Time series estimation models of tourists in the Netherlands 1990-2022

National and foreign tourism in campsites, motorhome parks and the like

Ágatha del Olmo Tirado

2024-03-30

Introduction

This paper focuses on analysing tourist accommodation data in the Netherlands from 1990 to 2022 obtained from Eurostat, specifically focusing on the total number of tourists, both domestic and foreign, who have stayed in caravan parks, campsites and the like. This study is especially interesting considering that this is a nation characterized by its famous canals and picturesque landscapes, making it an ideal destination to explore by caravan.

After having studied the behavior of the series: a constant growth in tourism from 2019 to the present and a seasonality marked with more tourism in summer; We can focus on finding the 'best' possible model to predict the future using exponential smoothing techniques.

Selecting the best model

After loading the necessary libraries and dating the database, we have cut the series to December 2019 to avoid the effect of Covid on the general study.

```
dataComplete<- read.csv2("./Netherlands_TO_OT.csv",
                         header = TRUE)

dataComplete<- ts(dataComplete[, 1],
                  start = 1990,
                  frequency = 12)

data <- window(dataComplete, end = c(2019, 12))
```

After preparing the data, we can get different prediction models with the ets() function by changing certain indicators.

The ets() function searches for the model that best suits our database according to the default optimization criteria (if we do not indicate anything): maximum likelihood of the MAPE to estimate the parameters and corrected Akaike for the selection of the models having already selected the optimal values of the parameters.

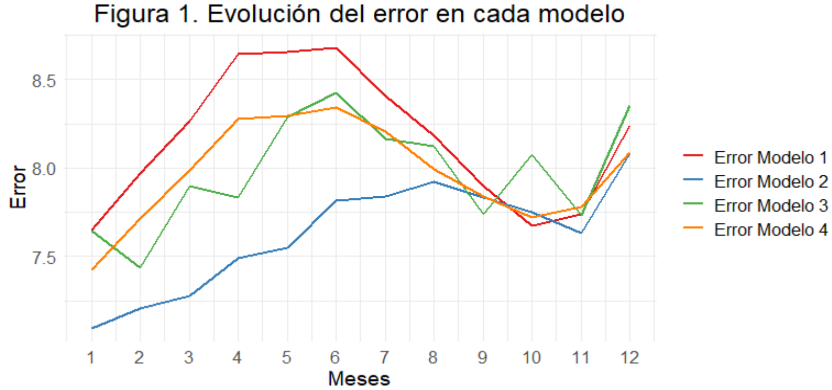
```
mod_1 <- ets(data)


mod_2 <- ets(data,damped
              = FALSE,opt.crit
              = "amse",
             nmse = 12)

mod_3 <- ets(data,damped
              =
             FALSE,lambda = 0)


mod_4 <- ets(data,
             damped = FALSE,lambda =

             0,opt.crit = "amse"
             ,nmse = 12)
```

The first and second models have come out MAM (Holt-Winters Multiplicative straightening), the third AAA (Holt-Winters Additive straightening) and the fourth ANA.

To know which of these models is the best predictor we cannot simply use accuracy(), as it would give us information about the fit, which does not imply a better predictive capacity. Therefore, we use the mobile prediction source method, also known as *cross validation*. This method consists of adjusting and predicting for a part of the sample (adjusts with the training set and evaluates with the test set) that moves forward and repeats the process *k* times.

After doing so, we can easily see which of the models gives us the lowest prediction error with a line graph:

Figura 1. Evolución del error en cada modelo



Clearly, the model that is least wrong when predicting is the second one. We see that the errors in the short and medium term are much lower, and although in the long term it is close to the rest, it remains the best. In addition, it starts with an error of 7.1% and ends with one of 8.1%, it is clear that the further away the time horizon, the worse the prediction will be, and we can expect it to grow at least one percentage point annually.

As this is the model that has shown the least error with the cross-validation method and by a considerable margin, we can imagine that in future predictions the same will happen compared to the rest. Therefore, I choose this model and, from now on, I am only going to study this one. We ask you to start RStudio with a summary of its features.

Study of the chosen model

```
summary(mod_2)

ETS(M,A,M)

Call:
 ets(y = data, damped = FALSE, opt.crit = "amse", nmse = 12)

  Smoothing parameters:
    alpha = 0.1073
    beta = 1e-04
    gamma = 0.2732

  Initial states:
    l = 616105.3406
    b = 1715.0509
    s = 0.4554 0.5321 0.8582 1.0242 2.1079 1.9089
          1.4299 1.5116 0.8451 0.528 0.4572 0.3415

  sigma: 0.0963

     AIC AICc BIC
10187.27 10189.06 10253.33

Training set error measures:
                   ME RMSE MAE MPE MAPE MASE ACF1
Training set -4726.22 101030.1 66817.05 -0.6193191 7.290692 0.788548 -
0.1541556
```

RStudio shows us that the model corresponds to a MAM model (multiplicative Holt-Winters exponential smoothing), i.e., multiplicative error, additive tendency, and multiplicative seasonality.

To begin with, we can analyze the quality of the model. As for bias, we see that there is -0.6%, so there is some bias, and as it is negative we know that it is bias from above, that is, the predictions made tend to be systematically higher than the real values. Regarding the quality of fit, it is wrong by around 100 thousand visitors according to the RMSE, and according to the MAE just over 66 thousand (MAE is always smaller and not directly comparable to the RMSE). We also see it with the MASE in percentage, 7.3%, which represents quite a mistake on the total number of tourists. Regarding the confidence intervals, we cannot trust their calculation, since the ACF1 is above 0.1 in absolute value, so we should use the bootstrapping method. Finally, it is improved, reducing the error by 21%, compared to the naïve method with seasonality (the simplest in models with seasonality).

After seeing the fit quality, it may be interesting to compare the prediction error we had obtained in the cross-validation method with the fit error we have just seen.

```
head(errorAlisado_2,1)

[1] 7.096358
```

We look at the MAPE, which is 7.29, which is somewhat greater than the prediction error, but they are very similar values, which we expected, and that one is greater than the other is of no importance for the study. If we had chosen the prediction error more years ahead, it would move away from the adjustment error, increasing as we have seen in the previous graph.

After this note, we can study the model again. The criterion by which the best model is chosen after estimating the optimal values is the one with the lowest AICc (corrected Acaike), which is a measure used to compare models without the number of parameters (complexity) affecting, because if we used another one such as the RMSE, the one with the most parameters would always be preferable, as if the saying "a més sucre més dolç" followed.

In this case we have a total of 7 parameters to estimate: alpha (smoothing of the level), beta (smoothing of the slope), gamma (smoothing of the seasonal component), l (level), b (slope) and s (seasonality).

The optimal values of the smoothing parameters have been estimated through the maximum likelihood method, and are alpha 0.1, beta 0 and gamma 0.2. All have very low values, which indicates that they change very slowly throughout the series. Specifically, alpha indicates that the level of the series remains almost constant, beta that the slope of the series is constant and gamma that the seasonality does not have many changes.≈≈≈

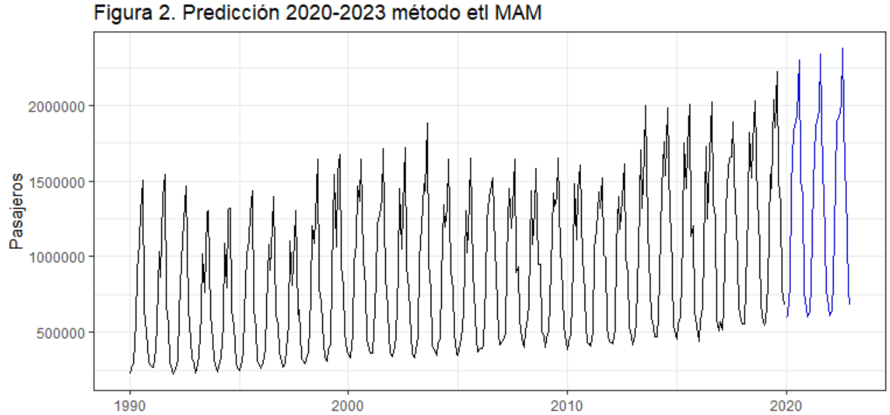
3-year prediction

To make the prediction we use the forecast() function indicating the number of months and the level of confidence we want it to show us.

```
pred_mod2 <- forecast(mod_2,h
                             = 3*12,
                            level = 90)

aggregate(pred_mod2$mean, FUN=sum)

Time Series:
Start = 2020
End = 2022
Frequency = 1
[1] 15740823 15998193 16255683
```

As we can see, annually we predict almost 15.75 million tourists in 2020, 16 million in 2021 and 16.25 million in 2022. We can see graphically that this model takes into account the trend unlike the naïve method with seasonality that we used in the previous practice.

Figura 2. Predicción 2020-2023 método etl MAM

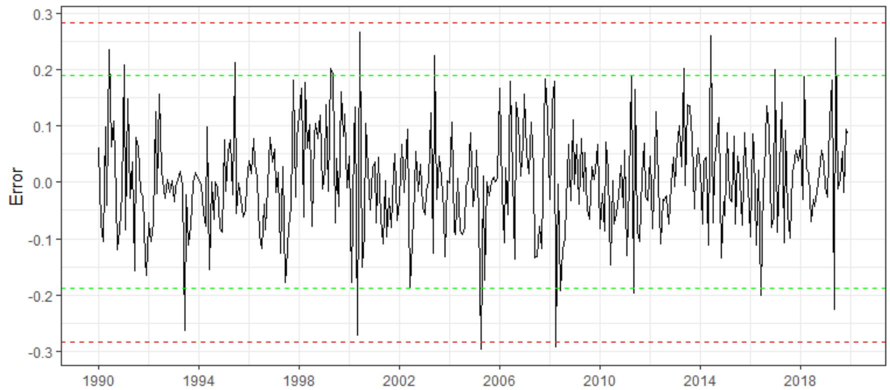


Analysis of the intervention

The intervention and the residual (stochastic component of the series) are studied together, and it is possible to measure whether the cause of an apparently noticeable rise or fall in the series is caused by the randomness of the error or by an intervention. The method I am going to use is that when an error exceeds at least 2.5 times its standard deviation, it is considered an *outlier value*. It should be noted that when it exceeds 3 times, whether or not the reason behind it is found, it is an intervention, on the other hand, if it does not exceed and a reason is not found, we could conclude that it is due to the stochastic nature of the error.

```
 Error <- Residual(mod_2)
SDerror <- SD(Error)
```

Figura 3. Detección de valores outliers en la serie




```
time(error)[!is.na(error) & abs(error) > 2.5 * sderror]

[1] 1993.417 2000.333 2000.417 2005.250 2008.250 2014.417 2019.417
```

As we can see, there are only two errors that exceed 3 times their standard deviation (if the complete series had been used, there would have been a clear intervention caused by Covid), and 5 below 3 but above 2.5 times.

The first year is 1993, in the month of June, regarding this error below, I can only attribute it to economic reasons, according to Mr. Bas B. Bakker, in his 1993 publication, 'II Crisis and recovery', the Netherlands suffered a small recession in the years 1993 and 1994 after what seemed to be an economic recovery (and which continued to be so two years later), so that the population that summer perhaps did not travel as usual.

The second year is 2000, with a big drop in May followed by a big rise in June. This can be given by the UEFA Champions League, whose hosts were Belgium and the Netherlands. Both the quarter-finals and semi-finals took place throughout June, and on July 2 the final took place. This could explain why the population decided not to travel in May to do so in June and to be able to watch the matches that took place in this country.

The third year is 2005, in the month of April, with a drop that exceeds 3 times the standard deviation. This can be due to the fact that, in that year, Holy Week fell in March, from the 20th to the 27th, unlike previous years, which fell in April.

The fourth year is 2008, in the month of April again, and the reason is the same, Easter occurred from March 16 to March 23, so that tourism in April fell compared to the other years, in which normally this week falls in this month

The fifth year is 2014, in the month of May, this time with an increase in tourism, for this mistake I have not been able to find any reason to support it, but it was a very brilliant year in terms of tourism in general.

The sixth and final year is 2019 in June, the only explanation I can give you is related to a tragic event that occurred on the morning of March 18, 2019: the Utrecht attack, where three people died and seven were seriously injured. This attack was considered the worst Islamic terrorist attack the country had experienced so far, and the city's threat level rose to grade 5. With Utrecht being the fourth most important city in the country, it is understandable that the effect that the attack had on visits to the country in the following weeks was notorious, in this way, it could also be that, by not traveling in April out of fear, families traveled in May, when the fear dissipated.

Prediction vs Covid

To obtain the effect of Covid, we subtract the prediction we have made ignoring Covid from what should have happened (the real series).

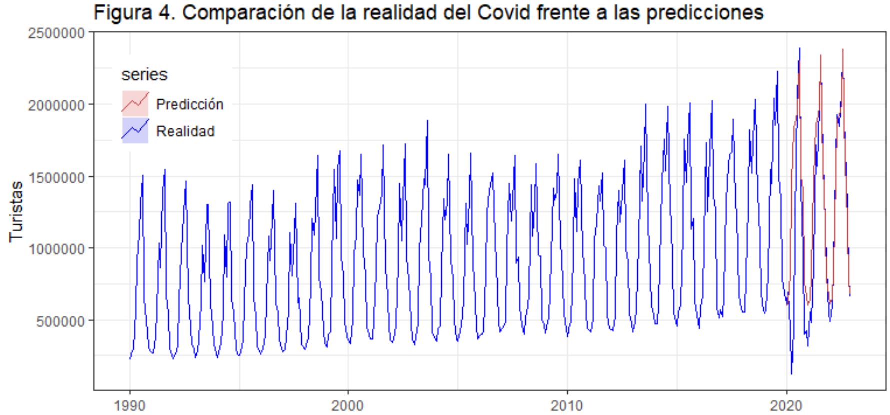
```
aggregate(dataComplete - pred_mod2$mean, FUN = sum)/1000000
```

```
Time Series:
Start = 2020
End = 2022
Frequency = 1
[1] -4.6034167 -1.7381559 0.1940969
```

The results are expressed in millions: in 2020 there were 4.6 million fewer tourists than expected, in 2021 1.7 million less, and in 2022, it recovers with 194 thousand tourists above expectations. Unlike when we use the naïve method, there is no exaggerated recovery in 2022 as it does take into account the effect of the trend of the series on predictions.

To see if the effect of Covid is considerable compared to the rest of the years, we can see it graphed more clearly:

Figura 4. Comparación de la realidad del Covid frente a las predicciones



As we can see, the decline in Covid is not very exaggerated, it does not seem that its effect was very noticeable in the Netherlands. In fact, the year 2020 is quite similar to the figures of the 90s, while in some European countries the pandemic caused record havoc well below historical ones. It should be noted that if we only considered foreign tourists, the difference would probably be greater. We can also see that the actual recovery is quite similar to the prediction as we have seen numerically. This may be due to a rebound effect in which people who did not travel before have done so in 2022, or because the trend from that year onwards is more exaggerated than we expected, which can be attributed to fashions (probably online) of traveling in caravans and the like.

In my opinion and based on this limited information, the effect of Covid-19 in the Netherlands has already worn off. I can conclude this because in 2022 the forecast is already exceeded even taking into account the growing trend of the series, therefore, I imagine that in 2023 it followed figures within what is expected, and even above if these desires or fashions to travel continue.

Conclusion

In summary, this study of tourism in the Netherlands from 1990 to 2022 in campsites, caravans and the like, reveals a growing trend, marked seasonalities that coincide with summer and possible interventions that affected the flow of tourists due to the economic recession, sporting events and attacks that have affected the country over the years. By analyzing data and applying exponential smoothing models, I was able to fairly accurately forecast tourism demand, highlighting the impact of Covid-19 in 2020 and the gradual recovery in the years that followed. This paper underscores the need to understand and address tourism fluctuations for informed decision-making such as improving infrastructure for caravan and camping tourism and implementing marketing policies in order to attract a greater number of visitors to the country.

Webography

Campers, K. (2023, December 4). How Long to Drive my Campervan Per Day. Kambu Campers. https://kambucampers.com/how-long-to-drive-campervan-per-day/

Statistics explained. (n.d.). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Tourism_trips_-_introduction_and_key_figures

Chapters 1 to 6: Tourism Market Trends 1993 - Americas (English version). (n.d.). Default Book Series. https://www.e-unwto.org/doi/abs/10.18111/9789284400829.2

June 1994 Calendar (With Holidays) - Calendarr. (n.d.). Calendarer. https://www.calendarr.com/united-states/calendar-june-1994/

Wikipedia contributors. (2024, April 26). 2014 in the Netherlands. Wikipedia. https://en.wikipedia.org/wiki/2014_in_the_Netherlands