FORECASTING RETAIL SALES DATA IN THE US

Pietro Ruffo

02/12/2021

—

Prediction and Decision Processes

—

António José Lopes Rodrigues

1. INTRODUCTION

The initial data refers to the Total monthly Retail sales data in the US (excluding Food Services) from January 1992 to December 2017. It accounts for a total of 312 observations.
When analysing an economic variable, such as sales, interest is often not focused on the absolute value of the variable, but rather on relative changes. In this case, a reference base has been identified and the dynamics of the phenomenon are assessed relative to the base.
Data was initially in nominal dollars and has been adjusted using consumer price index at 2017 to eliminate the effect of the inflation during the years. Having done that, the total monthly value has been divided for the number of days in each month to get the real value of sales per day:

$$Sales\ per\ day = \frac{(Total\ Monthly\ Sales\ /\ 2017\ Consumer\ Price\ Index)}{Number\ of\ days\ of\ that\ month}$$

The resulting series indicates the number of sales per day in the US in 2017 dollars and the goal of the analysis is to forecast the next two years of sales per day.

Theoretically a time series can be represented by the following model:

$$Y_t = f(t) + u_t$$

$$u_t \sim WN(0, \sigma_u^2)$$

It is assumed that the deterministic component, $f(t)$, is the result of the action of two components: trend and seasonality. In classical approach to the analysis of Time Series it is assumed that exists a law of temporal evolution of the phenomenon represented by $f(t)$.

The random component u_t represents the set of circumstances, each of negligible entity, that cannot be explicitly considered in Y_t. The residuals of Y_t., not explained by $f(t)$, are attributed to chance and assimilated to accidental errors.

From a statistical point of view, it is assumed that u_t is generated by a white noise process, i.e. by a succession of iid random variables, of zero mean and constant variance.

2. GRAPHICAL ANALYSIS

First of all, the series will be analysed graphically to detect any seasonal or trend components.

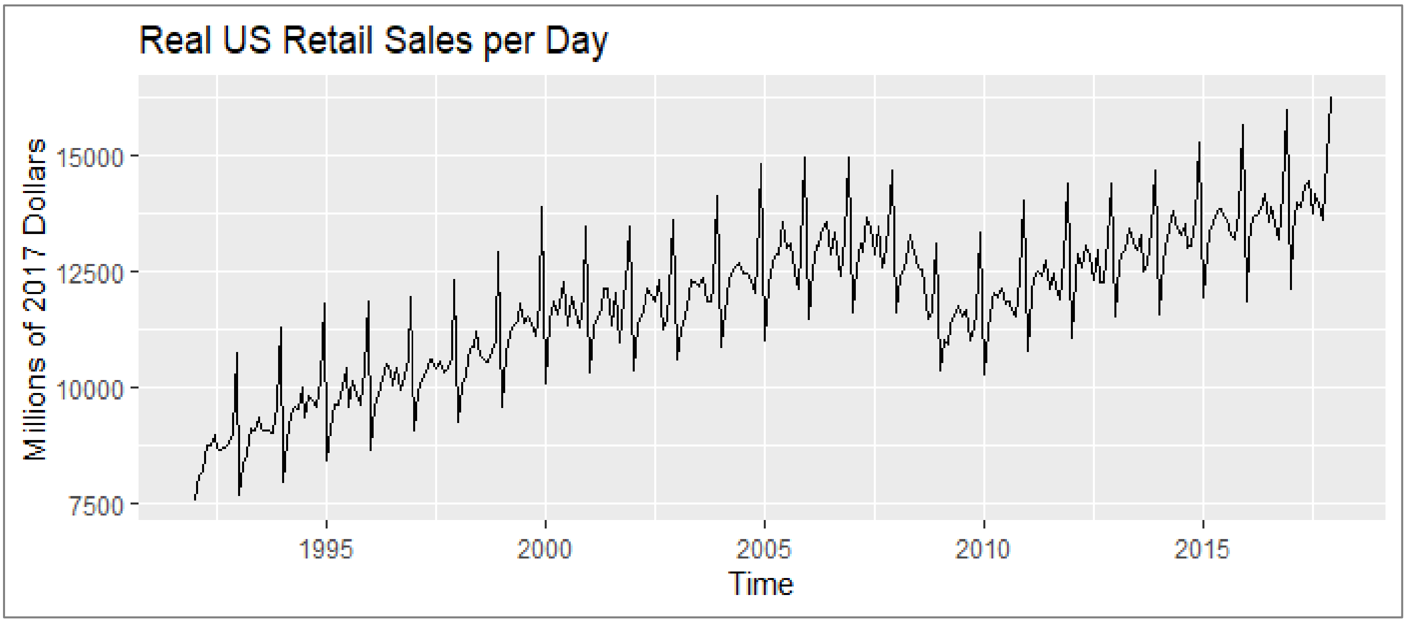

Figure 1 "Real US Retail Sales per Day in 2017 US Dollars"

The graph shows a seasonal pattern evolving on a positive trend. The negative effects of the financial crises on retailers are shown on the plot represented by a negative hump around the year 2008.

Both of the seasonal and trend components are not growing exponentially and for this reason it will be assumed an additive relationship between the deterministic components of the model:

$$Y_t = T_t + S_t + u_t$$

$$u_t \sim WN(0, \sigma_u^2)$$

3. PRE-PROCESSING

Trend and Seasonality will be estimated by using OLS method. Having done this, it will possible to obtain two different time series a series adjusted for seasonality as well as a series adjusted for trend.
Seasonality is made up of movements of the phenomenon during the year which tend to be repeated in an almost similar way during the same period in successive years. These movements can have various causes: succession of seasons, climatic variations, traditions, etc. The seasonal component of a time series can be represented by a periodic function *g(t)* whose value at time t is exactly reproduced at constant intervals, whose length *s* is the period.

The regression model can be used to estimate the seasonal component using the method of auxiliary variables (dummy variables).
It is assumed for the moment that only the seasonal component is present in the series:

$$Y_t = S_t + u_t$$

$$u_t \sim WN(0, \sigma_u^2)$$

Let us assume that the periodic function can be represented by:

$$g(t) = \sum_{j=1}^{s} \gamma_j d_{j,t} = \gamma_1 d_{1,t} + \gamma_2 d_{2,t} + \ldots + \gamma_s d_{s,t} \qquad t=1,2,\ldots,n$$

In the case of the data under analysis s=12 because are monthly observations and n=312 because that is the number of the total observations. Where *d* is a dummy variable:

$$d_{j,t} = \begin{cases} 1 & \text{if the observation } t \text{ corresponds to the month } j \\ 0 & \text{else} \end{cases}$$

By assuming that g(t)= S_t the following model (in matrix form) is used to estimate the raw coefficients of seasonality γ:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{u}$$

γ is the vector containing the coefficients of seasonality that have to be estimated. Those are regression coefficients representing the average level of the phenomenon at each of the 12 months of the year.

$\gamma = (\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_{12})^T$

D is a 312x12 matrix:

```
> head(D)
     d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12
[1,]  1  0  0  0  0  0  0  0  0   0   0   0
[2,]  0  1  0  0  0  0  0  0  0   0   0   0
[3,]  0  0  1  0  0  0  0  0  0   0   0   0
[4,]  0  0  0  1  0  0  0  0  0   0   0   0
[5,]  0  0  0  0  1  0  0  0  0   0   0   0
[6,]  0  0  0  0  0  1  0  0  0   0   0   0
> tail(D)
       d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12
[307,]  0  0  0  0  0  0  1  0  0   0   0   0
[308,]  0  0  0  0  0  0  0  1  0   0   0   0
[309,]  0  0  0  0  0  0  0  0  1   0   0   0
[310,]  0  0  0  0  0  0  0  0  0   1   0   0
[311,]  0  0  0  0  0  0  0  0  0   0   1   0
[312,]  0  0  0  0  0  0  0  0  0   0   0   1
```

The model estimated through linear regression is the following:

```
Call:
lm(formula = Y ~ d1 + d2 + d3 + d4 + d5 + d6 + d7 + d8 + d9 +
    d10 + d11 + d12 - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3429.5 -1038.7   184.1  1146.8  2759.4

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
d1   10291.4      287.7   35.77   <2e-16 ***
d2   11249.8      287.7   39.10   <2e-16 ***
d3   11623.1      287.7   40.39   <2e-16 ***
d4   11804.3      287.7   41.02   <2e-16 ***
d5   12051.3      287.7   41.88   <2e-16 ***
d6   12153.0      287.7   42.23   <2e-16 ***
d7   11715.7      287.7   40.72   <2e-16 ***
d8   11984.8      287.7   41.65   <2e-16 ***
d9   11555.5      287.7   40.16   <2e-16 ***
d10  11510.3      287.7   40.00   <2e-16 ***
d11  12124.0      287.7   42.13   <2e-16 ***
d12  13766.7      287.7   47.84   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1467 on 300 degrees of freedom
Multiple R-squared:  0.9855,    Adjusted R-squared:  0.9849
F-statistic:  1694 on 12 and 300 DF,  p-value: < 2.2e-16
```

Where the intercept has been subtracted not to fall in the dummies trap, i.e., linear dependent independent variables.

The seasonally adjusted time series with regressioncoefficients coincides with the series of (estimated) residuals.

$$y_t^{*d} = y_t - \hat{g}(t)$$

Graphically it has been scaled by the minimum of the above formula:

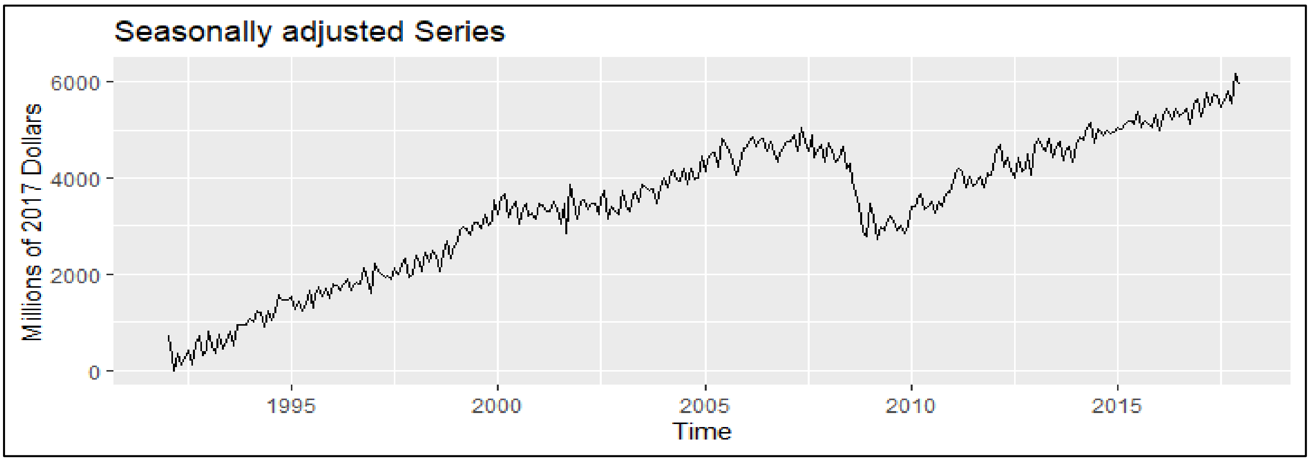

Figure 2 "Seasonally Adjusted Series"

Let us assume now that the systematic part consists only of the trend:

$$Y_t = T_t + u_t \qquad u_t \sim WN(0, \sigma_u^2)$$

The graphical analysis of the initial data suggested that a polynomial function could describe the evolution of the series. A reasonable criterion for choosing the order of the polynomial is based on the correct version of the coefficient of determination R^2. In particular, denoted by R_q^2 the adjusted coefficient of determination for the estimate of a polynomial of order q, if:

$$\overline{R}_q^2 \geq \overline{R}_{q+1}^2$$

then a polynomial of degree q is chosen, otherwise the search is continued for higher values of q. Obviously, if the above inequality is satisfied, it means that the t^{q+1} component did not make a significant contribution to the fit of the estimated model to the data.

The results show the trend is a 4th degree polynomial:

1st degree polynomial -> Adjusted R-squared: 0.6532
2nd degree polynomial -> Adjusted R-squared: 0.6905
3st degree polynomial -> Adjusted R-squared: 0.7313
4th degree polynomial -> Adjusted R-squared: 0.7446
5th degree polynomial -> Adjusted R-squared: 0.7441

Fitting the 4th degree polynomial:

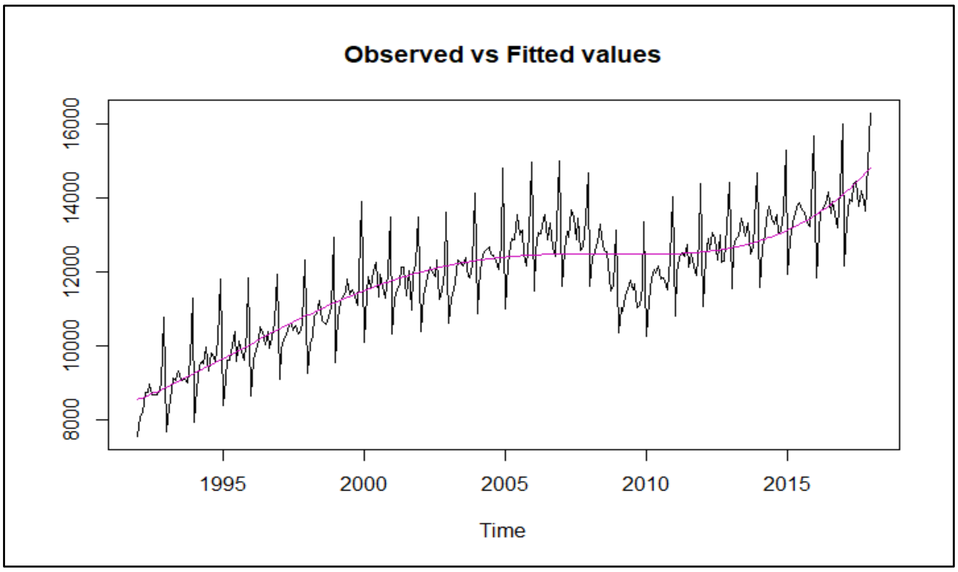

Figure 3 "Trend Estimation with a 4th degree polynomial: Observed vs Fitted Values"

By extracting the residual, the trend adjusted series is obtained:

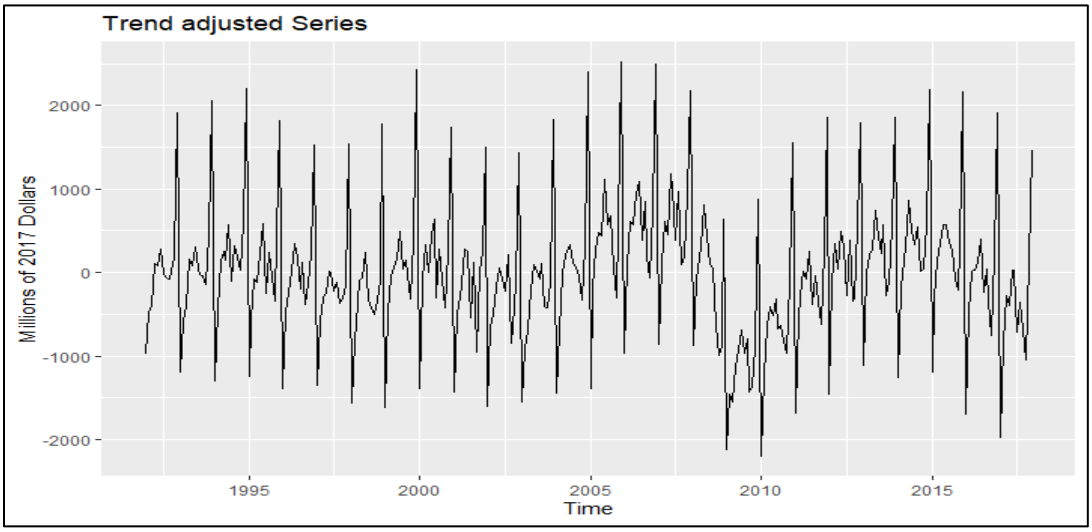

Figure 4 "Trend Adjusted Time Series"

It is possible to estimate the trend and the seasonal components at the same time combining the polynomial trend and the seasonal component expressed through the dummy variables.

The resulting model is:

```
Call:
lm(formula = Y ~ d1 + d2 + d3 + d4 + d5 + d6 + d7 + d8 + d9 +
    d10 + d11 + d12 - 1 + poly(t, 4, raw = T))

Residuals:
     Min       1Q   Median       3Q      Max
-1438.83  -193.13    20.85   261.41   913.21

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
d1                    7.198e+03  1.382e+02  52.070  < 2e-16 ***
d2                    8.137e+03  1.386e+02  58.728  < 2e-16 ***
d3                    8.492e+03  1.389e+02  61.149  < 2e-16 ***
d4                    8.654e+03  1.392e+02  62.185  < 2e-16 ***
d5                    8.881e+03  1.394e+02  63.694  < 2e-16 ***
d6                    8.964e+03  1.397e+02  64.164  < 2e-16 ***
d7                    8.507e+03  1.399e+02  60.786  < 2e-16 ***
d8                    8.756e+03  1.402e+02  62.465  < 2e-16 ***
d9                    8.307e+03  1.404e+02  59.171  < 2e-16 ***
d10                   8.242e+03  1.406e+02  58.622  < 2e-16 ***
d11                   8.835e+03  1.408e+02  62.761  < 2e-16 ***
d12                   1.046e+04  1.409e+02  74.196  < 2e-16 ***
poly(t, 4, raw = T)1  2.180e+01  5.166e+00   4.220 3.25e-05 ***
poly(t, 4, raw = T)2  2.453e-01  6.694e-02   3.664 0.000294 ***
poly(t, 4, raw = T)3 -2.141e-03  3.211e-04  -6.669 1.26e-10 ***
poly(t, 4, raw = T)4  4.268e-06  5.089e-07   8.386 2.07e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 405.5 on 296 degrees of freedom
Multiple R-squared:  0.9989,	Adjusted R-squared:  0.9988
F-statistic: 1.686e+04 on 16 and 296 DF,  p-value: < 2.2e-16
```

The adjusted R squared of the model is near 1 indicating a high goodness of fit. This can be confirmed by plotting both the observed and the estimated values:

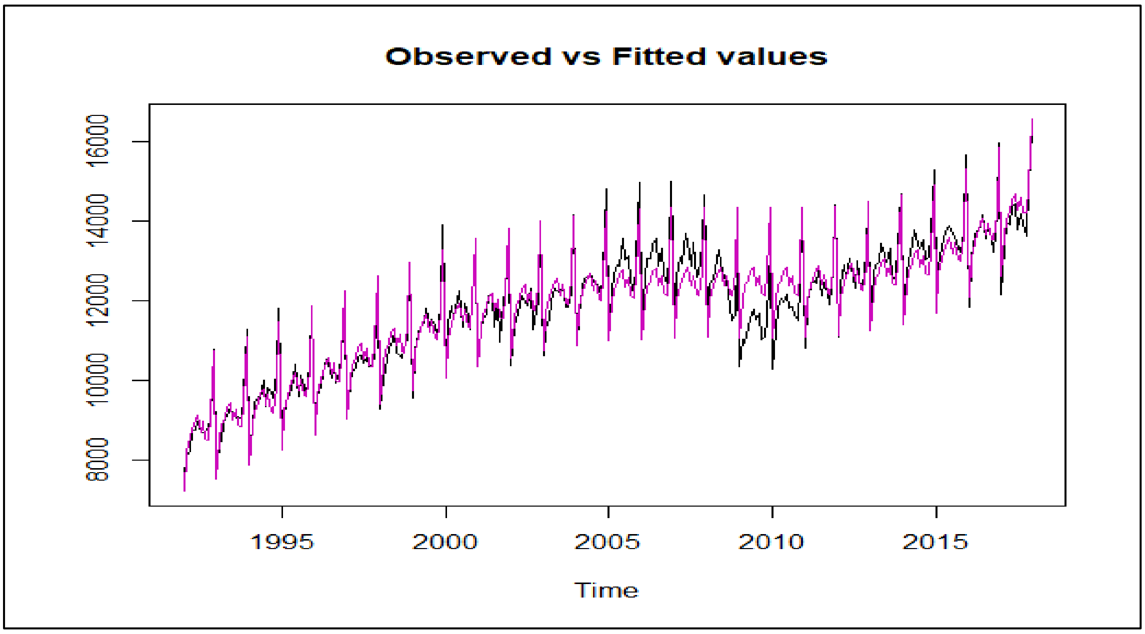

Figure 5 "Trend a d Seasonality Estimation, observed vs fitted Values"

4. FORECASTING

Forecasts will be performed on two time series: the one adjusted for seasonality (Figure 2) and on the original data (Figure 1).
The first one will be modelled using the Gadner-McKenzie Method while on the second a Holt-Winters model will be built; Both of them belong to the exponential smoothing methods which are statistical techniques and procedures for discrete time series data used to forecast the immediate future.
For such purpose the analysis is going to be conducted using the library "forecast" which accounts with methods and tools for displaying and analysing univariate time series forecasts including exponential smoothing.
Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations get older. In other words, the more recent the observation the higher the associated weight.

ACF for the original data:

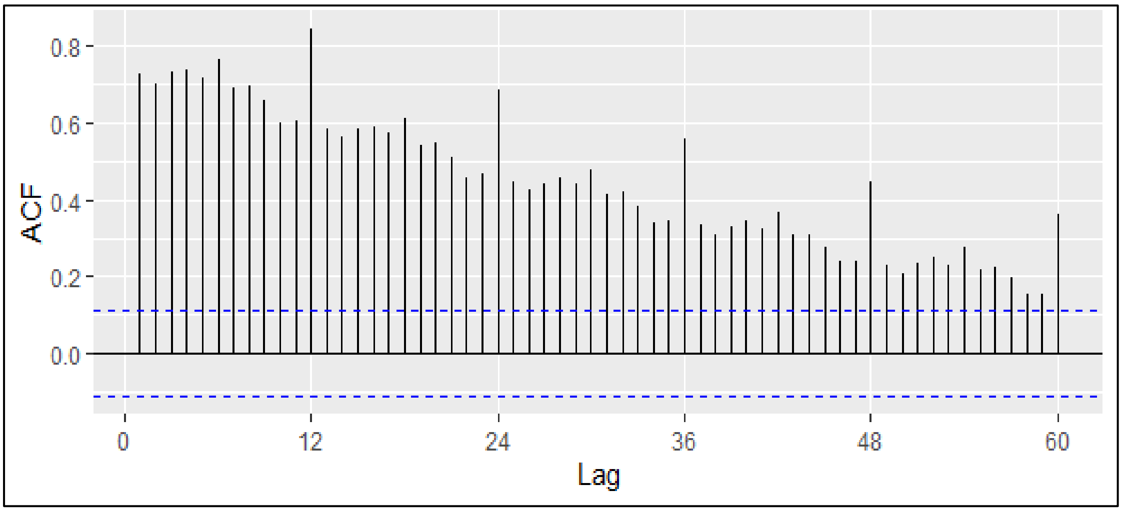

Figure 6"ACF for the original data"

ACF for the seasonality adjusted time series:

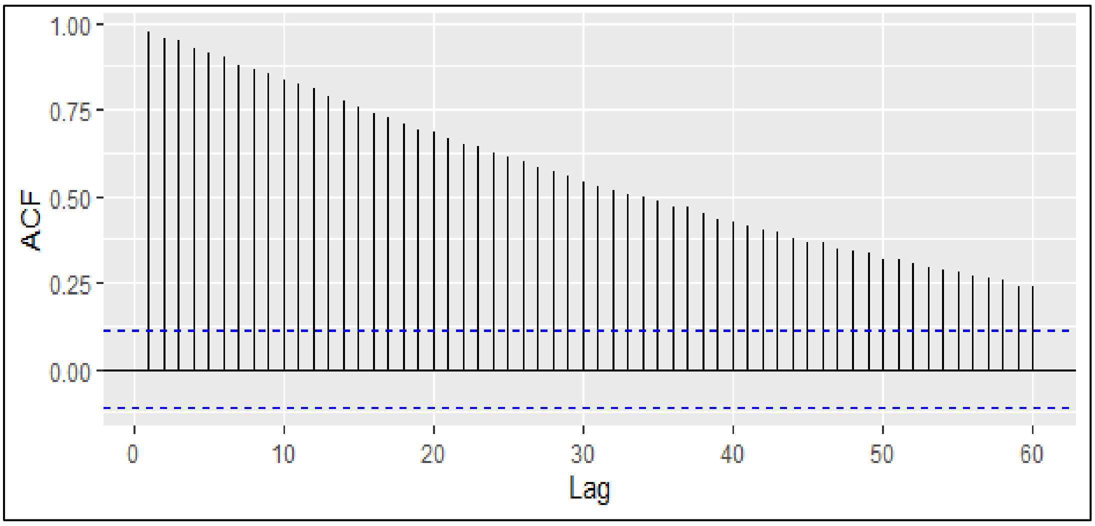

Figure 7 "ACF seasonality adjusted time series"

4.1 GADNER-MCKENZIE METHOD

Since the seasonal adjusted time series seems to evolve on a damped trend, it will be modelled using the Gadner-McKenzie method. The Gadner-McKenzie method is a variation of the Holt method specifically suited for damped trends. Such methods support changing trends with time in different ways, either additively or multiplicatively, counted on if the trend is linear or exponential correspondingly.

This method accounts with three parameters: alpha, beta and phi.

Alpha parameter is used to determine the smoothing factor for a particular level. The rate at which the weights decrease is controlled by the parameter alpha. For any alpha between 0 and 1, the weights attached to the observations decrease exponentially as time goes back, (that is the reason why is named "exponential smoothing.") If alpha is small (close to 0), more weight is given to observations from the more distant past. If alpha is large (approximately 1), more weight is given to the more recent observations.

On the other hand, Beta is used to control the decay of the impact of trends changes while phi is the damping parameter (with values between 0 and 1). If phi=1, the method is identical to Holt's linear method. For values between 0 and 1, phi dampens the trend so that it approaches a constant some time in the future. This means that short-run forecasts are trended while long-run forecasts are constant.

The level and the trend equations, needed for the forecasts, are computed optimizing parameters three parameters above as follows:

$$a_k = \alpha\, y_k + (1-\alpha)\left(a_{k-1} + \phi\, b_{k-1}\right) \qquad (0 < \alpha \le 1) \qquad \text{Level equation}$$

$$b_k = \beta\left(a_k - a_{k-1}\right) + (1-\beta)\, \phi\, b_{k-1} \qquad (0 < \beta \le 1) \qquad \text{Trend Equation}$$

Where $k = 1,\ldots,312$

Where a_k denotes an estimate of the level of the series at time k and b_k denotes an estimate of the trend (i.e. slope) of the series at time k.

To perform the Gadner-McKenzie on the time series adjusted for seasonality the R's function holt() will be used. This function returns forecasts and other information for exponential smoothing forecasts applied to the time series. The unknown parameters and the initial values for any exponential smoothing method can be estimated by minimising the SSE. Since the adjusted for seasonality time series seems to evolve on a damped trend the parameter of the function damped is set to TRUE, while the damping parameter phi has been automatically estimated as well as the smoothing parameter for the level alpha.

Different values of beta have been plugged into the function and the results can be seen plotting the cost function associated with the different trials:

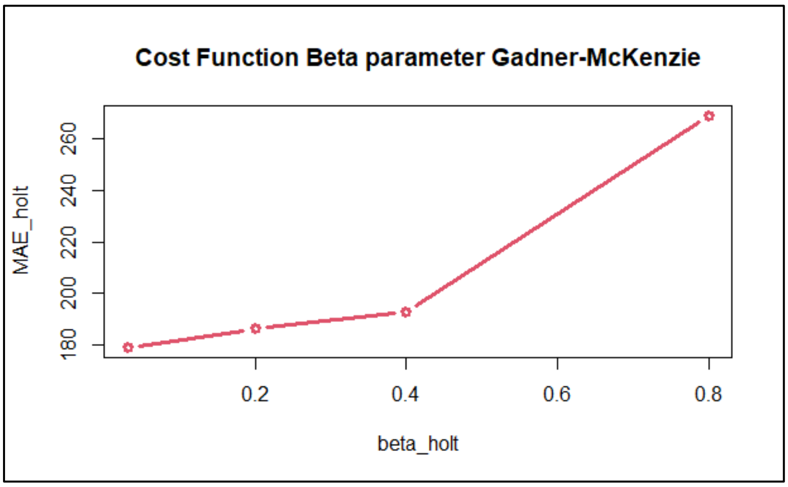

Figure 8 "Cost Function Beta Parameter Gadner-McKenzie"

The optimal model, performing the lowest MAE (MAE=180.0497) assume the following values for the smoothing parameters:

alpha = 0.3946
beta = 0.0315
phi = 0.9457

By plugging these values into both the Level and the Trend equations it will be possible to obtain the initial state values:

a_0 = 245.0074
b_0 = 24.0969

Once the parameters have been estimated it will be possible to compute the estimations for the observed values as well as the forecasted values 24 step ahead (2-year forecasting) according to the formula:

$$\hat{y}_{k+p|k} := a_k + \left(\sum_{i=1..p} \phi^i\right) b_k \qquad (0 \leq k \leq N;\ p \geq 0)$$

Graphically:

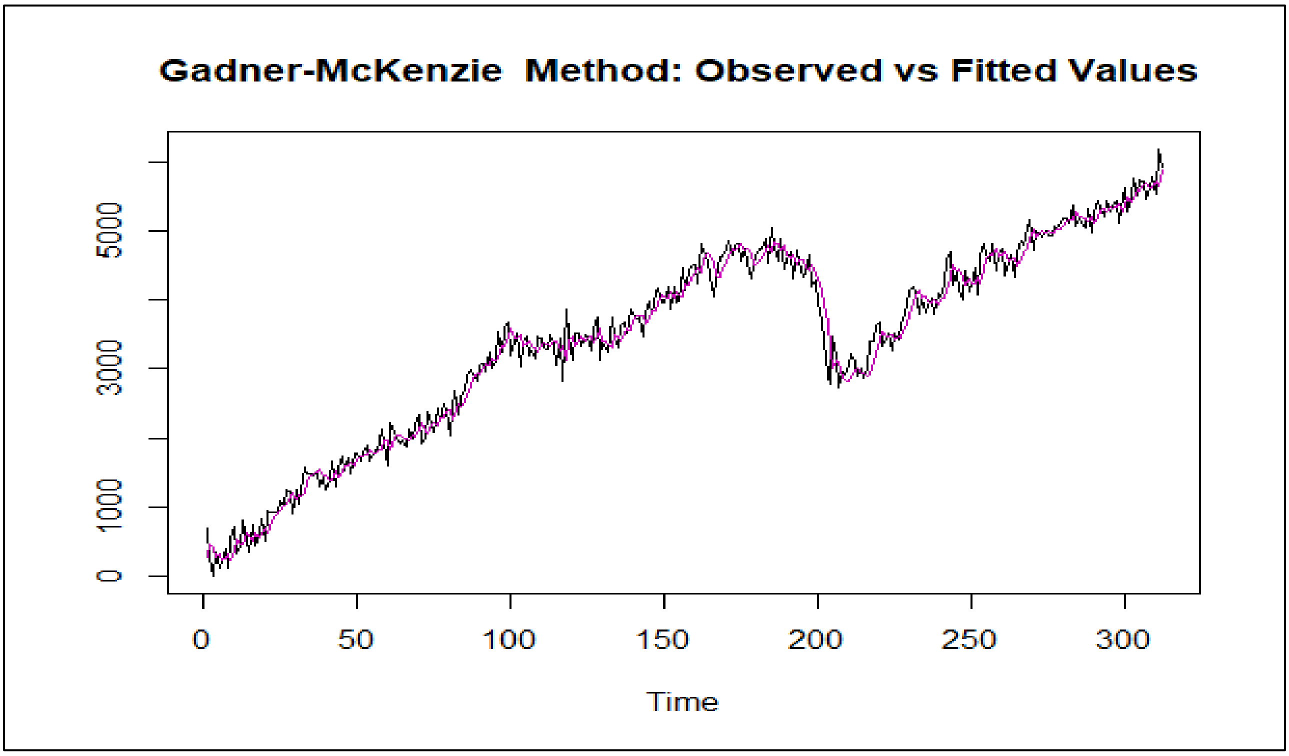

Figure 9"Gadner-McKenzie Method: Observed vs Fitted Values."

Forecasting for the next two years using the Gadner-McKenzie Method on the Seasonal Adjusted Time Series:

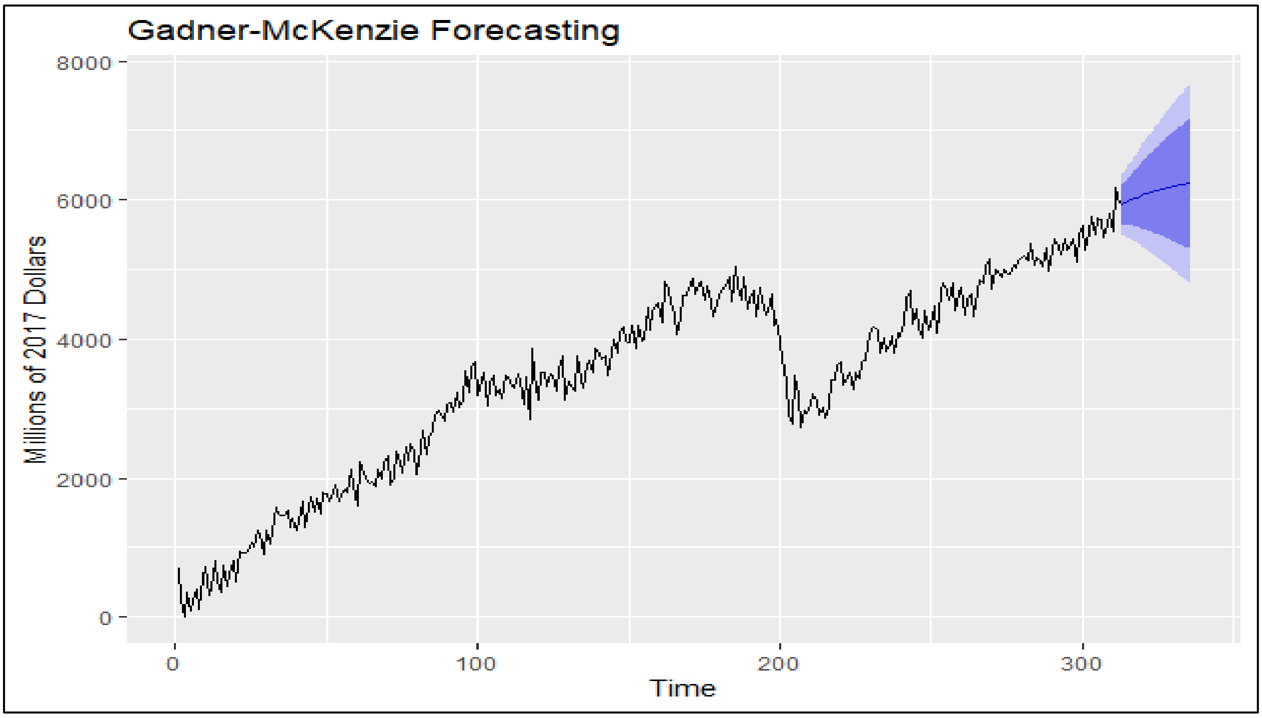

Figure 10 "Gadner-McKenzie Forecasting"

According to the graph at the end of 2019 US retailers will sell a daily amount of 6244.645 USD dollars (of 2017) with a 95% confidence interval of (Lower bound: 4796.889, Upper Bound 7692.402).

For completeness the same time series as been modelled using the Holt Method, which accounts with just the parameters Alpha and Beta. The estimates for the two parameters of the final model were Alpha=0.449, Beta=1e-04. The performances of the two model are pretty similar:

Gadner-McKenzie	**Holt Method**
MAE=180.0497	MAE = 178.8047
AIC=5178.622	AIC= 5176.671
BIC=5201.080	BIC=5195.386

The slightly lower AIC and BIC in the Holt Model[1] (linear trend) are justified by the fact that such model has a parameter less (phi) compared to the Gadner-McKenzie final model. Holt linear model tends to over-forecast, especially for longer forecast horizons.

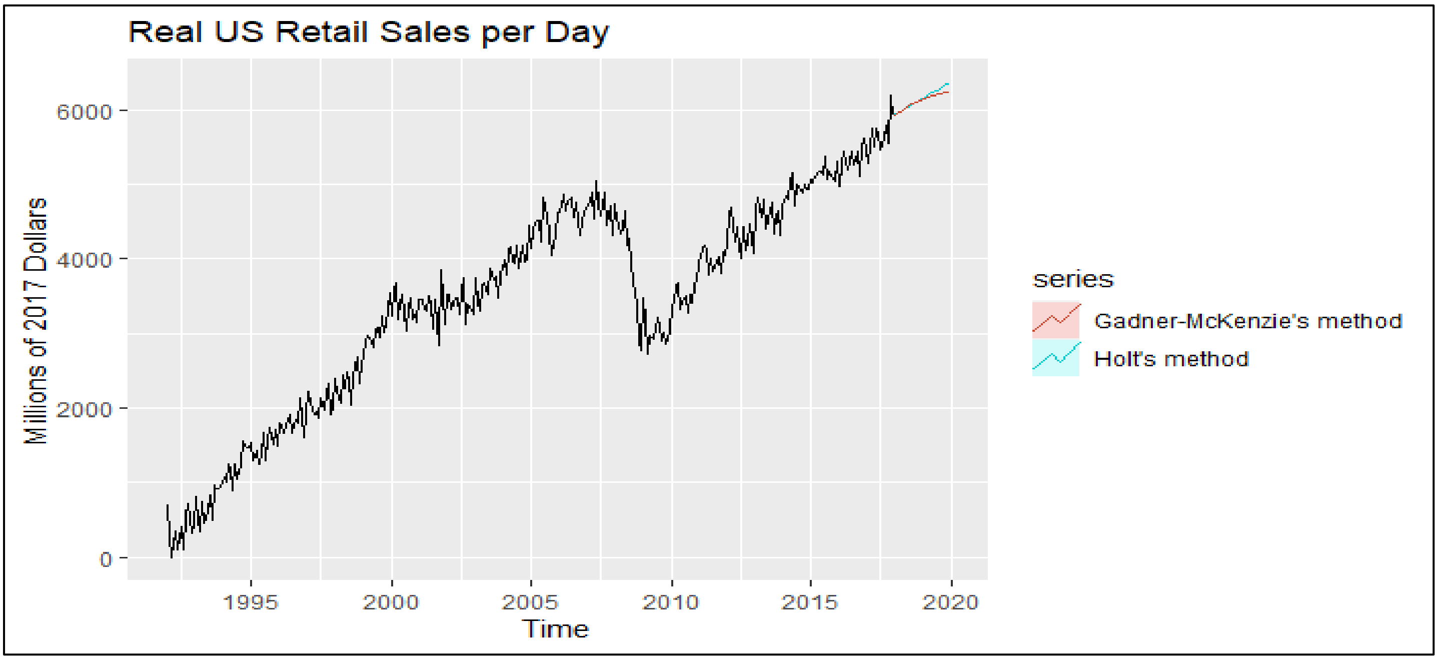

Figure 11 "Damped vs Linear Holt"

[1] The code part related to the application of Holt Linear Model is contained in the "Extra Ideas" section of the R code.

HOLT-WINTERS

The Holt-Winters exponential smoothing models the level, trend and seasonality patterns to change over time as it is an adaptive method. Since the seasonal variations are roughly constant through the series the additive version of the method will be used.
Beside the two smoothing factors, alpha and beta, an additional new factor is introduced, called gamma in order to control the impact on the seasonal element.

The parameters of the models are estimated as follow optimizing the three smoothing parameters:

$$a_k = \alpha\left(y_k - c_{k-m}\right) + (1-\alpha)\left(a_{k-1} + b_{k-1}\right) \qquad (0 < \alpha < 1) \qquad \text{Level equation}$$

$$b_k = \beta\left(a_k - a_{k-1}\right) + (1-\beta)\, b_{k-1} \qquad (0 < \beta < 1) \qquad \text{Trend equation}$$

$$c_k = \gamma\left(y_k - a_k\right) + (1-\gamma)\, c_{k-m} \qquad (0 < \gamma < 1) \qquad \text{Equation for the seasonal component}$$

Where m denotes the frequency of the seasonality.

The level equation shows a weighted average between the seasonally adjusted observation and the non-seasonal forecast for time k. The trend equation shows that b_k is a weighted average of the estimated trend at time k based on the previous estimate of the trend. The seasonal equation shows a weighted average between the current seasonal index, and the seasonal index m periods ago.
In correspondence with the trend, seasonality can be modelled in the particular of additive or multiplicative process for the linear and exponential variation in the seasonality.
This time the analysis will be performed on the initial data that accounts with both seasonal and trend components.
The seasonal plot emphasizes the seasonal pattern plotting the data against the individual "seasons" in which the data were observed:

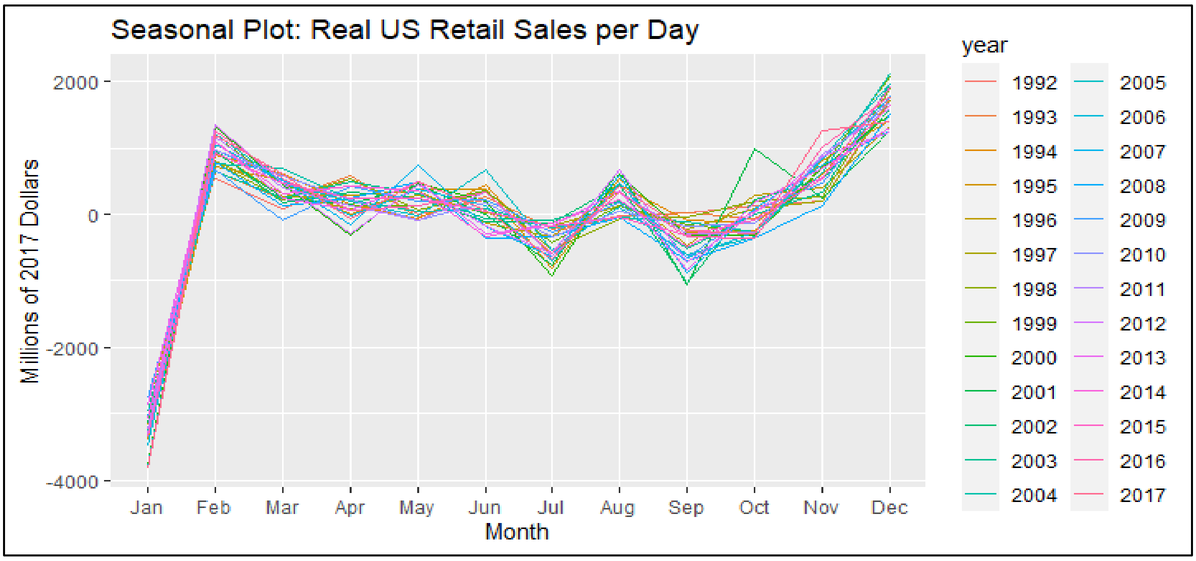

Figure 12 "Seasonal Plot"

For this purpose, the function hw () will be used plugging in the original time series. The parameter “seasonal” is set to additive; the smoothing parameter for the trend beta and the smoothing parameter for seasonality gamma will be optimized by the function minimizing the SSE. Different values for the smoothing factor for levels alpha will be plugged into the function:

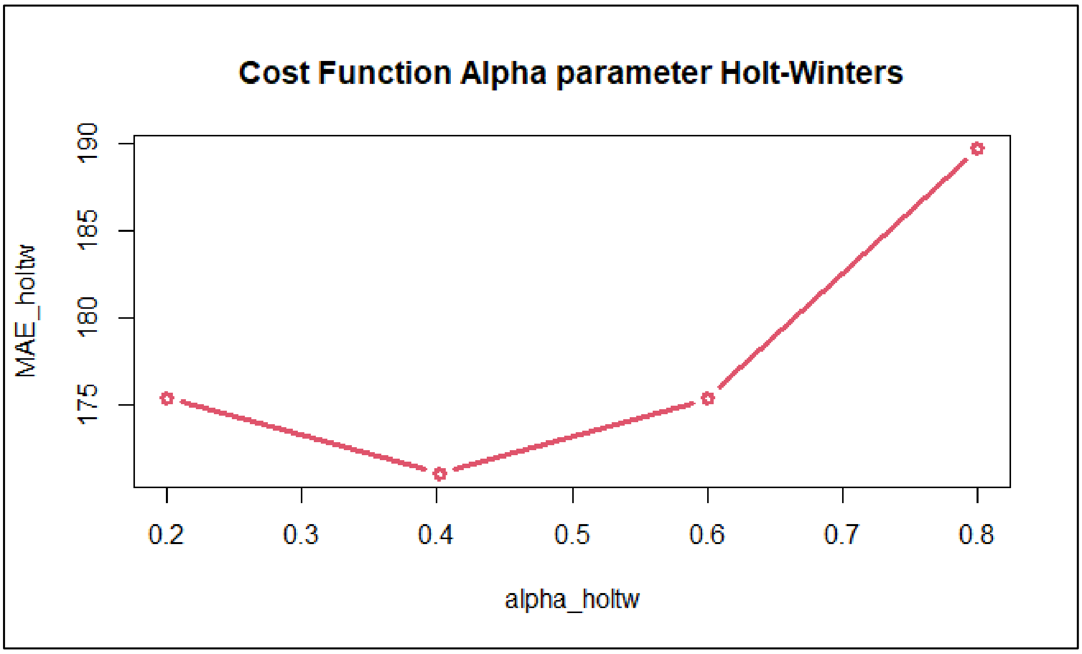

Figure 13" Cost Function Alpha Parameter Holt-Winters"

The optimal value for alpha is 0.402 corresponding to a mean absolute error equal to 171.0885. Alpha close to 0 indicates slow learning meaning past observations have a large influence on forecasts.
The smoothing parameters of the final model are:

Alpha = 0.402
Beta = 0.0327
Gamma = 1e-04

For what concerns the initial states of level, trend and seasonal adjustment terms:

a_0 = 8655.6405

b_0 = 31.3738

c_0 = (1837.03; 209.4959; -361.8488; -312.1387; 140.9807; -104.3604; 346.5446; 268.097; 30.2294; -120.6607; -487.1425; -1446.226)

The observations as well as the forecasted values 24 step ahead (2-year forecasting) will be estimated according to the following procedure:

$$\hat{y}_{k+p|k} := \left(a_k + p\, b_k\right) + c_j, \quad j = k + p - \left\lceil \tfrac{p}{m} \right\rceil \times m \quad (0 \le k \le N;\ p \ge 0)$$

It is possible to appreciate the results graphically:

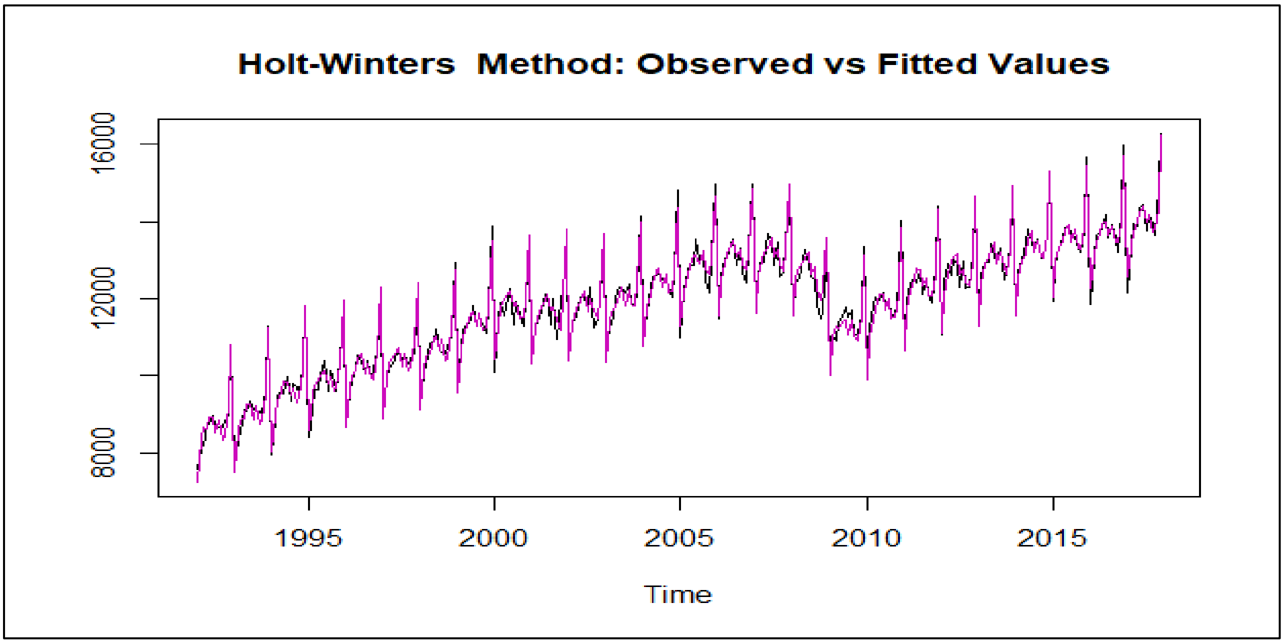

Figure 14 "Holt-Winters Observed vs Fitted Values"

The resulting forecasting using the Holt-Winters Method are the following:

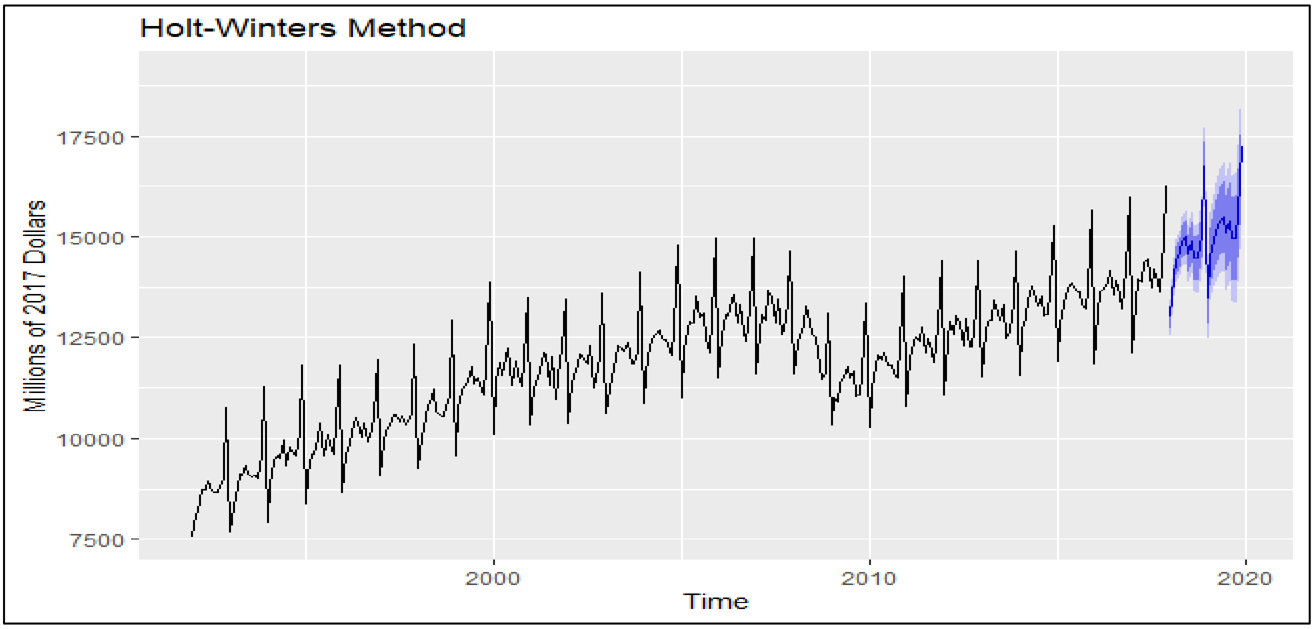

Figure 15 "Holt-Winters Method Forecasting"

Using Holt-Winters Method daily sales in January 2018 are expected to be 12995.13 $ with a 95% c.i. (Lower bound: 12562.91, Upper bound: 13427.36) whereas in December 2019 the values are predicted to be 17246.35 $ (Lower Bound 15486.11, Upper Bound 19006.59).

5. CONCLUSION

The series under analysis, indicating the number of sales per day in the US in 2017 dollars, has been modelled in order to study the deterministic components as well as forecasting the next two years of sales per day. Trend and Seasonality have been estimated by using OLS method. Having done this, it was possible to obtain two different time series a series adjusted for seasonality as well as a series adjusted for trend.

Since the seasonal adjusted time series seemed to evolve on a damped trend, it was modelled using the Gadner-McKenzie method whereas the original data, accounting with both trend and seasonal components have been modelled using the Holt-Winters Method.

Both the final models, obtained by optimizing the smoothing parameters, predict that sales per day will increase robustly over the next two years.