

**Caso de uso experimental:
Predicciones de viajeros y
gasto a partir de datos de
FRONTUR, ETR y EGATUR**

Metodología y aplicación

Fecha de publicación: enero 2025

Contenido

I. Introducción.....	2
II. Metodología	3
a. Series temporales: fuente de origen.....	3
b. Modelos de predicción.	4
c. Proceso train-test.....	4
d. Hipótesis aceptadas en el sistema de predicciones.....	6
III. Aplicación.....	7
a. Obtención de predicciones.....	7
b. Actualización periódica de predicciones.	8
IV. Conclusiones	8
Notas metodológicas	9

Lista de ilustraciones

Ilustración 1 Esquema series sobre las que realizar predicciones	2
Ilustración 2 Proceso train-test.....	5
Ilustración 3 Esquema proceso obtención predicciones para cada serie.....	7
Ilustración 4 Gráfico evolución observaciones y predicciones	7
Ilustración 5 Esquema proceso actualización de predicciones	8

Lista de tablas

Tabla 1 Resumen de series incluidas en el programa	3
--	---

I. Introducción

El propósito de este documento es explicar la metodología elaborada por **SEGITTUR** y su aplicación para la generación de predicciones con series temporales de viajeros y de gasto de **FRONTUR**, **EGATUR** y **ETR**.

En un inicio, el objetivo planteado fue establecer un *programa de código en R Software libre* que permitiera conocer una predicción de cómo continuarían las series nacionales de turistas y gasto.

Una vez comenzado el desarrollo se decide ampliar el planteamiento. Utilizando la programación se estudiarán, tanto series de *visitantes*, *turistas* y *excursionistas* como series de principales orígenes y destinos. En concreto, además de predecir sobre la serie nacional, también se realizará para cinco países de origen y cinco Comunidades Autónomas de destino considerados los principales en términos de *turistas*. A continuación, se sintetiza dicho planteamiento.

Ilustración 1 Esquema series sobre las que realizar predicciones



Fuente: Elaboración propia.

Por lo tanto, la personalización de la predicción en cada serie es una premisa que forma parte del proyecto. Para ello, el *programa de código* elaborado analiza los resultados de una lista de modelos de predicción preestablecidos con cada serie y con ellos, determina el mejor modelo de predicción para cada serie. Tanto los modelos como la manera de establecer el mejor modelo se explican en el apartado II. Metodología.

En la siguiente tabla se puede conocer la naturaleza de las diferentes series sobre las que se han elaborado predicciones a través del *programa* elaborado. En total, se trata de 45 series sobre [FRONTUR](#), [EGATUR](#) y [ETR](#). Además, de otras 22 para *excursionista* que se obtienen de otro modo.

Tabla 1 Resumen de series incluidas en el programa

	Internacionales			Nacionales		Total series
	Visitantes	Turistas	Gasto	Viajes	Gasto	
Serie Total	1	1	1	1	1	5
Serie por Mercado de origen	5	5	5	0	0	15
Serie por Ciudad de destino	5	5	5	5	5	25
Número total series	11	11	11	6	6	45

Fuente: Elaboración propia.

II. Metodología

La base para establecer una metodología en este caso ha sido el análisis de cómo funciona la predicción con las series seleccionadas. Se pretende comprender la casuística sobre la que se va a trabajar analizando las series temporales. Se trata de intentar responder cuestiones cómo qué componentes tienen las series temporales, cómo influye la estacionalidad, estacionariedad o la varianza en ellas para así poder generar predicciones de calidad aceptable.

a. Series temporales: fuente de origen.

En primer lugar, se han seleccionado series relacionadas con viajes y gasto turístico del Instituto Nacional de Estadística ([INE](#)), en concreto:

- [Estadística de movimientos turísticos en frontera \(FRONTUR\)](#): estadística de publicación mensual que proporciona estimaciones mensuales y anuales del número de visitantes no residentes en España que llegan a nuestro país.
- [Encuesta de gasto turísticos \(EGATUR\)](#): tiene como objetivo principal conocer el gasto turístico de los visitantes extranjeros a su salida de España.

- Encuesta de turismo de residentes (ETR): encuesta continua cuyo objetivo principal es proporcionar estimaciones mensuales, trimestrales y anuales de los viajes realizados por la población residente en España.

Estas series se han extraído a través de las consultas a la [API de Dataestur](#). El estudio para establecer las bases de la metodología se ha centrado en la serie nacional de visitantes. Sobre ésta se han generado transformaciones que van desde la exclusión de determinados años a transformaciones logarítmicas y otras.

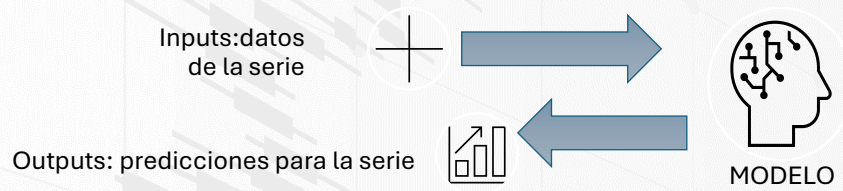
b. Modelos de predicción.

El estudio de series va ligado al establecimiento de modelos para la predicción de valores. Se analizan las diferentes combinaciones de modelos y series que van desde el modelo *ARIMA*, habitual en las series temporales, a otros como la regresión lineal. El modelo *ARIMA* será en el que se encuentre un buen balance entre calidad de resultados y facilidad de ejecución para las series evaluadas. Respecto a *ARIMA*, cuenta con tres componentes: *autorregresión (AR)*, *diferenciación (I)* y *medias móviles (MA)*. La autorregresión captura la relación entre una observación y una serie de observaciones rezagadas. La diferenciación hace que la serie temporal sea estacionaria. Y las medias móviles modelan la relación entre una observación y un error residual a partir de un modelo de media móvil aplicado a observaciones rezagadas.

c. Proceso train-test

El método de evaluación de las combinaciones de series y modelos se establece con un proceso de *train-test split*. Éste consta de unos inputs y outputs. Los inputs son datos de la serie que se enseñan al modelo para que se entrene con ellos. Los outputs son las estimaciones que el modelo realiza para otros períodos de los que tenemos el dato, pero no se lo mostramos al modelo.

Ilustración 2 Proceso train-test



Fuente: Elaboración propia.

Para este estudio, se han tomado dos enfoques de *train-test* en función de la fase de evaluación. En una primera evaluación, se determinan como inputs o datos de entrenamiento el 80% de la serie. Finalmente, se definen como datos de entrenamiento datos del año actual, en el momento de realización de este trabajo, era 2024. Se ha de contrastar el dato que propone el modelo con el dato que tiene la serie que se ha ocultado al modelo para conocer la calidad con la que cada modelo predice. Así, se mide la desviación, se han empleado indicadores como:

- *Error Cuadrático Medio (RMSE)*: diferencia media cuadrática entre los valores reales y los valores predichos. En base a este indicador se ha seleccionado el que consideramos “mejor modelo”. Se calcula con la fórmula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

donde:

- Y_i es el valor real.
- $\hat{Y}_i$ es el valor predicho.
- n es el número de observaciones.
- *Error porcentual absoluto medio (MAPE)*: expresa la precisión como un porcentaje y se calcula con la fórmula:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \times 100$$

donde:

- A_t es el valor real.

- F_t es el valor pronosticado.
- n es el número de observaciones.

d. Hipótesis aceptadas en el sistema de predicciones.

Una vez evaluadas las combinaciones de series y modelos dentro de la serie nacional, se han extraído una serie de conclusiones que han determinado la configuración del sistema de predicciones:

- Exclusión de la serie de las observaciones que van del año 2020 a 2022.
Debido al efecto la pandemia, sin estos años mejoran los indicadores.
- Exclusión de la serie de las observaciones del año 2015.
Las series comienzan en octubre 2015, los indicadores varían ligeramente al alza o a la baja con estas observaciones.
- Empleo de series no transformadas y predicción con modelos *ARIMA*.
Tras pruebas con modelos con transformaciones logaritmo, box-cox y otros como la regresión lineal, se encuentra más ágil y con resultados aceptables la selección de modelos *ARIMA* sin transformaciones. Se ha acotado a los siguientes:
 - *AUTOARIMA*: busca automáticamente los mejores valores para los parámetros p , d y q , basándose en criterios estadísticos como el AIC (Criterio de Información de Akaike).
 - *ARIMA con grid search*: supone la búsqueda sistemática de los mejores hiperparámetros para el modelo mediante un proceso de prueba y error.
 - *ARIMA con early stopping*: utiliza grid search para encontrar los mejores parámetros reduciendo el tiempo de cálculo al detenerse la búsqueda si se encuentra un modelo con un AIC significativamente mejor que los previos.
 - *ARIMA manual*: se especifican los parámetros, en este caso, según el tipo y nombre de la serie temporal se han establecido dos casos:
 - Sin componente autoregresivo ni de media móvil en la parte no estacional (0,1,0) (1,1,1).
 - Con componentes autoregresivo en la parte no estacional y la estacional y de media móvil (1,1,1) (1,1,1).

- Serie de excursionistas calculada sobre la serie visitantes y la serie turistas.

La precisión en *excursionistas* es aceptable en la *serie total*, si bien en series de algunos mercados se reduce considerablemente al ser *excursionista* poco representativo para ese mercado. Como la precisión en *visitantes* y *turistas* es aceptable, *excursionistas* se estimará en base a éstas, del siguiente modo:

$$\text{Predicción excursionistas} = \text{Predicción Visitantes} - \text{Predicción Turistas}$$

III. Aplicación

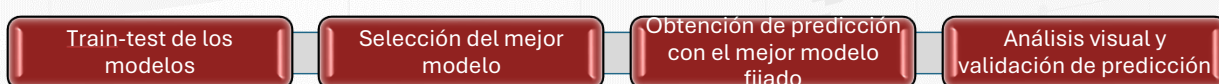
a. Obtención de predicciones.

Una vez definido en la *Metodología* el marco sobre el que trabajar, se inicia el proceso de generación de predicciones.

Para ello, es necesario capturar el mejor modelo para esa serie y entrenarlo. Se ha definido como mejor modelo para cada serie, aquel cuyo test ha obtenido el menor *RMSE*. Además, se ha analizado que se sitúe dentro de unos valores considerados aceptables.

Este mejor modelo se ha empleado para generar predicciones. Además, de los indicadores se ha elaborado una gráfica de control con la evolución mensual en los últimos años.

Ilustración 3 Esquema proceso obtención predicciones para cada serie



Fuente: Elaboración propia.

Ilustración 4 Gráfico evolución observaciones y predicciones



Fuente: Elaboración propia.

b. Actualización periódica de predicciones.

Pasado un tiempo, la serie original del INE se actualizará y contará con nuevas observaciones. El proceso de evaluación con el que hemos determinado el mejor modelo se volverá a ejecutar para obtener el mejor modelo con los nuevos inputs que el modelo tendrá en cuenta.

Se conocerá si el mejor modelo sigue siendo el que había quedado establecido anteriormente y se podrá valorar si se debe modificar como propone la actualización para obtener las nuevas predicciones.

Ilustración 5 Esquema proceso actualización de predicciones



Fuente: Elaboración propia.

IV. Conclusiones

Para concluir, con este caso, se ha podido construir un sistema de predicciones de series de alta importancia en el sector turístico como son de [FRONTUR](#), [EGATUR](#) y [ETR](#).

Se trata de estimaciones obtenidas mediante modelos ARIMA, uno de los enfoques más habituales para las series temporales.

Se ha procurado obtener una calidad aceptable en los resultados, para ello, se han establecido mecanismos de control. Entre otros, se dispone de indicadores como el *error cuadrático medio* o el gráfico de evolución de las estimaciones.

Se ha tratado de automatizar gran parte del proceso, los datos se actualizan desde la [API](#) de [Dataestur](#) y obtención del mejor modelo y las predicciones se obtiene de la ejecución del código elaborado. Con estas automatizaciones se consigue la obtención en un tiempo reducido las predicciones de más de cuarenta series diferentes.

Notas metodológicas

Este documento se crea con el propósito de compartir una metodología para la elaboración de predicciones turísticas. Es posible su elaboración desde otro enfoque o mediante otras técnicas.

En el tratamiento de datos y la construcción del índice se ha empleado el *programa R de Software libre*, siendo posible la utilización de otros.

Para consultar o sugerencias específicas pueden contactar a través de info.sit@segittur.es.

Fecha de publicación: enero 2025

DATAESTUR

