^d X_t) < Var(\Delta^{d+1} X_t) \tag{5}$$

Table 2 shows the variance and the percentage reduction of the variance considering the various transformations carried out on the totalnsa time series.

Table 2: **Variance and reduction in variance %**

Series	**Variance**	**Reduction in variance %**
y_t	41.94202	-
$\Delta_{12} y_t$	20.81683	0.503676
$\Delta\Delta_{12} y_t$	14.65821	0.6505125

Another criterion for verifying stationarity in a time series is to carry out a test. Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given time series is stationary or not. It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series.

Table 3: **Augmented Dickey-Fuller Test (alternative hypothesis: stationary)**

Series	**Dickey-Fuller**	**Lag order**	**p-value**
$\Delta\Delta_{12}y_t$	-8.7314	8	0.01

As we can see from Table 3, $\Delta\Delta_{12}y_t$ is a stationary series. The p-value(0.01) is less than 0.05, so the null hypothesis H0(non-stationary) is rejected.

3.2 Model Identification

Now let's move on to the model identification phase. Model identification involves determining the order of the model.

Using plots of the data, ACF, PACF, and other information, a class of simple ARIMA models is selected. It is usually not possible to tell, simply from a time plot, what values of p and q are appropriate for the data. However, it is sometimes possible to use the ACF plot, and the closely related PACF plot, to determine appropriate values for p and q.

The data may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced data show the following patterns:

- the ACF is exponentially decaying or sinusoidal
- there is a significant spike at lag p in the PACF, but none beyond lag p

The data may follow an ARIMA(0,d,q) model if the ACF and PACF plots of the differenced data show the following patterns:

- the PACF is exponentially decaying or sinusoidal
- there is a significant spike at lag q in the ACF, but none beyond lag q

For the seasonal ARIMA models the seasonal part of an AR or MA model will be seen in the seasonal lags of the PACF and ACF.

The modelling procedure is almost the same as for non-seasonal data, except that we need to select seasonal AR and MA terms as well as the non-seasonal components of the model.

In Figure 15 we analyze the ACF and PACF of the $\Delta\Delta_{12}y_t$ stationary series. Our aim now is to find an appropriate ARIMA model based on the ACF and PACF shown in Figure 15.

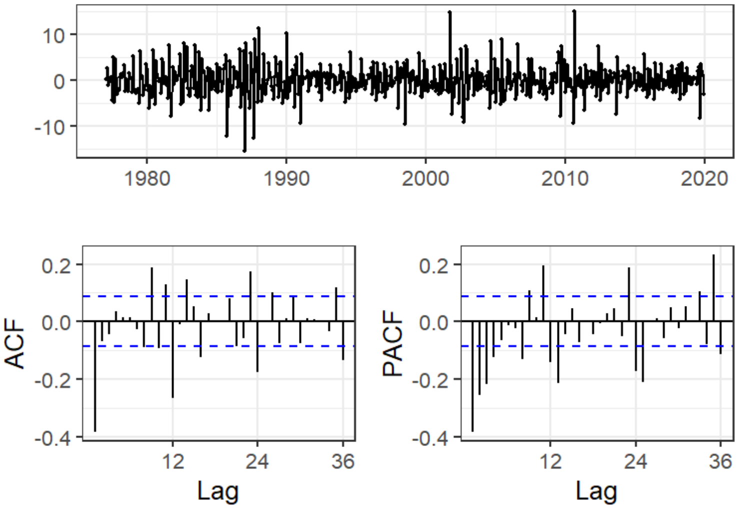


Figure 15: **Acf and Pacf of the combined non-seasonal and seasonal differences on the Training set (January 1976 – December 2019)**

The significant spike at lag 1 in the ACF suggests a non-seasonal MA(1) component, and the significant spike at lag 12 and lag 24 in the ACF suggests a seasonal MA(2) component.

In PACF there seems to be an exponential decay of partial autocorrelations.

Consequently, this initial analysis suggests that a possible model for these data is an $ARIMA(0,1,1)(0,1,2)_{12}$.

We fit this model, along with some variations on it, and compute the various information criteria shown in the Table 4.

Table 4: **Arima models with information criteria**

Model	**AICc**	**BIC**	**HQIC**
$ARIMA(1,1,2)(1,1,2)_{12}$	2448.69	2478.18	2456.451
$ARIMA(1,1,2)(0,1,2)_{12}$	2445.44	2470.74	2451.593
$ARIMA(1,1,1)(0,1,2)_{12}$	2445.62	2466.72	2450.154
$ARIMA(0,1,2)(0,1,2)_{12}$	2445.27	2466.38	2449.81
$ARIMA(0,1,1)(0,1,2)_{12}$	2445.39	2462.29	2448.304

Of these models, the best are $ARIMA(0,1,2)(0,1,2)_{12}$ model (it has the smallest AICc value) and $ARIMA(0,1,1)(0,1,2)_{12}$ model (it has the smallest BIC and HQIC value).

The corrected AIC (Akaike's Information Criterion) is useful for determining the order of an ARIMA model. It can be written as:

- $AICc = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}$

Where AIC can be written as:

- $AIC = -2log(L_{max}) + 2(p+q+k+1)$

$log(L_{max})$ is the maximized value of the log-Likelihood for the estimated model, while the last term in parentheses is the number of parameters in the model.

Another information criteria is the Bayesian Information Criterion (BIC). It can be written as:

- $BIC = AIC + [log(T) - 2](p+q+k+1)$

The difference between BIC and AIC manifests itself when we add a very large number of parameters in order to increase the goodness of fit of the model. In this case, the BIC penalizes this increase in parameters more (compared to the AIC).

Finally, the last information criterion in Table 4 is the Hannan-Quinn information criterion (HQIC). It can be written as:

- $HQIC = -2log(L_{max}) + 2(p+q+k+1)\ log(log(T))$

$log(L_{max})$ is the maximized value of the log-Likelihood for the estimated model, while the second term in parentheses is the number of parameters in the model. T is the number of observations.

Good models are obtained by minimising the AICc, BIC or HQIC.

Now, let us analyze the residuals of the two models chosen by the information criteria. The main objective is to check the adequacy of the models selected in the previous step by looking at the residuals. They should be small and no systematic or predictable patterns should be left in the residuals.

As we can see from Figure 16, the residuals do not have a systematic or predictable patterns. The majority of the global autocorrelations of the residuals (ACF) are within the 95% confidence interval.

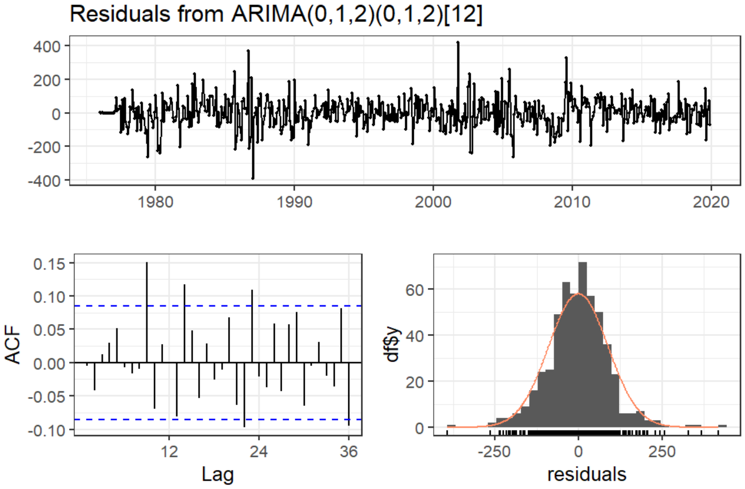


Figure 16: **Check residuals for ARIMA (0,1,2) (0,1,2)$_{12}$**

A very useful test is that of Ljung Box. The Ljung–Box test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags, and is therefore a portmanteau test. Table 5 shows the results of the Ljung Box test.

Q*	**Total lags used**	**p-value**
49.93	24	0.0002267

Table 5: **Ljung Box test for ARIMA (0,1,2) (0,1,2)$_{12}$**

The p-value(0.0002267) is less than 0.05, so the null hypothesis H0 is rejected. This means that at least one of the global autocorrelations of the residuals up to lag 24 is different from 0. We can also see this from the ACF in Figure 16. If we consider the first 8 lags for example, all the global autocorrelations of the residuals are equal to 0, in this case the null hypothesis H0 is not rejected.

To understand whether the residuals exhibit a Normal distribution, we can analyze the Normal Q-Q Plot and perform a test such as Jarque-Bera. Figure 17 shows the Normal QQ-plot. If all the points plotted on the graph perfectly lies on a straight line (y = x) then we can clearly say that this distri- bution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot. In this case the residuals do not have a normal distribution.

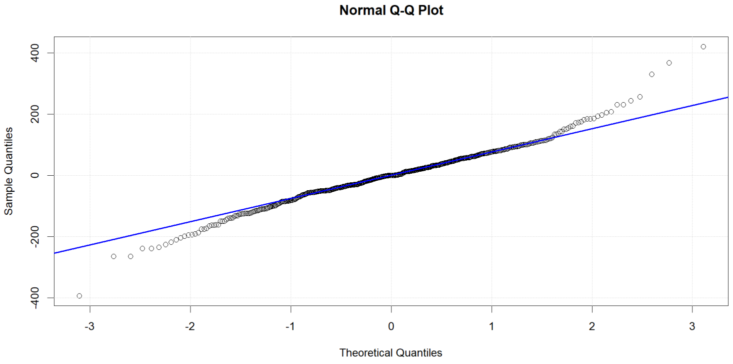


Figure 17: **Normal QQ PLOT**

A very useful test is Jarque–Bera. In statistics, the Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test statistic is always nonnegative. If it is far from zero, it signals the data do not have a normal distribution.

The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero. Samples from a normal distribution have an expected skewness of 0 and an expected excess kurtosis of 0 (which is the same as a kurtosis of 3). Table 6 shows the results of the Jarque-Bera test.

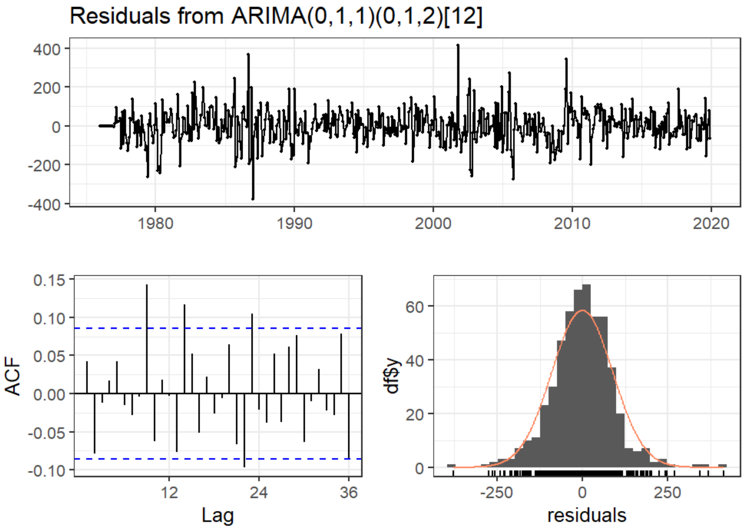
X-squared	**p-value**
114.73	2.2e-16

Table 6: **Jarque-Bera test**

The p-value(2.2e-16) is less than 0.05, so the null hypothesis H0 is rejected. The residuals do not have a normal distribution.

Now, let us analyze the residuals of the $ARIMA(0,1,1)(0,1,2)_{12}$ model. Steps are the same. Again, as we can see from Figure 18, the residuals do not have a systematic or predictable patterns. The majority of the global autocorrelations of the residuals (ACF) are within the 95% confidence interval.

Table 7 shows the results of the Ljung Box test, the p-value(0.0003754) is less than 0.05, so the null hypothesis H0 is rejected. Again, at least one of the global autocorrelations of the residuals up to lag 24 is different from 0.

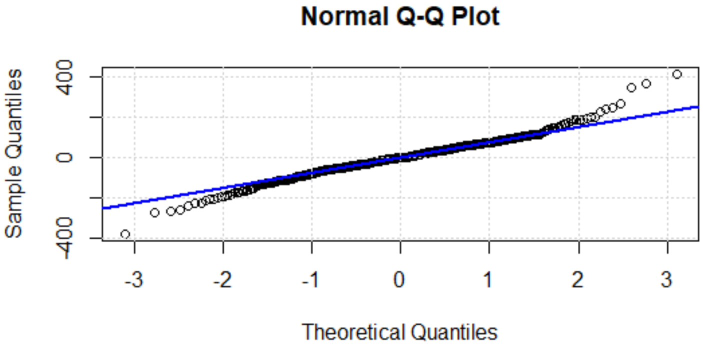


Q*	Total lags used	p-value
49.911	24	0.0003754

Table 7: Figure 18: **Ljung Box test and check residuals and for ARIMA (0,1,1) (0,1,2)**$_{12}$

Again, to understand whether the residuals exhibit a Normal distribution, we can analyze the Normal Q-Q Plot and perform a test such as Jarque-Bera.

As we can see from Figure 19, the quantiles of our residuals and theoretical quantiles do not follow the QQ line almost perfectly. For this reason, the QQ plot indicates that our residuals do not have a normal distribution, while the Table 8 shows the results of the Jarque-Bera test, the p-value (2.2e-16) less than 0.05, also the Jarque-Bera test confirms that the residuals do not have normal distribution.



X-squared	p-value
110.22	2.2e-16

Table 8: Figure 19: **Jarque-Bera test and Normal QQ-PLOT**

3.3 Forecasting ARIMA models

Now, we will compare some models fitted so far using a test set consisting of the last two years of data. Thus, we fit the models using data from January 1976 to December 2019, and forecast the total vehicle sales for January 2020 – December 2021. The results are summarised in Table 9.

Table 9: **Arima models with accuracy measures**

Model	RMSE	MAE	MAPE	MASE
$ARIMA(1,1,2)(1,1,2)_{12}$	299.77	226.69	20.89	2.1658
$ARIMA(1,1,2)(0,1,2)_{12}$	299.98	227.01	20.92	2.168
$ARIMA(1,1,1)(0,1,2)_{12}$	299.62	226.61	20.887	2.1651
$ARIMA(0,1,2)(0,1,2)_{12}$	299.49	226.53	20.880	2.164
$ARIMA(0,1,1)(0,1,2)_{12}$	300.21	227.21	20.94	2.17

All accuracy measures in table 9 choose the $ARIMA(0,1,2)(0,1,2)_{12}$ model which has the lowest RMSE, MAE, MAPE and MASE value on the test set.

The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

- Mean absolute error: MAE $= mean\ (|\hat{e}_t|)$
- Root mean squared error: RMSE $= \sqrt{mean\ (\hat{e}_t^2)}$

Mape is a measure of accuracy based on the concept of percentage errors.

Percentage errors have the advantage of being unit-free, and so are frequently used to compare forecast performances between data sets.

Measures based on percentage errors have the disadvantage of being infinite or undefined if $x_t = 0$ for any t in the period of interest, and having extreme values if any x_t is close to zero.

The percentage error is given by $p_t = 100\ \frac{\hat{e}_t}{x_t}$. The most commonly used measure is:

- Mean absolute percentage error: MAPE $= mean\ (|p_t|)$

An alternative to using percentage errors when comparing forecast accuracy across series with different units are scaled errors, for example MASE. The mean absolute scaled error is simply:

- Mean absolute scaled error: MASE $= mean\ (|q_j|)$

Where:

$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|x_t - x_{t-1}|}$

For a non-seasonal time series, and:

$q_j = \frac{e_j}{\frac{1}{T-m}\sum_{t=m+1}^{T}|x_t - x_{t-m}|}$

For seasonal time series.

Figure 20 shows the $ARIMA(0,1,2)(0,1,2)_{12}$ model that minimized the various accuracy measures seen in Table 9. This model manage to follow the true trend of test set data quite closely.

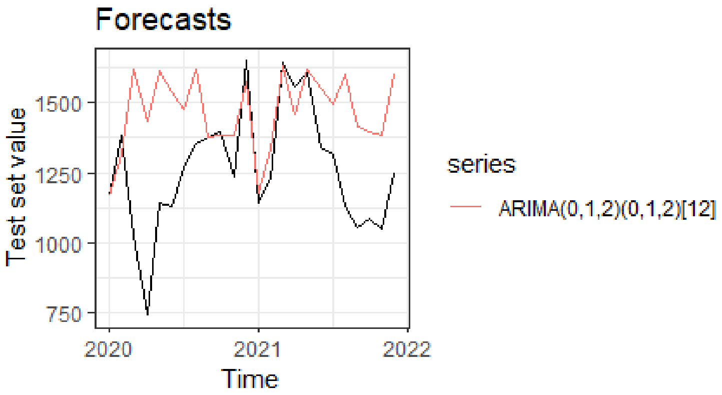


Figure 20: **Forecasts on the Test set (January 2020–December 2021)**

The predictions obtained in Figure 20 and the accuracy measures obtained in Table 9 refer to the original scale. After choosing a transformation, we need to predict using the transformed data. Then, we need to invert the transformation (back-transform) to obtain predictions on the original scale.

The inverse Box-Cox transformation in this case is given by:

$$x_t = \left\{(0.5y_t + 1)^{\frac{1}{0.5}}\right. \tag{6}$$

Where 0.5 is the lambda parameter and y_t is the time series to which the Box-Cox transformation has been applied.

However, the point forecast from the "back-transformation" can be considered the median and not the mean of the distribution of the forecasts (assuming that the distribution over the transformed space is symmetric). Obviously this is acceptable, but in some cases an "average forecast" is required.

The back-transformed mean in this case is given by:

$$x_t = \left\{(0.5y_t + 1)^{\frac{1}{0.5}} \left[1 + \frac{\sigma_h^2(1-0.5)}{2(0.5y_t+1)^2}\right]\right. \tag{7}$$

Where σ_h^2 is the h-step forecast variance. The larger the forecast variance, the bigger the difference between the mean and the median. The difference between the simple back-transformed forecast and the back-transformed mean is called the bias.

Figure 21 shows the difference between the forecast medians and the forecast means on the test set with $ARIMA(0,1,2)(0,1,2)_{12}$ model.

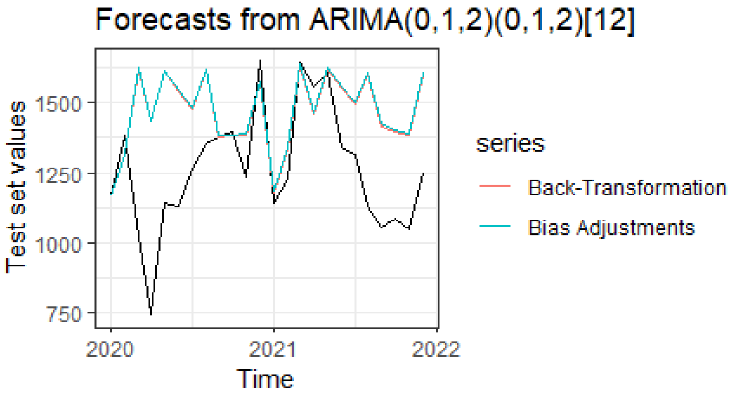


Figure 21: **Forecast medians and forecast means on the test set with ARIMA (0,1,2)(0,1,2)$_{12}$ model**

There is a small difference between back-transformation and bias adjustment predictions.

Lastly, we use $ARIMA(0,1,2)(0,1,2)_{12}$ model to predict the number of total vehicle sales considering a time horizon of two years(h = 24), to get an idea of how the phenomenon will evolve in the near future.

Figure 22 shows the point forecast (Back-Transformed) for the possible number of total vehicle sales in the period January 2022-December 2023, in addition to the point forecast, the 95% and 80% confidence intervals are also shown.

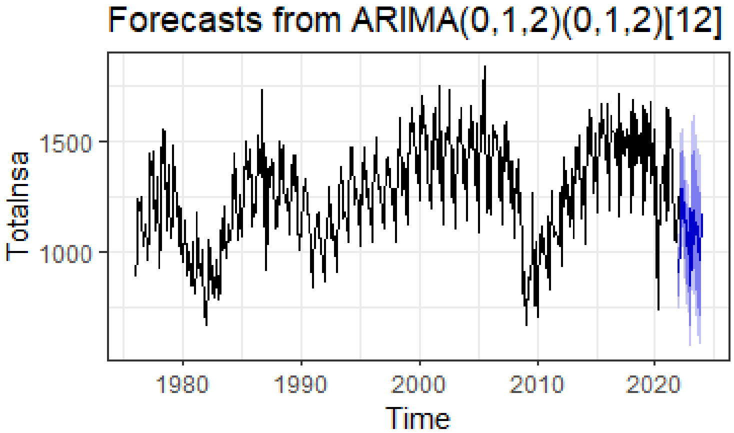


Figure 22: **Forecasts considering a time horizon of two years(January 2022-December 2023)**