Modelos de estimación en serie temporal de turistas en Países Bajos 1990-2022

Turismo nacional y extranjero en campings, aparcamientos de autocaravanas y similares

Ágatha del Olmo Tirado

2024-03-30

Introducción

Este trabajo se centra en analizar los datos de alojamiento turístico en los Países Bajos desde 1990 hasta 2022 obtenidos del Eurostat, específicamente enfocándose en el total de turistas, tanto nacionales como extranjeros, que se han alojado en aparcamientos de caravanas, campings y similares. Este estudio es especialmente interesante teniendo en cuenta que esta es una nación caracterizada por sus canales famosos y pintorescos paisajes, convirtiéndose en un destino ideal para explorar en caravana.

Tras haber estudiado el comportamiento de la serie: un crecimiento constante del turismo desde 2019 hasta la actualidad y una estacionalidad marcada con más turismo en verano; podemos enfocarnos en encontrar el 'mejor' modelo posible para predecir a futuro usando técnicas de alisado exponencial.

Selección del mejor modelo

Tras cargar las librerías necesarias y fechar la base de datos, hemos cortado la serie hasta diciembre 2019 para evitar el efecto del Covid sobre el estudio general.

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

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

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

Después de preparar los datos, podemos obtener diferentes modelos de predicción con la función ets() cambiando ciertos indicadores.

La función ets() busca el modelo que mejor se adapte a nuestra base de datos según el criterio de optimización por defecto (si no le indicamos nada): máxima verosimilitud del MAPE para estimar los parámetros y Akaike corregido para la selección de los modelos ya habiendo seleccionado los valores óptimos de los parámetros.

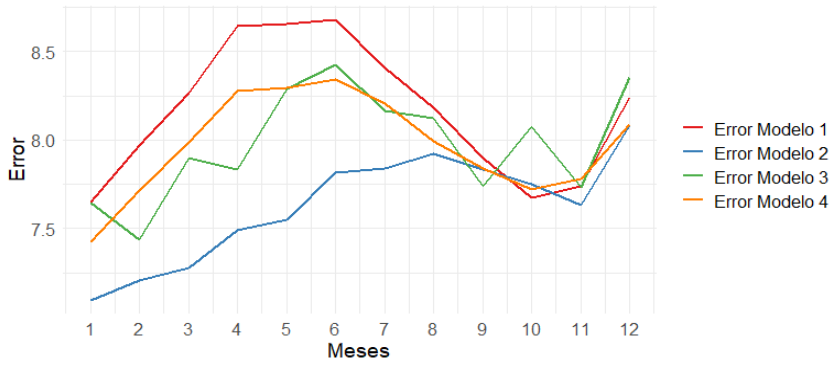
```
mod_1 <- ets(data)


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)
```

El primer y segundo modelo ha salido MAM (alisado de Holt-Winters Multiplicativo), el tercero AAA (alisado de Holt-Winters Aditivo) y el cuarto ANA.

Para saber cuál de estos modelos es el que mejor predice no podemos simplemente usar el accuracy(), pues nos daría información acerca del ajuste, que no implica una mejor capacidad de predicción. Por lo tanto, usamos el método de origen de predicción móvil, también conocido como validación cruzada (*cross validation*). Este método consiste en ajustar y predecir para una parte de la muestra (ajusta con el training set y evalúa con el test set) que se desplaza hacia delante y repite el proceso *k* veces.

Tras realizarlo podemos ver fácilmente cuál de los modelos nos da menor error de predicción con un gráfico de líneas:

Figura 1. Evolución del error en cada modelo



Claramente el modelo que menos se equivoca al predecir es el segundo. Vemos que los errores a corto y medio plazo son mucho menores, y aunque a largo se acerque al resto se mantiene como el mejor. Además, empieza con un error de 7.1% y acaba con uno de 8.1%, está claro que cuanto más lejano esté el horizonte temporal, la predicción será peor, y podemos esperar que mínimo crezca un punto porcentual de forma anual.

Como es este el modelo que menos error ha mostrado con el método de validación cruzada y con bastante diferencia, podemos imaginar que en futuras predicciones ocurrirá lo mismo comparado con el resto. Por lo tanto, escojo este modelo y, a partir de ahora, únicamente voy a estudiar este. Le pedimos para empezar a RStudio un resumen de sus características.

Estudio del modelo escogido

```
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 nos muestra que el modelo se corresponde a un modelo MAM (alisado exponencial de Holt-Winters multiplicativo), es decir, de error multiplicativo, tendencia aditiva y estacionalidad multiplicativa.

Para empezar, podemos analizar la calidad del modelo. En cuanto al sesgo, vemos que hay un -0.6%, de forma que sí que hay algo de sesgo, y al ser negativo sabemos que es sesgo por arriba, es decir, las predicciones realizadas tienden a ser sistemáticamente mayores que los valores reales. Respecto a la calidad de ajuste, se equivoca por alrededor de 100mil visitantes según el RMSE, y según el MAE algo más de 66mil (MAE siempre es más pequeña y no directamente comparable al RMSE). También lo vemos con el MASE en porcentaje, un 7.3%, lo cual representa bastante error sobre el total de turistas. Acerca de los intervalos de confianza, no nos podemos fiar de su cálculo, ya que el ACF1 está por encima del 0.1 en valor absoluto, de forma que deberíamos usar el método bootstrapping. Por último, se mejora, reduciendo en un 21% el error, respecto del método ingenuo con estacionalidad (el más sencillo en los modelos con estacionalidad).

Tras ver la calidad de ajuste puede ser interesante comparar el error de predicción que habíamos obtenido en el método de validación cruzada con el error de ajuste que acabamos de ver.

```
head(errorAlisado_2,1)
```

```
[1] 7.096358
```

Nos fijamos en el MAPE, que es 7.29, que es algo mayor que el error de predicción, pero son valores muy similares, lo cual esperábamos, y que uno sea mayor que el otro no tiene ninguna importancia para el estudio. Si hubiéramos escogido el error de predicción a más años vista se iría alejando del de ajuste, aumentando como hemos visto en el gráfico anterior.

Tras este apunte, podemos volver a estudiar el modelo. El criterio por el cual escoge el mejor modelo tras estimar los valores óptimos es el de menor AICc (Acaike corregido), que es una medida utilizada para comparar modelos sin que el número de parámetros (complejidad) afecte, pues si usáramos otro como el RMSE el que más parámetros tuviera siempre sería preferible, como si siguiera el dicho “a més sucre més dolç”.

En este caso tenemos un total de 7 parámetros a estimar: alpha (suavizado del nivel), beta (suavizado de la pendiente), gamma (suavizado de la componente estacional), l (nivel), b (pendiente) y s (estacionalidad).

Los valores óptimos de los parámetros de suavizado se han estimado a través del método de máxima verosimilitud, y son alpha ≈ 0.1, beta ≈ 0 y gamma ≈ 0.2. Todos tienen unos valores muy bajos, lo cual indica que se modifican a lo largo de la serie muy lentamente. En concreto, alpha indica que el nivel de la serie permanece casi constante, beta que la pendiente de la serie es constante y gamma que la estacionalidad no tiene muchos cambios.

Predicción a 3 años vista

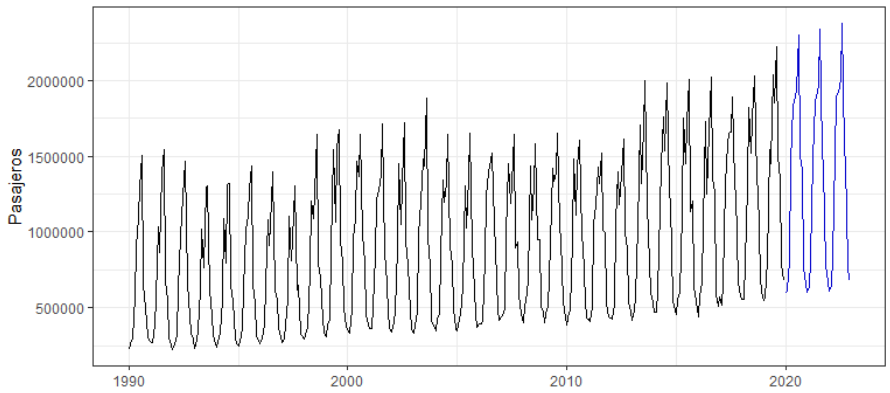
Para realizar la predicción utilizamos la función forecast() indicándole el número de meses y el nivel de confianza que queremos que nos muestre.

```
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
```

Como vemos, anualmente predecimos casi 15.75 millones de turistas en 2020, 16 millones en 2021 y 16.25 millones en 2022. Podemos ver gráficamente que este modelo tiene en cuenta la tendencia a diferencia del método ingenuo con estacionalidad que usamos en la anterior práctica.

Figura 2. Predicción 2020-2023 método etl MAM

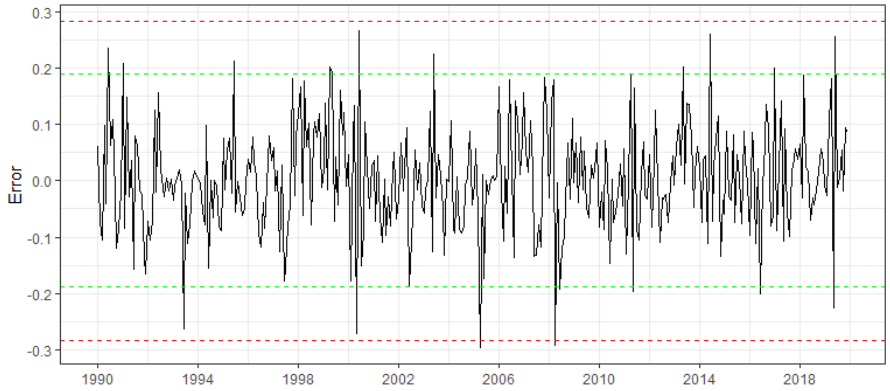


Análisis de la intervención

La intervención y el residuo (componente estocástico de la serie) se estudian de forma conjunta, y se puede medir si la causa de una subida o bajada aparentemente notoria en la serie es causa de la aleatoriedad del error o de una intervención. El método que voy a usar consiste en que cuando un error sobrepasa mínimo 2.5 veces su desviación típica, se considera un valor *outlier*. Cabe destacar que cuando sobrepasa las 3 veces, se encuentre o no la razón detrás, se trata de una intervención, en cambio, si no sobrepasa y no se encuentra una razón podríamos concluir que se debe a la naturaleza estocástica del error.

```
error <- residuals(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
```

Como vemos, solo hay dos errores que sobrepasen 3 veces su desviación típica (de haber usado la serie completa habría una clara intervención provocada por el Covid), y 5 por debajo de 3 pero por encima de 2.5 veces.

El primer año es 1993, en el mes de junio, respecto a este error por debajo, solo puedo atribuirlo a razones económicas, según Mr. Bas B. Bakker, en su publicación de 1993, 'II Crisis and recovery', los Países Bajos sufrieron una pequeña recesión en los años 1993 y 1994 después de lo que parecía una recuperación económica (y que siguió siéndolo dos años más tarde), de forma que la población aquel verano tal vez no viajó como acostumbraba.

El segundo año es el 2000, con una gran bajada en mayo seguida de una gran subida en junio. Esto se puede dar por la Liga de Campeones de la UEFA, cuyos anfitriones fueron Bélgica y Países Bajos. Tanto los cuartos de final como las semifinales se llevaron a cabo a lo largo de junio, y en el 2 de julio tuvo lugar la final. Esto podría explicar que la población decidiera no viajar en mayo para hacerlo en junio y poder ver los partidos que se llevaron a cabo en este país.

El tercer año es 2005, en el mes de abril, con una bajada que supera las 3 veces la desviación típica. Esto se puede dar al hecho de que, en ese año, la semana santa cayó en marzo, del 20 al 27, a diferencia de los años anteriores, que cayó en abril.

El cuarto año es 2008, en el mes de abril de nuevo, y la razón es la misma, la semana santa ocurrió del 16 de marzo al 23 de marzo, de forma que el turismo de abril cayó respecto a los demás años, en los que normalmente esta semana cae en este mes

El quinto año es 2014, en el mes de mayo, esta vez con una subida del turismo, para este error no he podido encontrar ninguna razón que lo respalde, pero sí que fue un año muy brillante en cuanto al turismo en general.

El sexto y último año es 2019 en junio, cuya única explicación que puedo darle está relacionada a un evento trágico que ocurrió la mañana del 18 de marzo de 2019: el

atentado de Utretcht, donde tres personas murieron y siete quedaron gravemente heridas. Este atentado se consideró el peor ataque terrorista islámico que el país había vivido hasta el momento, y el nivel de amenaza de la ciudad subió a grado 5. Siendo Utrecht la cuarta ciudad más importante del país, es comprensible que el efecto que tuvo el atentado sobre las visitas al país las siguientes semanas fuera notorio, de esta forma, también pudo darse que, al no viajar el mes de abril por miedo, las familias viajaron en mayo, cuando el miedo se disipó.

Predicción vs Covid

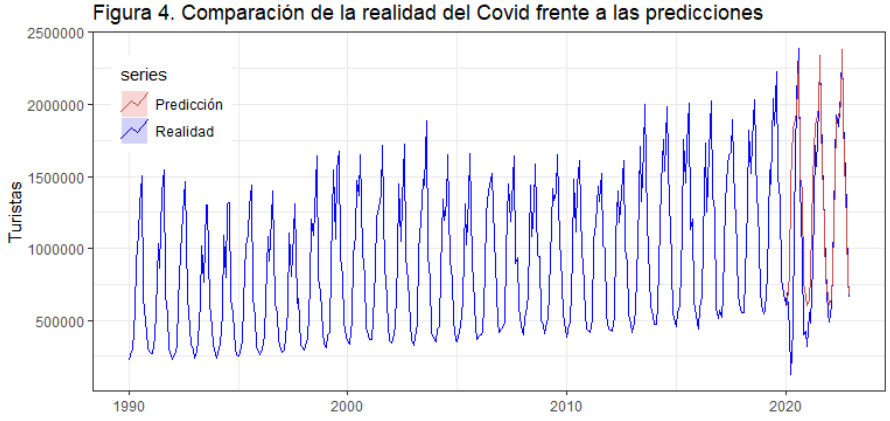
Para obtener el efecto del Covid, restamos la predicción que hemos realizado ignorando el Covid a lo que debiera haber pasado (la serie real).

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

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

Los resultados están expresados en millones: en 2020 hubo 4.6 millones de turistas menos de los que se esperaba, en 2021 1.7 millones menos, y en 2022, se recupera con 194 mil turistas por encima de lo esperado. A diferencia de cuando utilizamos el método ingenuo, no hay una exagerada recuperación en 2022 ya que este sí tiene en cuenta el efecto de la tendencia de la serie en las predicciones.

Para ver si respecto al resto de los años el efecto del Covid es considerable podemos verlo graficado de forma más clara:

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



Como vemos, el descenso del Covid no es muy exagerado, no parece que su efecto fuera muy notorio en los Países Bajos. De hecho, el año 2020 se asemeja bastante a las cifras de los años 90, mientras que en algunos países europeos la pandemia causó estragos record muy por debajo de los históricos. Debe tenerse en cuenta que si solo consideráramos los turistas extranjeros, la diferencia sería probablemente mayor.

También podemos ver que la recuperación real es bastante similar a la predicción tal como hemos visto numéricamente. Esto se puede dar por un efecto rebote en el que la gente que no viajó antes lo ha hecho en 2022, o porque la tendencia a partir de ese año es más exagerada de lo que esperábamos, lo cual se puede atribuir a modas (probablemente online) de viajar en caravanas y similares.

En mi opinión y basándome en esta información limitada, el efecto del Covid-19 en Países Bajos ya ha desaparecido. Esto lo puedo concluir porque en 2022 ya se supera la previsión incluso teniendo en cuenta la tendencia creciente de la serie, por lo tanto, imagino que en 2023 siguió unas cifras dentro de lo esperable, e incluso por encima si siguen estas ganas o moda por viajar.

Conclusión

En resumen, este estudio del turismo en los Países Bajos desde 1990 hasta 2022 en campings, caravanas y similares, revela una tendencia creciente, estacionalidades marcadas que coinciden con el verano y posibles intervenciones que afectaron al flujo de turistas por la recesión económica, eventos deportivos y atentados que han afectado al país a lo largo de los años. Mediante el análisis de datos y la aplicación de modelos de alisado exponencial, logré prever con bastante precisión la demanda de turismo, destacando el impacto del Covid-19 en 2020 y la recuperación gradual en los años posteriores. Este trabajo subraya la necesidad de comprender y abordar las fluctuaciones del turismo para la toma de decisiones informadas como la mejora de infraestructuras destinadas al turismo en caravanas y campings, y la implementación de políticas de marketing con el fin de atraer un mayor número de visitantes al país.

Webgrafía

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.). Calendarr. 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