DATAESTUR

Experimental use case: Traveller and spending forecasts based on data from FRONTUR, ETR and EGATUR

Methodology and application

Publication date: January 2025

Content

List of illustrations

List of tables

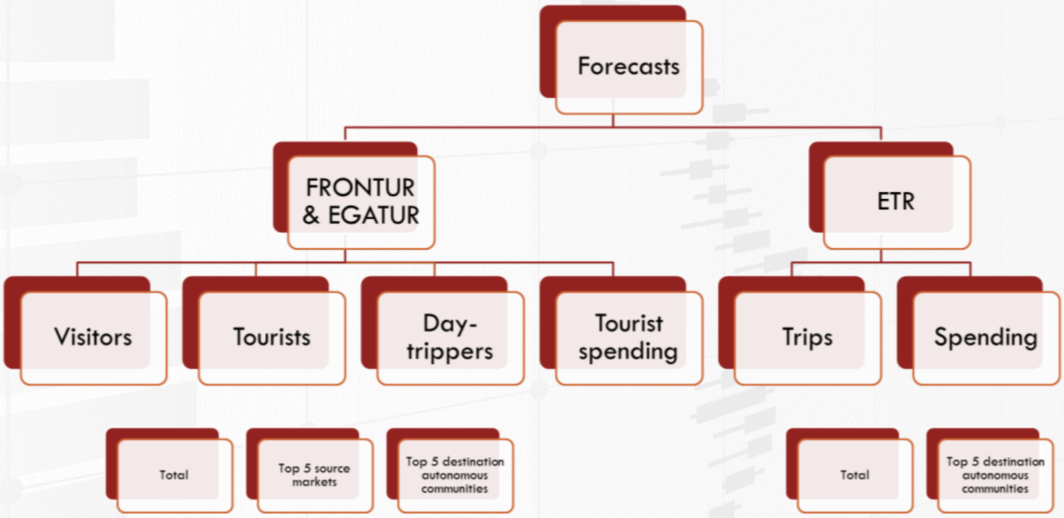
I. Introduction

The purpose of this document is to explain the methodology developed by SEGITTUR and its application for generating series of traveller and spending forecasts via FRONTUR, EGATUR and ETR.

At the outset, the goal was to establish an *open-source R software program* that could provide a forecast of how national tourist and expenditure trends would continue.

Once development began, it was decided to expand the approach. Using programming, both *visitor, tourist, and day-tripper* data, as well as data on key source and destinations, will be analysed. Specifically, in addition to predicting data for Spain, this will also be done for five source countries and five destination autonomous communities considered to be the main ones in terms of *tourists*. This approach is summarised below.

Illustration 1 Diagram of the data used to create the predictions



Source: by the authors.

Therefore, personalising the prediction for each data series is a premise that is part of the project. To achieve this, the *code program* used analyses the results of a pre-established list of prediction models for each series and, based on these, determines the best prediction model for each series of data. Both the models and how to establish the best model are explained in section II. Methodology.

The following table shows the nature of the different series for which predictions have been made using the *program* developed. In total, there are 45 series on FRONTUR, EGATUR and ETR data. There are also another 22 for *day-trippers* that are obtained in another way.

Table 1 Summary of series included in the program

	Internationals			**Nationals**		**Total series**
	Visitors	Tourists	Spenditure	Trips	Spenditure	
Total Series	1	1	1	1	1	5
Series by source market	5	5	5	0	0	15
Series by destination city	5	5	5	5	5	25
Total number of series	11	11	11	6	6	45

Source: by the authors.

II. Methodology

The basis for establishing a methodology in this case has been the analysis of how predictions function with the series selected. The aim is to understand the specific cases to be worked on by analysing the time series. It is about trying to answer questions such as what components time series have, how seasonality, stationarity or variance influence them in order to generate predictions of acceptable quality.

a. Time series: source of origin.

Firstly, series related to travel and tourist spending from the National Institute of Statistics (INE), specifically:

- Statistics on tourist movements at the border (FRONTUR): Monthly publication statistics that provide monthly and annual estimates of the number of non-resident visitors to Spain.
- Tourism spending survey (EGATUR): its main objective is to understand tourism spending among foreign visitors when they leave Spain.
- Resident tourism survey (ETR): ongoing survey whose main objective is to provide monthly, quarterly and annual estimates of the trips taken by the resident population in Spain.

These series have been extracted through queries to the Dataestur API. The study to establish the bases of the methodology has focused on the national series of visitors. Transformations have been generated on this, ranging from the exclusion of certain years to logarithmic transformations, and others.

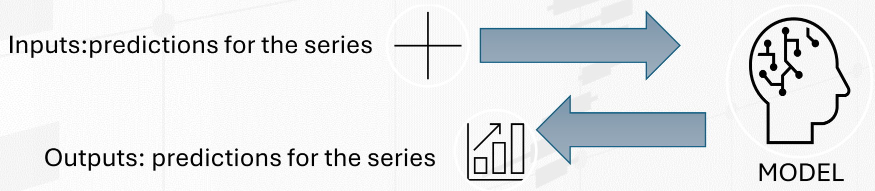
b. Prediction models.

The study of series is linked to the development of models for predicting values. The different combinations of models and series are analysed, ranging from the *ARIMA* model, which is common in time series, to others such as linear regression. The *ARIMA* model will be the one that finds a good balance between quality of results and ease of execution for the series evaluated. *ARIMA* has three components: *autoregression (AR), differencing (I)* and *moving averages (MA)*. Autoregression captures the relationship between an observation and a series of lagged observations. Differencing makes the time series stationary. Moving averages model the relationship between an observation and a residual error from a moving average model applied to lagged observations.

c. Train-test process

The evaluation method for series and model combinations is established with a train-test split process. It consists of inputs and outputs. Inputs are data from the series that are fed into the model for training. Outputs are the estimates that the model makes for other periods for which we have data, but we do not provide it to the model.

Illustration 2 Train-test process


Inputs:predictions for the series
Outputs: predictions for the series
MODEL

Source: by the authors.

For this study, two train-test approaches were taken depending on the evaluation phase. In a first evaluation, 80% of the series is determined as inputs or training data. Finally, training data is defined as data from the current year at the time of this work, which was 2024. The data proposed by the model must be compared with the data in the series that has been withheld from the model to determine the accuracy with which each model predicts. In this way, the deviation is measured, and indicators such as the following have been used:

- ○ *Root Mean Squared Error (RMSE):* mean squared difference between actual values and predicted values. Based on this indicator, we have selected what we consider to be the "best model". It is calculated with the formula:

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

where:

- Y_i is the real value.
- $\hat{Y}_i$ is the predicted value.
- n is the number of observations.

- ○ *Mean absolute percentage error (MAPE):* expresses accuracy as a percentage and is calculated using the formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_{t-\,F_t}}{A_t}\right| x\;100$$

where:

- A_t is the real value.
- F_t is the predicted value.
- n is the number of observations.

d. *Hypotheses accepted in the prediction system.*

Once the combinations of series and models within the national series were evaluated, a number of conclusions were drawn that have determined the configuration of the prediction system:

- <u>Exclusion of observations ranging from 2020 to 2022 from the series.</u>
  Due to the effects of the pandemic, the indicators improve without these years.
- <u>Exclusion of observations from the year 2015 from the series.</u>

The series begins in October 2015; the indicators vary slightly up or down with these observations.

- Using non-transformed series and prediction with *ARIMA models*.

  After testing models with logarithmic, Box-Cox and other transformations such as linear regression, the selection of ARIMA models without transformations was found to be more agile and yielded acceptable results. They have been narrowed down to the following:

  - *AUTOARIMA*: Automatically searches for the best values for the parameters p, d and q, based on statistical criteria such as the AIC (Akaike Information Criterion)
  - *ARIMA with grid search*: involves the systematic search for the best parameters for the model through a trial and error process.
  - *ARIMA with early stopping:* uses grid search to find the best parameters, reducing the computation time by stopping the search if a model with an AIC significantly better than the previous ones is found.
  - *ARIMA manual:* parameters are specified; in this case, according to the type and name of the time series, two cases have been set up:
    - No autoregressive component or moving average in the non-seasonal part (0,1,0) (1,1,1).
    - With autoregressive components in both the non-seasonal and seasonal parts, and moving average (1,1,1) (1,1,1).

- *Data set of day-trippers calculated on the basis of the data set of visitors and the data set of tourists.*

  The accuracy in *day-trippers* is acceptable in the *total data set*, although in data sets from some markets it is considerably reduced because *day-trippers* are not very representative for that market. Since the accuracy in *visitors* and *tourists* is acceptable, *day-trippers* will be estimated based on these, as follows:

$$\text{Prediction for day-trippers} = Prediction\ Visitors\ - Prediction\ Tourists$$

III. Application

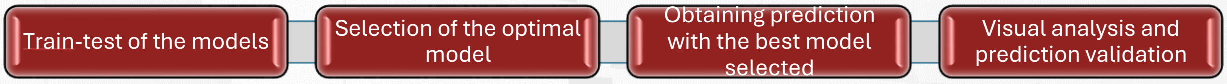
a. *Obtaining predictions.*

Once the working framework has been defined in the *Methodology*, the process of generating predictions begins.

To do this, it is necessary to capture the best model for that series and train it. The best model for each series has been defined as the one whose test has obtained the lowest *RMSE*. In addition, it has been analysed to ensure it falls within the acceptable range of values.

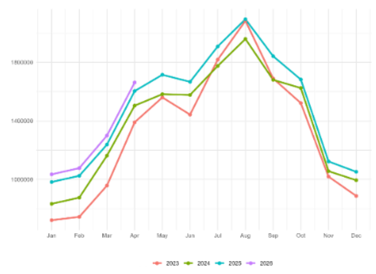
This optimal model has been used to generate predictions. In addition, a control chart has been drawn up for the indicators showing the monthly evolution in recent years.

Illustration 3 Diagram of the process for obtaining predictions for each series


Train-test of the models
Selection of the optimal model
Obtaining prediction with the best model selected
Visual analysis and prediction validation

Source: by the authors.

Illustration 4 Chart of progression of observations and predictions



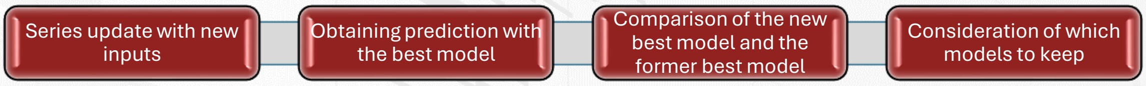
Source: by the authors.

b. *Periodic update of predictions.*

After some time, the original INE series will be updated to include new observations. The evaluation process with which we have determined the best model will be executed again to obtain the best model with the new inputs that the model will take into account.

It will be determined whether the best model is still the one previously established, and it will be possible to assess whether it should be modified as proposed by the update in order to obtain new predictions.

Illustration 5 Diagram of the prediction update process


Series update with new inputs
Obtaining prediction with the best model
Comparison of the new best model and the former best model
Consideration of which models to keep

Source: by the authors.

IV. Conclusions

In conclusion, through this case study, it has been possible to build a system of predictions of series of high importance in the tourism sector such as FRONTUR, EGATUR and ETR.

These are estimates obtained using ARIMA models, one of the most common approaches for time series.

Efforts have been made to achieve an acceptable quality of results, and control mechanisms have been established for this purpose. Among others, indicators such as the *mean square error* or the graph of the trend in estimates are available.

An attempt has been made to automate a large part of the process; the data is updated from the *API* from *Dataestur* and the best model and predictions are obtained from the execution of the developed code. With these automations, it is possible to obtain predictions for more than forty different series in a short time.

NOTES

Methodological notes

This document aims to share a methodology for the creation of tourism forecasts. It is possible to develop it using a different approach or other techniques.

The *open-source R program* was used for data processing and index construction, although other programs can be used.

For specific questions or suggestions, please contact us via info.sit@segittur.es.

Publication date: January 2025

DATAESTUR