

# Tourism in Spain

Nicolás Jiménez Muñoz and Aina Vila Arbusà

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Basic description of the dataset</b>	<b>3</b>
<b>3</b>	<b>Summary of the methodology used for data analysis</b>	<b>4</b>
<b>4</b>	<b>Results and interpretation</b>	<b>5</b>
4.1	Identification . . . . .	5
4.1.1	Transformations for stationarity . . . . .	5
4.1.2	ACF and PACF analysis . . . . .	7
4.2	Estimation . . . . .	9
4.3	Validation . . . . .	10
4.3.1	Residual analysis . . . . .	10
4.3.2	Model expressions . . . . .	13
4.3.3	Forecasting accuracy . . . . .	15
4.4	Forecasting . . . . .	16
4.5	Calendar Effects . . . . .	16
4.6	Outlier Treatment . . . . .	17
4.6.1	Automatic outlier detection . . . . .	17
4.6.2	Forecasts for the original series . . . . .	20
<b>5</b>	<b>Conclusions and discussion</b>	<b>22</b>
<b>6</b>	<b>References</b>	<b>23</b>
<b>7</b>	<b>Annex</b>	<b>23</b>

# 1 Introduction

Spain is ranked as one of the world's top tourist destinations, attracting millions of visitors every year due to its diverse landscapes, rich gastronomy and cultural heritage. Tourism is a key economic sector in Spain, significantly contributing to the country's employment and. Understanding the fluctuations in tourist arrivals and anticipating movements is very important for effective planning and decision-making in both public and private sectors.

In this project, we will apply ARIMA models and the acquired knowledge of time series analysis in order to generate meaningful and accurate forecasts for future tourists arrivals. The main purpose is develop a reliable forecasting model based on the available data of the selected time series: the number of tourist arrivals in Spain (in millions).

By following Box-Jenkins methodology, which consists of a structured process of description and model identification, estimation, diagnosis and residual analysis, validation, final model selection and forecasting, the project aims to provide both a comprehensive analysis of the historical data and projections for future periods. This study can be crucial for authorities and businesses to prepare for future demand and optimizing their strategies accordingly.

## 2 Basic description of the dataset

The selected dataset contains monthly observations of the total number of tourist arrivals in Spain, expressed in millions of visitors. It spans from January 2000 to November 2018, comprising 227 observations.

The variable of interest is the number of tourist arrivals per month. This provides valuable insights into seasonal patterns, trends and fluctuations in the tourism sector, which are crucial for understanding Spain's tourism industry.

The data is sourced from the Spanish Ministry of Industry, Commerce and Tourism, specifically from the FRONTUR survey. The data collection process, in this case, involves surveys and administrative records at border entry points, ensuring comprehensive and accurate tracking of international tourist movements.

### 3 Summary of the methodology used for data analysis

In this analysis, the Box-Jenkins methodology is used to perform time series forecasting employing ARIMA models, with additional extensions for detecting and adjusting outliers and incorporating calendar effects.

ARIMA models (Autoregressive Integrated Moving Average) are statistical tools that are used for analyzing and forecasting time series data. These models take into account past values and prediction errors to make predictions of future periods.

They consist of three key components:

- **AR (autoregressive)**: captures the linear relationship between the current value of the series and its past values.
- **MA (moving average)**: models the relationship between the current value and past prediction errors.
- **I (integrated)**: used to remove trends and achieve stationarity in the series.

The **Box-Jenkins** methodology involves several steps:

1. **Identification**: to apply the ARIMA models, the time series must be stationary, meaning its statistical properties remain constant over time. First, a visual analysis is done to determine the appropriate transformations to achieve stationarity. Following this, a correlation analysis using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) is carried out to identify potential ARIMA models for the series.
2. **Estimation**: it involves estimating the parameters for the identified models. Then, a significant analysis of the parameters is conducted, where insignificant parameters are detected and, consequently, removed from the model.
3. **Validation**: it includes several aspects. One of them is the residual analysis to ensure they resemble white noise, meaning they are independent, have zero mean and constant variance. Additionally, model properties, invertibility and stability are assessed. Furthermore, so as to assess how well our model explains the data, metrics such as AIC, BIC, RMSE, MAE, RMSP and MAPE are used. Finally, the last 12 observations will be reserved to evaluate the model's forecasting accuracy. By comparing each model's performance, we will select the best one for real forecasting. In case the models are bad, the procedure has to be done again. Otherwise, the following step can be performed.
4. **Forecasting**: the chosen model is used to make predictions for the next 12 months beyond the last recorded observation.
5. **Calendar Effects**: to improve forecast accuracy, calendar-related factors such as public holidays or the number of weekends in a month are incorporated as external regressors. These effects are modeled, tested for significance, and used to adjust the series accordingly.
6. **Outlier Treatment**: after calendar adjustment, the outdetec function is applied to detect outliers in the data, which can be classified in three types:
  - AO (additive outlier): effects a single observation.
  - TC (temporary change): has a transitory effect that disappears over time.
  - LS (level shift): causes a permanent change in the level of the series.

By carefully analyzing these outliers, their impact on the model is assessed. The time series is then adjusted accordingly, and the linearized time series is obtained. The forecasting process is repeated using the transformed model, and the results of the new forecasts are compared with those from the initial one to evaluate improvements.

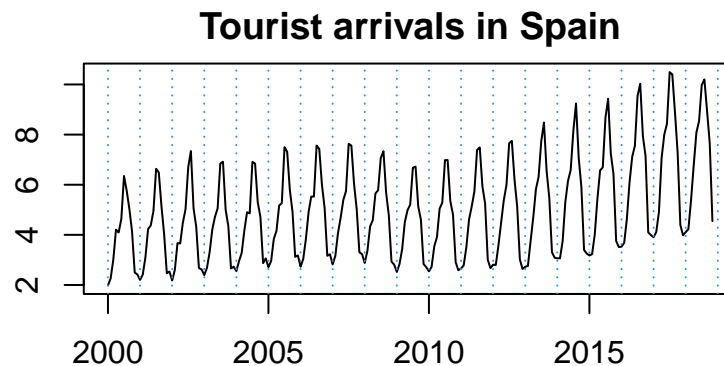
## 4 Results and interpretation

### 4.1 Identification

#### 4.1.1 Transformations for stationarity

Plotting the time of tourist arrivals in Spain makes easier to detect an increasing trend, particularly evident since 2013. This upward trajectory reflects the growth in tourism over the years.

Additionally, a clear seasonal pattern is observed, with peaks each year that can be attributed to the increase in tourism during summer holiday period. This seasonal variation highlights the cycle nature of tourism, with higher numbers of arrivals during warmer months.



In time series analysis, as it has been explained, it is essential for the series to be stationary, meaning its statistical properties do not change over time.

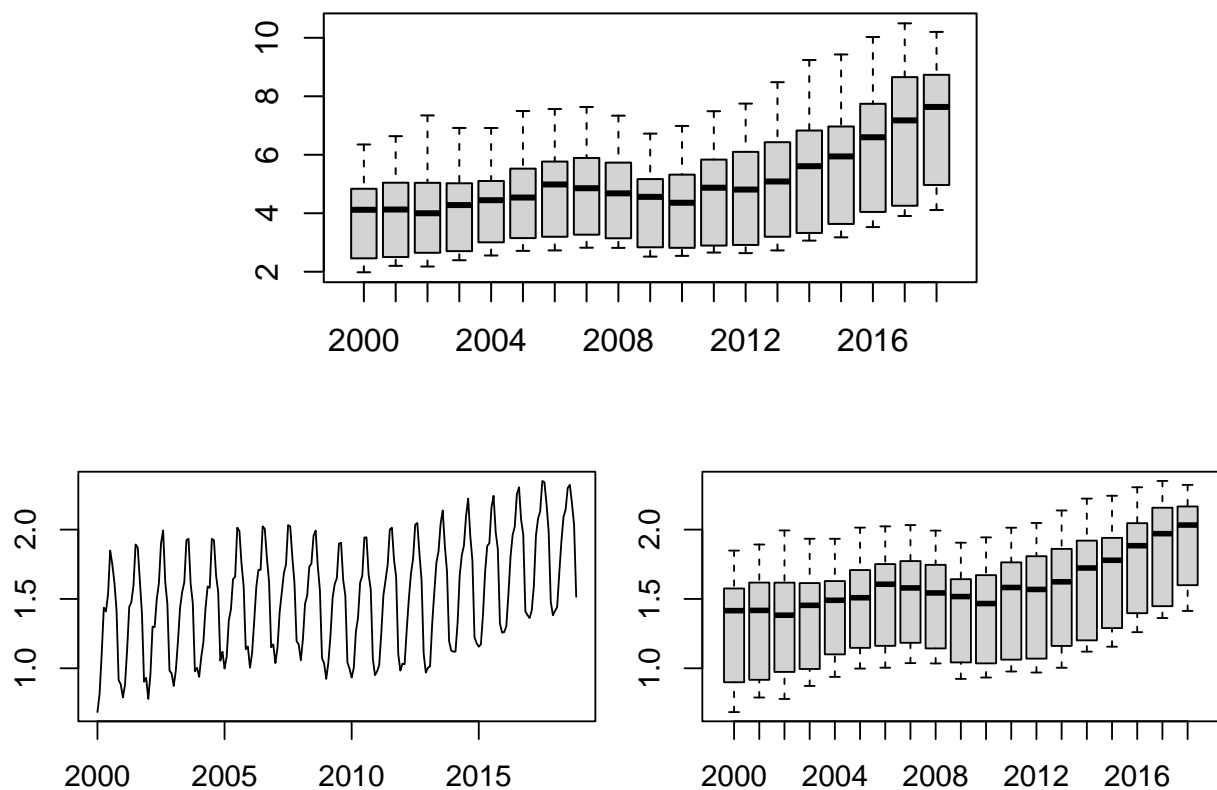
The first step in preparing a time series for modeling is to transform it into a stationary form. To achieve this, we address three different aspects.

1. **Constant variance:** although the initial time series plot already suggests that variance might not be constant over time, building a box plot by periods offers a clearer visualization of how variance behaves throughout years. It can be seen in the first picture.

The height of the boxes, which represents the interquartile range (IQR), tends to increase in the later years, indicating growing dispersion in the data. However, there are some exceptions, such as 2009 and 2004, where variance appears unusually low. In 2009, the lower variance can likely be attributed to the global financial crisis that started in 2008. During economic downturns, travel tends activity to decrease sharply worldwide due to reduced income and economic uncertainty. As a result, not only did tourism numbers decline, but they also fluctuated less, as the impact was global and affected the general population. Similarly, in 2004, the Madrid train bombings in March led to a temporary shock in tourism due to concerns over safety and security.

Nevertheless, excluding these exceptional events, the overall pattern suggests a potential heteroscedasticity in the series, meaning that the variance changes over time. To stabilize the variance, a logarithmic transformation is applied to the series.

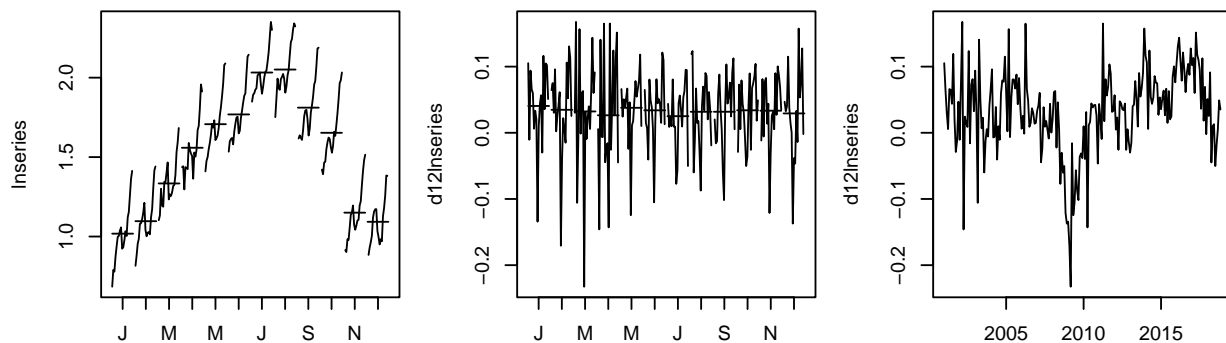
We repeat the plots and, as a result, we see that the issue has been fixed. The changes can be seen in the two plots below.



2. **Seasonal pattern:** as previously mentioned, there is a clear seasonal pattern in the data. However, to better highlight the pattern, we can revise the monthplot of the series (left plot)

Each year, peaks are observed during summer months, especially in July and August, likely due to the increase in tourism during the holiday period.

Consequently, we have to apply a seasonal difference to the series and we repeat the monthplot (figure in the middle) and the plot of the transformed series (figure on the right side), where we can see that the transformation has effectively attenuated the seasonal fluctuations in the series.



2. **Constant mean:** the final aspect to be checked is the presence of a trend. As it has been mentioned at the beginning, it seems, at first glance, that there the real series shows an upward trajectory that reflects the growth in tourism over the years.

In order the series to be stationary, it is necessary to remove this trend component. However, this should be done carefully, ensuring that variance doesn't increase as a result of the transformation. We proceed to visually inspect the transformed data obtained in the previous steps (with logarithmic transformation and a seasonal difference).

Since determining the need for a regular difference is not straightforward, we evaluate the variance of the series under the different transformations: logarithmic transformation, logarithmic transformation + seasonal difference and logarithmic transformation + seasonal + regular difference.

Comparing the variances, we observe that in the previous step the seasonal difference has decreased variance, while adding a regular difference results in an increase in variance, suggesting that it may not be beneficial.

Transformation	Variance
lnseries	0.156137
d12lnseries	0.004071
diff(d12lnseries)	0.004267

**Series:** Based on the analysis, we conclude that the optimal approach is to use the series with both logarithmic transformation and seasonal difference. This version of the series will serve as our stationary series for modeling.

The expression for the stationary series ( $W_t$ ) in function of the original series ( $X_t$ ) and the backshift operator ( $B$ ), is written as:

$$W_t = (1 - B^{12})\log(X_t)$$

#### 4.1.2 ACF and PACF analysis

After applying the logarithmic transformation and one seasonal difference, we are working with a model of the form:

$$ARIMA(p, 0, q)(P, 1, Q)_{12}$$

.

To determine the appropriate model parameters, we examine the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots shown above.

For the **seasonal part** (represented by the uppercase letters P and Q), we focus on the lags colored in red, corresponding to multiples of the seasonal period (lag 12, 24, ...).

It is difficult to detect a clear decreasing pattern. We observe that the seasonal lags in the PACF plot cut off after lag 2, suggesting a seasonal autoregressive structure. We interpret this pattern as  $P = 2$ , meaning a seasonal AR(2) component. Nevertheless, since the second significant lag is very close to the blue confidence line, we can also propose an AR(1),  $P = 1$ .

Alternatively, we can consider the PACF to have the infinite lags different from zero. We propose a MA(0) due to the lack of significant lags multiple to the seasonal part in the ACF plot.

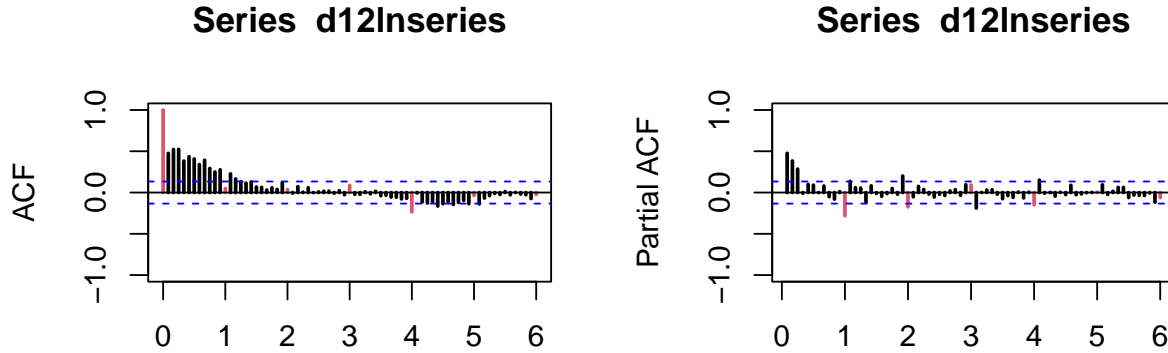
On the other hand, for the **regular part**, at first glance the autocorrelation function (ACF) shows that for the first few lags (up to around lag 12), the autocorrelations are strong and significant. Nevertheless, it shows a slowly decreasing pattern, gradually tapering off which is typical of an AR process. Since, the PACF has the last significant lag at position 3,  $p = 3$  and, consequently we propose a regular AR(3) component.

Based on the above, we have several model proposals:

- $ARIMA : (3, 0, 0)(2, 1, 0)_{12}$
- $ARIMA : (3, 0, 0)(1, 1, 0)_{12}$
- $ARIMA : (3, 0, 0)(0, 1, 0)_{12}$



Nevertheless, as we have mentioned, the ACF shows a suspicious shape with too many significant lags, which leads us to think that it might be either because of non-stationarity of the transformed series or outliers that could be affecting the autocorrelation structure.



In order to make sure that we are doing a correct analysis, we can use an alternative approach: the **auto.arima()** function from the forecast package in R.

This function automates the process of model selection by iterating over different combinations of parameters and selecting the best ARIMA model based on criteria such as the AIC. It has been proven to be a more efficient way of identifying an appropriate model, especially in cases where manually analyzing ACF and PACF plots might be ambiguous.

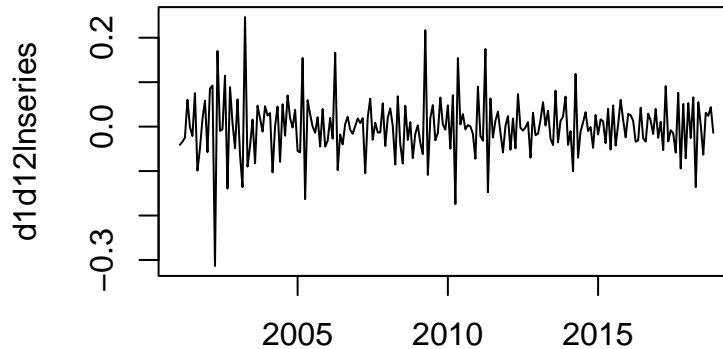
See Figure 1.

The automatic function, suggests an  $ARIMA(2, 1, 0)(2, 0, 0)_{12}$ . The non-seasonal component of the model has AR(2) with a regular differencing of 1 ( $d = 1$ ), and the seasonal component suggests AR(2) without seasonal differencing (since we have already applied a seasonal difference of 12 in the data).

Therefore, taking into consideration the already done transformations we have an  $ARIMA(2, 1, 0)(2, 1, 0)_{12}$  proposal.

Going back to the analysis we have done when regular differencing is applied, the increase in variance from this operation might be negligible and, indeed, an appropriate transformation to be carried out.

As a result, we change our working series and we now work with a series that has gone through a logarithmic transformation, a seasonal differencing and a regular differencing.



We repeat the ACF and PACF analysis. We are now working with a model of the form:

$$ARIMA(p, 1, q)(P, 1, Q)_{12}$$

At first glance, we notice that ACF no longer displays the strong autocorrelations in the first few lags that were previously present, allowing for a more reliable analysis.

For the **seasonal part**, we focus on the lags highlighted in red.

It is still challenging to detect a clear, decreasing pattern again. However, in the PACF plot, the seasonal lags appear to cut off after lag 1, suggesting a seasonal autoregressive structure. We interpret this as  $P = 1$ , indicating a seasonal  $AR(1)$ . Nevertheless, after the first two lags (12 and 24), there are still some observations out of the significance bounds. Although these are not very prominent, it has to be taken into consideration.

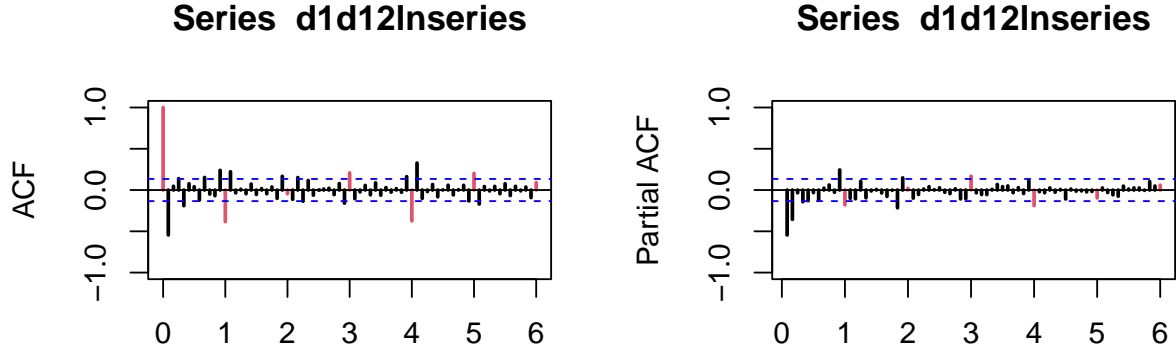
This observation, leads us to an alternative proposal. If we view the PACF as displaying many significant lags, while the ACF shows the last significant seasonal lag at lag 1, this could support the idea of a  $MA(1)$ . However, it is difficult to make a consistent diagnosis for this part.

Turning on the **regular part**, now we can observe a decreasing pattern in the PACF, especially within the first year of observations. So, if we focus on the last lag different from zero in the ACF function it corresponds to lag 4. Since lag 2 is not significant but lags 3 and 4 are, we propose a  $MA(4)$  for this part.

However, if we revise again the plots we can also deduce a decreasing tendency in the ACF, while the PACF shows that the first significant lag is at lag 2, suggesting an  $AR(2)$  structure.

Based on what we have said, we now have new model proposals:

- $ARIMA : (2, 1, 0)(2, 1, 0)_{12}$  {proposed by the automatic function}
- $ARIMA : (0, 1, 4)(1, 1, 0)_{12}$
- $ARIMA : (2, 1, 0)(1, 1, 0)_{12}$
- $ARIMA : (0, 1, 4)(0, 1, 1)_{12}$
- $ARIMA : (2, 1, 0)(0, 1, 1)_{12}$



## 4.2 Estimation

We now proceed with the estimation of the models. Since our series `d1d12Inseries` has already been differenced, with both seasonal and regular differencing applied beforehand, we set  $d = 0$  and  $D = 0$  in the model specification for the estimation and we use the transformed stationary series to obtain the mean estimation (Fig 2)

In order to determine if the mean (intercept) is significant in the proposed models we do a t-test:

- Hypothesis:

$$H_0 : \mu_{W_t} = 0$$

$$H_1 : \mu_{W_t} \neq 0$$

- Test Statistic:

$$\hat{t} = \frac{\hat{\mu} - 0}{S_{\hat{\mu}}} \sim N(0, 1)$$

- Decision rule:

$$|\hat{t}| > 2 \Rightarrow H_1$$

$$|\hat{t}| \leq 2 \Rightarrow H_0$$

Since means are not significant (Fig 3), we re-estimate the model for the original series (ln transform) specifying the differentiations in the estimation method (Fig 4) and we verify the significance of coefficients (Fig 5, 6, 7, 8, 9).

The coefficients of models 1, 3 and 5, which are  $ARIMA : (2, 1, 0)(2, 1, 0)_{12}$  {proposed by the automatic function},  $ARIMA : (2, 1, 0)(1, 1, 0)_{12}$  and  $ARIMA : (2, 1, 0)(0, 1, 1)_{12}$ , respectively, are all significant.

On the other hand, model 2 ( $ARIMA : (0, 1, 4)(1, 1, 0)_{12}$ ) has several non-significant coefficients: MA(2), MA(3) and MA(4). The least significant is MA(3), with a t-value = 0.8009. In order to fix it, we remove this coefficient and re-estimate the model.

Similarly, model 4 ( $ARIMA : (0, 1, 4)(0, 1, 1)_{12}$ ) has non-significant coefficients, MA(2) and MA(3). In this case, MA(2) is the least significant, and it will be removed in the re-estimation.

After the re-estimation (Fig 10), t-test is done again.

Removing the non-significant coefficients mentioned is not sufficient for any of these models, as they still include some that are not statistically significant: specifically, MA(4) coefficient in model 2 (Fig 11), and MA(3) and MA(4) coefficients in model 3 (Fig 12).

To improve the models, we will reduce the order of model 2 by decreasing its MA degree to eliminate the non-significant MA(4) term. For model 3, we will remove MA(3) term, since it is the least significant among the remaining coefficients.

After the new re-estimation (Fig 13), t-test is carried out again (Fig 14, 15).

Since new non-significant coefficients arise in both models after removing some of the initial ones, this indicates potential overparametrization. Removing parameters can cause the model to redistribute variance explanation, which leads to other parameters becoming non-significant. This suggests that the model specification may not be appropriate.

For this reason, these two models are ruled out, and we will focus on:

- Models 1:  $ARIMA : (2, 1, 0)(2, 1, 0)_{12}$
- Model 3:  $ARIMA : (2, 1, 0)(1, 1, 0)_{12}$
- Model 5:  $ARIMA : (2, 1, 0)(0, 1, 1)_{12}$

## 4.3 Validation

### 4.3.1 Residual analysis

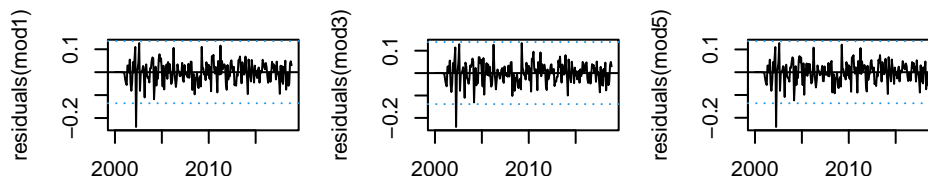
In order to do the validation of the models we first need to check if the residuals behave like white noise:

- **Homocedasticity (constant variance)**
- **Normality**
- **Independence**

In order to verify if **variance** is constant, we start by plotting the residuals.

In all three models, the residuals behave very similarly. Most of the residuals lie within the confidence bounds, except for one observation between 2002 and 2003, which appears to deviate from the overall pattern, this can be considered to be an outlier that we are going to take into account from now on.

Among the models, model 1 shows slightly narrower residual dispersion, indicating potentially better variance behavior.

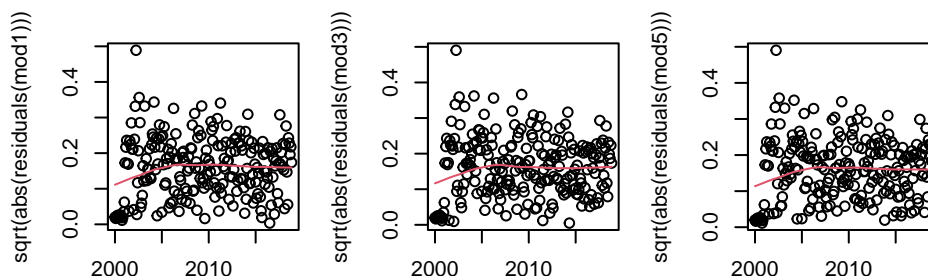


To further examine variability, we plot the square root of the absolute residuals and add a smooth trend line to observe potential changes in the spread of residuals over time.

In the ideal case of homoscedastic variance, we would expect to see a flat and stable smooth line.

Nevertheless, we observe a slight increase in variance during the initial years of the data, between 2000 and 2005, where we have previously observed the potential outlier. After this period, the line flattens and stabilizes in all three models, achieving a more stable variance in the later observations.

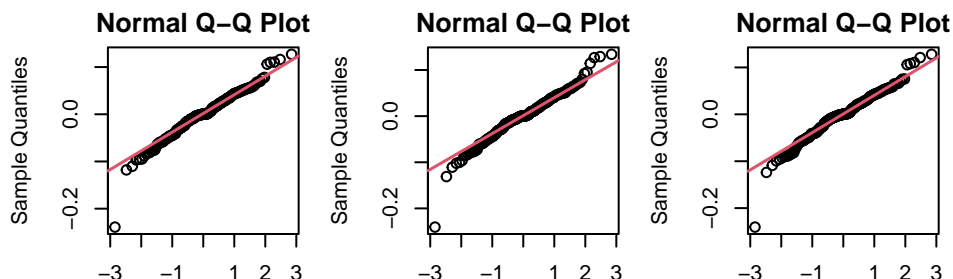
We can consider that variance is constant.



Then, to assess the **normality** of residuals, we proceed with the Q-Q plot analysis. On the left there is the mod1, in the middle mod3 and finally mod5.

Across all three models, the residual patterns are fairly consistent. The tails show deviations from normality, showing volatility, with model 3 displaying higher degree of dispersion compared to others.

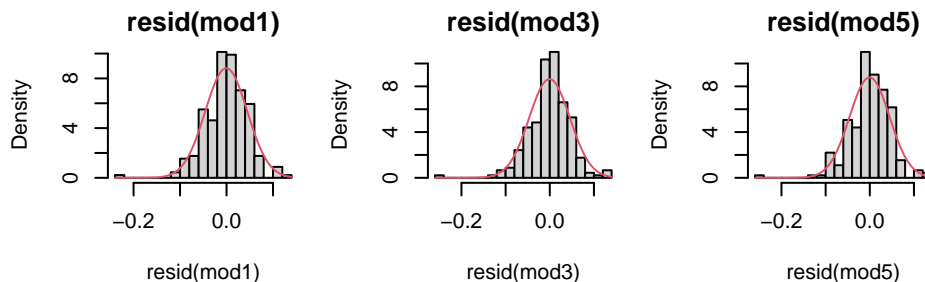
Particularly there is the presence of a single, isolated observation that lies far from the reference line, strongly indicating the influence of an outlier.



To further examine the normality of the residuals, we also plot the histogram.

In this case, we can see a small, isolated bar on the left tail of the histogram, which is likely associated with the isolated dot identified in the Q-Q plot. Additionally, it has to be mentioned that the central bars around the mean appear more concentrated and taller than the reference normal curve, suggesting a distribution that is slightly more peaked than the normal one.

These observations suggest that the presence of outliers could be altering the residual distribution. Probably, addressing them we could improve normality and overall model adequacy.



The previous graphical observations can make us think that the normality assumption might not hold for any of these three models, primarily due to the presence of outliers. To formally and numerically test it, we conduct the Shapiro-Wilk test on the residuals (Fig 16). In all three cases, p-value is below the 0.05 significance threshold, indicating evidence against normality assumption.

However, it is important to recognize that the Shapiro-Wilk test is sensitive to outliers. A single outlier can alter the test outcome, leading to a false rejection of normality. Considering what we have observed on the previous plot of residuals, where we identified certain isolated points that seem to behave differently from the rest, we can say that probably addressing these outliers could result in residuals that follow a normal distribution.

Finally, to assess if the model adequately captures the autocorrelation structure of data, we examine the **independence** of residuals.

First, we begin by plotting the ACF and PACF of the residuals. Ideally, if the model captures time dependence correctly, all autocorrelations should lie within the confidence bands.

In our case, the three models look pretty similar. We observe a few significant spikes at higher lags, far from the present. These, can be considered satellites, that suggest that the model may not fully explain the temporal structure of the data (Fig 17, 18, 19).

Moreover, to complement our assessment of residual independence, we conduct a Ljung-Box test (Fig 20, 21, 22)

This test evaluates whether any group of autocorrelations of the residuals is significantly different from zero. In all three models, from lag 48 onwards, the p-values remain consistently below the reference line, suggesting that there is a lack of independence among the residuals. This, again, may be due to the presence of an outlier that could be affecting the residuals' autocorrelation structure.

For earlier lags, most p-values are above 0.05, indicating that the models adequately capture the autocorrelation structure of the residuals. However, in models 1 and 3, a few observations are very close to the threshold.

Nevertheless, given that outlier treatment could fix the mentioned issues, we can assume that residuals are independent in all three models once the outliers are properly handled.

After conducting the residual analysis, we can confidently conclude that, after an appropriate outlier treatment, the residuals will likely satisfy the assumptions of constant variance, normality and independence, behaving like white noise.

#### 4.3.2 Model expressions

To express the structure of each model formally, we use the backshift operator  $B$ . The general form of a seasonal ARIMA model is:

$$\phi_p(B) \Phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B) \Theta_Q(B^s) Z_t$$

Now we are going to analyse our three models one by one. :

- **Model 1:**  $ARIMA(2, 1, 0)(2, 1, 0)_{12}$

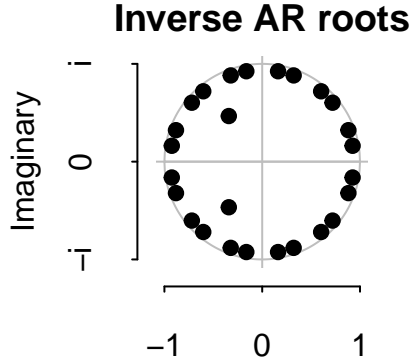
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12})X_t = Z_t$$

This model includes both non-seasonal and seasonal autoregressive terms of order 2, with first-order regular and seasonal differencing. There are no moving average (MA) components. Because of this, invertibility is not a concern in this case. However, the model admits an infinite moving average representation.

In order to assess stationarity and stability of the model, using the characteristic polynomials of the AR and MA components, we are going to calculate the modulus of their roots.

For the model to be stationary, all roots must lie outside the unit circle. Since the condition is met, we can conclude that model 1 is well-specified (Fig 23).

In the following plot we can visualize it. Since we are doing the **inverse**, all roots are inside the unit circle as expected.



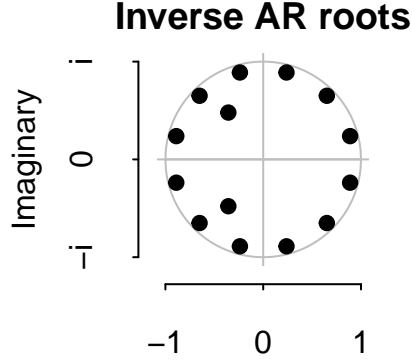
- **Model 3:**  $ARIMA : (2, 1, 0)(1, 1, 0)_{12}$

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})X_t = Z_t$$

This one is a simpler seasonal structure with a seasonal AR term at lag 12. Like the previous one, it admits an infinite MA representation.

In order to assess stationarity of the model, we examine the roots of the characteristic polynomial. Since invertibility is not applicable, we are going to focus on the autoregressive operators.

Since all roots lie outside the unit circle (Fig 24), we can conclude that model 3 is stable and well-specified.



- **Model 5:**  $ARIMA : (2, 1, 0)(0, 1, 1)_{12}$

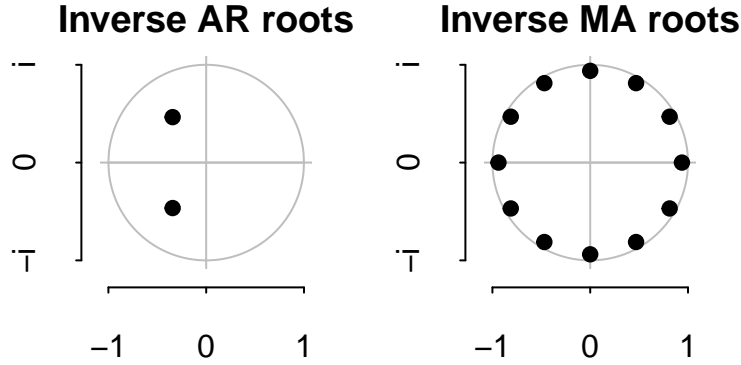
$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})X_t = (1 + \Theta_1 B^{12})Z_t$$

This model differs from the previous ones as it contains a seasonal moving average component (MA) instead of a seasonal AR term.

Because of the presence of both non-seasonal AR terms and seasonal MA term, we must check stationarity and invertibility conditions.

To ensure causality and stationarity, the roots of the non-seasonal AR polynomial must lie outside the unit circle as well as the roots of the seasonal MA polynomial, to ensure invertibility.

In this case, all roots have modulus greater than 1 (Fig 25). Consequently, we can conclude that the AR part is causal, the MA invertible and the model well-defined and stable.



Then, we assess and compare the model fit metrics for the three models.

Examining the AIC and BIC values, we observe that all three models have very similar metrics. Nevertheless, Model 5, has the lowest AIC and BIC values, indicating a better fit relative to other models. These results suggest that model 5 strikes a favorable balance between goodness of fit and model complexity, making it a more efficient choice for this data set.

Model	AIC	BIC
Model 1	-692.7635	-675.9336
Model 3	-686.9478	-673.4839
Model 5	-692.7936	-679.3297

### 4.3.3 Forecasting accuracy

In order to assess model stability and forecasting performance, we reserve the last 12 observations for validation. These are excluded from model training and used to evaluate how well each model predicts unseen data.

Down below, we present plots comparing the actual series with the forecasts produced by each model. The black line represents the original time series, the red line shows the forecasted values for the last 12 periods, and the blue lines indicate the 95% confidence intervals. Predictions from all models closely follow the same pattern, with some deviations, specially in predictions for summer months, where the peak is observed.

At the end, a summary table shows the main evaluation metrics: RMSE, MAE, RMSPE, MAPE and AIC for both the full and the cut models, residual variance, and the mean width of the confidence intervals.

As far as model stability is concerned, we observe that the differences between the AIC values of the full and cut models are consistent for the three models. This suggests that the model structures remain stable even when we have less data. Residual variances (sig2\_f and sig2\_c) also support this idea, since they remain almost the same across full and cut versions for each model.

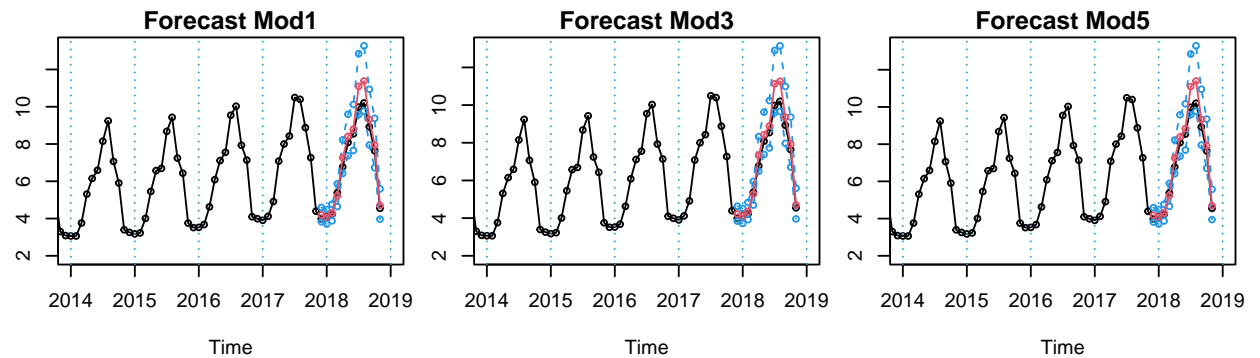
When it comes to forecasting accuracy, all three models have almost identical RMSE, MAE, RMSPE and MAPE values. Although differences are minimal, model 5 has the lowest error metrics. Since percentages are lower than 10%, our models are making reasonably accurate forecasts.

On the other hand, the mean width of the confidence intervals are also very similar but, in this case, model 1 produces slightly narrower intervals compared to others.

All the three models exhibit acceptable performance metrics. Nevertheless, since we need to select a single model for forecasting, it is essential to consider all the evaluation criteria discussed previously.

First, we can rule out model 3, as it is the one with weakest performance across the metrics.

If we compare model 1 and 5, model 5 has the lowest AIC for the full model and the lowest error metrics. However, model 1 slightly outperforms model 5 in terms of AIC for the cut model, as well as sigma squared and the width of the confidence intervals. Given that the differences where model 5 is better are relatively small compared to those where model 1 performs better, the choice finally falls on model 1:  $ARIMA(2, 1, 0)(2, 1, 0)_{12}$



Model	AIC_full	AIC_cut	sig2_f	sig2_c	RMSE	MAE	RMSPE	MAPE	CIWidth
Model 1	-692.7635	-651.8284	0.0022	0.0022	0.5535	0.4098	0.0602	0.0497	2.0162
Model 3	-686.9478	-645.2995	0.0023	0.0023	0.5524	0.4193	0.0612	0.0511	2.0388
Model 5	-692.7936	-651.7643	0.0022	0.0022	0.5545	0.4074	0.0600	0.0490	2.0212



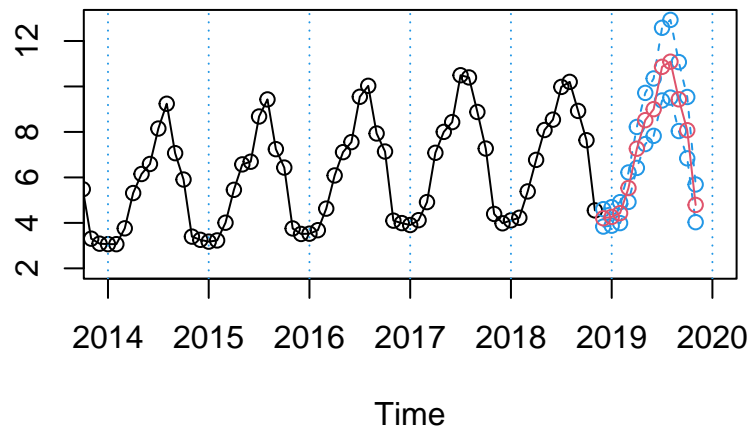
## 4.4 Forecasting

Based on the evaluation of forecasting accuracy and model stability, we now proceed generate long-term forecasts using the selected model:  $ARIMA(2, 1, 0)(2, 1, 0)_{12}$ .

Specifically, we are going to forecast the number of tourism arrivals in Spain for th 12 months following the last available observation, which corresponds to November 2018.

So, the forecast covers the period from December 2018 to November 2019 and it includes 95% confidence intervals to quantify the uncertainty associated with the predictions.

As it was expected, the months with the highest number of tourist arrivals correspond to summer season, while the lowest numbers occur during the winter months. (for more specific info: Fig 26)



## 4.5 Calendar Effects

Tourism activity can be influenced by various calendar-related factors, such as public holidays, school vacations or even the number of weekends in a month.

To take into consideration these potential influences, we propose modified models that explicitly incorporate calendar effects into the analysis by including them as external regressors (Fig 27).

So as to assess the relevance of these variables, we do a t-test on the corresponding coefficients (Fig 28) where the results show that both coefficients for the calendar effects are statistically significant.

Additionally, we detect that adding these effects influences the original model parameters: the coefficient for  $sar2$  is no longer significant.

Since the model that considers both calendar effects has the lowest AIC (Fig 27), we choose this one and remove the  $sar2$  coefficient.

So, the final model we adopt is  $ARIMA(2, 1, 0)(1, 1, 0)_{12}$  (Fig 29).

With this final model, we clean our original series of the Easter effect and traditional calendar effects and, doing this, we obtain an adjusted series that will undergo the same regular and seasonal differencing. (the comparison between the original and the adjusted series is shown in Fig 30).

## 4.6 Outlier Treatment

### 4.6.1 Automatic outlier detection

To conclude the analysis, we apply an automatic outlier detection procedure to the series already adjusted for calendar effects. This allows us to identify and interpret anomalies that may still be present in the data, thereby refining the model's accuracy.

Some outliers have been identified and the output shows several informative columns (Fig 31):

- Obs: index of the observation where the outlier is detected.
- Type: kind of outlier detected (AO, LS, TC)
- Weight coefficient: measures the magnitude and direction of the anomaly's effect, the larger, the stronger influence.
- Absolute likelihood ratio: significance of the outlier.
- Date: moment when the outlier is detected.
- PercVar: quantifies the deviation of the outlier from the expected series.

First, in March 2002 a positive additive outlier is detected. That year the Easter holidays occurred in March, which likely led to an unusual spike in tourism activity, despite the prior removal of average Easter effects.

Another positive additive outlier appears in August 2002, suggesting an exceptionally high volume of arrivals during the summer holiday period, probably due to favorable conditions or events that boosted travel beyond typical seasonal patterns. In February 2004, there is a positive additive outlier too. This can be linked to the fact that 2004 was a leap year, the additional day in February may have contributed increased tourist arrivals during this month.

Moving to April 2007, there is a negative transitory change observed. This may be related to early signs of the global financial crisis, which began affecting consumer behavior and, consequently, the tourism sector at that point.

In April 2010, a negative additive outlier appears. This likely reflects the effects of the volcanic ash cloud in Iceland, which led to the disruption of air traffic across Europe for several days in April. This caused a sharp decline in tourism flows during that month.

Furthermore, in April 2012 another negative additive outlier is detected, likely due to the increase in airport taxes introduced that month, which may discouraged travel and reduced inbound tourism.

Finally there is a positive level shift identified in April 2013, suggesting a sustained increase in the level of tourist arrivals from that point onward. This can be attributed to structural changes in the tourism sector, marketing strategies, or improvements in economic conditions that made tourism in Spain more attractive.

After analyzing the detected outliers, we now propose the linearized series, that is the version of the adjusted series for calendar effects now also adjusted for outliers.

In the Fig 32 plot, the red line shows the series we would expect if outliers had not occurred in the adjusted series for calendar effects and it represents the theoretical behavior of the data in the absence of anomalous events.

Fig 33 displays the difference between the log-transformed adjusted series for calendar effects and the linearized version for outliers and calendar effects. It shows the impact of outliers over time, showing where and how much the series has been adjusted. The peaks indicate points where significant outliers were detected and corrected.

Fig 34 shows the differenced linearized series (first a seasonal and then a regular difference), which have been done in order to achieve stationarity.

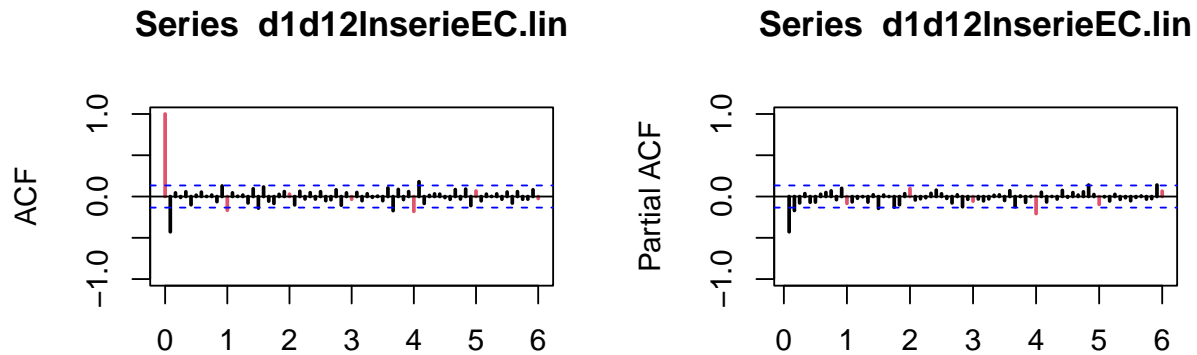
Since the series has changed, we now need to re-examine the ACF and the PACF plots to determine if we need to reformulate the model.

For the **seasonal part**, we focus on the lags highlighted in red.

Although it is still challenging to detect a clear and decreasing pattern, the ACF plot shows unusual behavior at seasonal lags. While the PACF does not show any significant seasonal spikes, the behavior in the ACF suggests a potential seasonal AR(1), leading us to consider  $P = 1$ .

Turning to the **regular part**, the lags appear to cut off after lag 2 in the PACF plot, indicating an AR(2).

Based on what we have said, we can confirm that the model we were working with is good for the adjusted series.



To validate our new adjustments and ensure the model is appropriate for the linearized series, we repeat the estimation of the model and the residual analysis.

Since our series d1d12InserieEC.lin has already been differenced, with both seasonal and regular differencing, we set  $d = 0$  and  $D = 0$  in the model specification for the estimation.

Like before, we first specify the transformed stationary series (Fig 35) in order to estimate the mean and check whether it is significant or not. Since the mean is not significant, we re-estimate the model without the corresponding differentiations. Instead, we need to specify them in the estimation method (Fig 36). The model will be called mod3EaTD.lin.

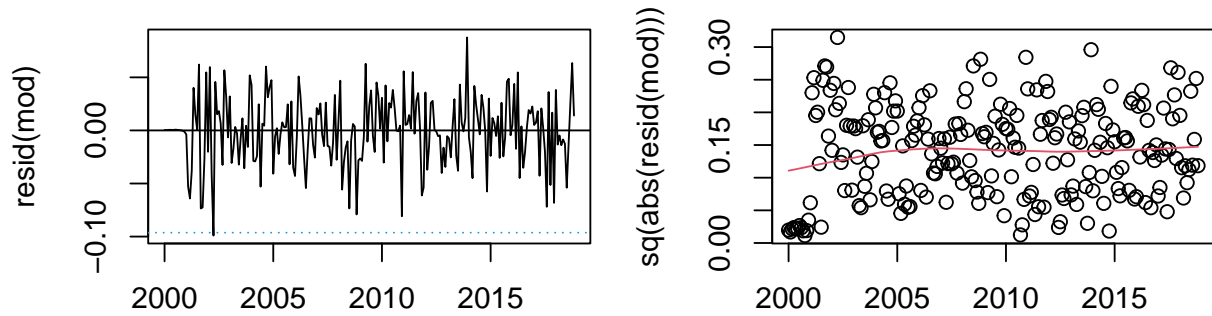
Afterwards, we proceed once again to verify the significance of coefficients carrying out the t-test, where we observe that all of them are significant (Fig 37).

In order to validate the model, as we have said before, we repeat the residual analysis where we have to check:

- **Homocedasticity (constant variance)**
- **Normality**
- **Independence**

To verify if **variance** is constant over time, we plot both the residuals and the square root of their absolute values, along with a smoothed trend line. This helps us to observe potential changes in the spread of the residuals.

We can see that now the red trend line is slightly flatter compared to the one obtained in the initial analysis. This means that the variance of the residuals is more stable with the currently adjusted series, suggesting improved homoscedasticity.

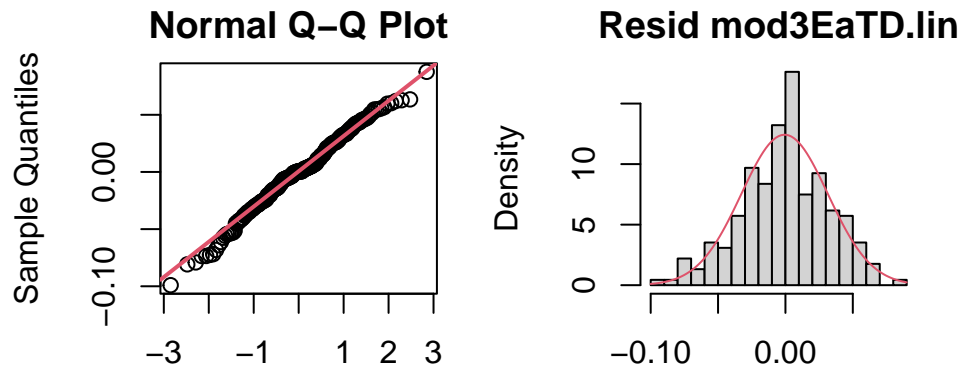


To assess the **normality** of the residuals, we proceed with the Q-Q plot and histogram analysis.

In this case, normality has also improved. The Q-Q plot shows minimal volatility, and the isolated point observed in the previous analysis is no longer present, indicating that the outlier has been removed and, consequently, we obtain a better fit to the normal distribution.

Concerning the histogram, the tails are now more aligned with the normal curve. Although the center still displays slightly higher bars than expected, the overall shape is much closer to a normal distribution. Therefore, we can conclude that the assumption of normality is satisfied.

The Shapiro-Wilk normality test confirms our visual analysis, since the p-value is 0.2349, higher than 0.05.



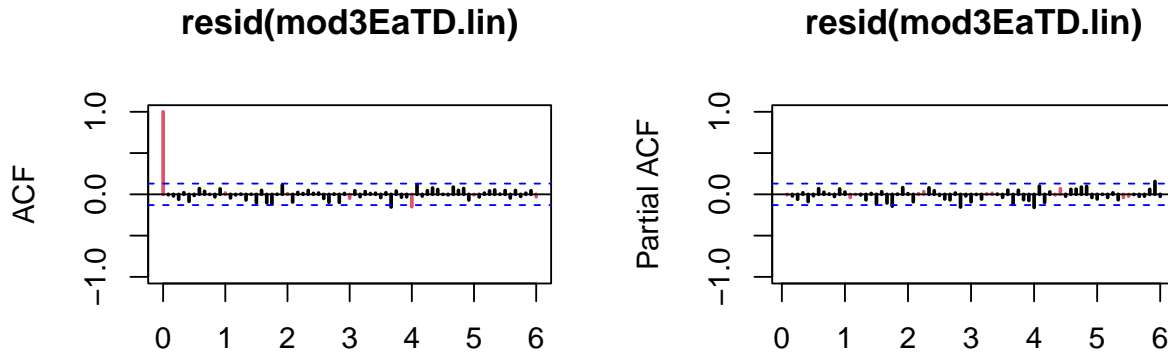
Finally, to assess whether the model adequately captures the autocorrelation structure of the data, we examine the independence of residuals.

First, we inspect the ACF and PACF plots of the residuals. The significant spikes that were previously observed in the first analysis have now been reduced and almost all of them lie within the confidence bounds, suggesting a better fit.

To complement our assessment of residual independence, we conduct a Ljung-Box (Fig 38). Unlike the earlier analysis, where p-values from lag 48 onwards remained consistently below the significance threshold, this time all p-values above 0.05.

This indicates that there is no significant autocorrelation in the residuals, and thus the model captures the autocorrelation structure well.

We can now firmly say that residuals are independent.



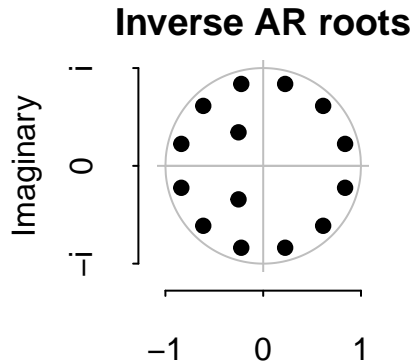
After conducting the residual analysis, we can confidently conclude that now residuals satisfy the assumptions of **constant variance, normality and independence**, behaving like white noise.

In order to assess stationarity of the model, we examine the roots of the characteristic polynomial.

Since invertibility is not applicable (there are no moving average terms), we are going to focus on the autoregressive components.

Since all roots lie outside the unit circle (Fig 39), we can conclude that the model is stable and well-specified.

As previously explained, the plot displays the inverse of the roots. That's the reason why, as expected, the dots lay inside the unit circle, which confirms that the actual roots lie outside.



Now that the series has been adjusted for outliers, we generate forecasts for the original series using the transformed model, which has gone through both calendar effects and outlier correction.

#### 4.6.2 Forecasts for the original series

To assess forecasting accuracy, we reserve the last 12 observations for validation. The model is estimated using the adjusted series, the one corrected for calendar effects and outliers, while forecasts are compared against the original (unadjusted) series.

Since we have all the information of the initial model (fitted on the unadjusted series), we can now compute the same performance metrics for the transformed model and compare them.

The AIC for the full model is significantly lower when using the cleaned series, indicating a better overall fit. This improvement is consistent when comparing both the full and the cut versions of the models, which confirms that the model is stable.

Similarly, the sigma squared, also called residual variance, is lower in both versions of the model fitted on the linearized series, suggesting a reduced unexplained variability.

The values for error metrics are also considerably lower for the model fitted on the cleaned series, which supports the conclusion that it does more accurate forecasts. Additionally, the mean confidence interval width is narrower in the updated model, implying greater precision and reliability in the forecasts.

Model	AIC_full	AIC_cut	sig2_f	sig2_c	RMSE	MAE	RMSPE	MAPE	CIWidth
mod1	-692.7635	-651.8284	0.0022	0.0022	0.5535	0.4098	0.0602	0.0497	2.0162
mod3EaTD.lin	-844.1085	-798.0691	0.0011	0.0011	0.2877	0.2688	0.0480	0.0428	1.4547

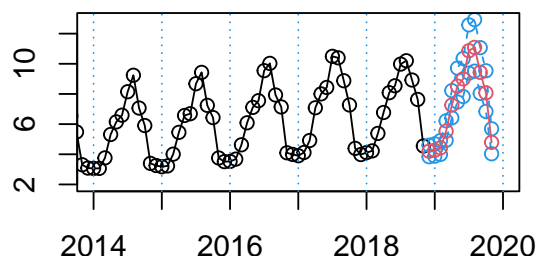
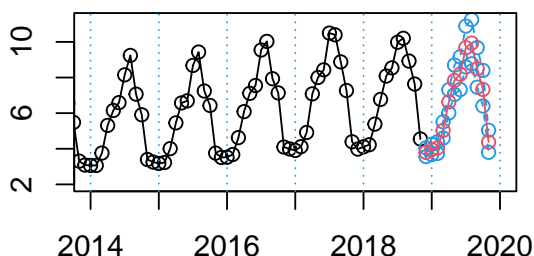
Finally, we compare the forecasts from both models for the future. Since we don't have the actual future data to check forecasting accuracy numerically, we perform a visual comparison of the predictions.

On the left plot there are the forecasts from the final model,  $ARIMA(2, 1, 0)(1, 1, 0)_{12}$ , fitted to to the linearized series, which has been adjusted for both outliers and calendar effects.

On the right, there are the forecasts from the previously selected model,  $ARIMA(2, 1, 0)(2, 1, 0)_{12}$ , fitted to the original, unadjusted series.

The forecasts generated by the adjusted model (left) show smooth and regular seasonal pattern, closely aligned with the historical behavior of the series. There are no abrupt spikes or dips and prediction intervals colored in blue seem reasonably tight.

In contrast, the forecasts from the model fitted on the original series show a larger amplitude, with higher peaks and variability. This can be likely caused due to the model being influenced by outliers and calendar effects in the original data. Forecast confidence intervals seem to be slightly wider, indicating greater uncertainty.



## 5 Conclusions and discussion

After completing the required analysis for model selection and validation, we conclude that the model that adjusts best our data is the  $ARIMA(2, 1, 0)(1, 1, 0)_{12}$ , which has been applied to a stationary series that was adjusted for both calendar effects and outliers, ensuring that the input data was clean and representative of the pattern.

The residuals of the selected model show good properties, resembling white noise, indicating that the model has successfully captured the structure of the series.

The model demonstrated good predictive performance, suggesting it generalizes well beyond the training data.

Through this process, we have developed a model that performs well not only on historical data but also in forecasting future values, which reinforces the reliability of our approach.

Apart from the statistical analysis, we also gained insight into various external factors that can significantly influence series, such as economic shifts, political events, holidays or even environmental conditions.

We can conclude that this series is highly sensitive to calendar effects and outliers, which must be properly addressed so as to produce accurate forecasts. Moreover, the series presents a clear seasonal pattern, with a pronounced increase in tourist arrivals during summer months, followed by a notable decline in colder seasons.

All of these conclusions have been made possible thanks to rigorous analysis, model selection and cleaning of the data. The ARIMA models that we have been able to propose, and particularly the final one, have enabled us to model the true structure of the series effectively.

## 6 References

Ministerio de Industria, Comercio y Turismo. <https://sedeaplicaciones.minetur.gob.es/Badase/BadasiUI/lstSeriesInformesPostBack.aspx?U64A2:FRONTUR+ENTRADA+DE+TURISTAS>

RPUbs - Modelos ARIMA. (s.f.). <https://www.rpubs.com/Meca/386432>

1.3.5.17. Detection of outliers. (n.d.). <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>

Historical events by year - on this day. (n.d.). On This Day. <https://www.onthisday.com/events-by-year.php>

España - Turismo internacional 2002. (n.d.). Datosmacro.com. <https://datosmacro.expansion.com/comercio/turismo-internacional/espana?anio=2002>

Wikipedia contributors. (2025, April 15). 2010 eruptions of Eyjafjallajökull. Wikipedia. [https://en.wikipedia.org/wiki/2010\\_eruptions\\_of\\_Eyjafjallaj%C3%B6kull](https://en.wikipedia.org/wiki/2010_eruptions_of_Eyjafjallaj%C3%B6kull)

## 7 Annex

In the file *TourismSpainScript.Rmd*, you will find the code used to conduct the analyses and studies related to the project.

Additionally, a PDF document titled *TourismSpainAnnex* contains all the figures that have been referenced throughout the report.