Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Modelling and prediction of a destination's monthly average daily rate and occupancy rate based on hotel room prices offered online

Noelia Oses, Jon Kepa Gerrikagoitia, Aurkene Alzua

Centro de Investigación Cooperativa en Turismo - CICtourGUNE

Donostiako Parke Teknologikoa

Mikeletegi Pasealekua, 71 - 3.Solairua

20009 Donostia (Spain)

Email: {NoeliaOses, JonKepaGerrikagoitia, AurkeneAlzua}@tourgune.org

Tel. +[34] 943 010 885 - Fax +[34] 943 010 846

Alternative email address for corresponding author: noelia@nofcon.com

Acknowledgements

This research was supported by the Basque Government's Department of Industry, Innovation, Commerce, and Tourism's "Etortek" program 2014-2015.

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Modelling and prediction of a destination's monthly average daily rate and occupancy rate based on hotel room prices offered online

Abstract

Tourism metrics are essential for managing a destination. Hotel performance metrics such as average daily rate and occupancy rate are two of the most prominent metrics for the industry. Our group works on developing methods for estimating tourism metrics based on digital footprint. Data available publicly on the Internet, including hotel room prices, is collected daily. This paper shows that the prices offered online have a high positive correlation with the prices reported by official statistics at the NUTS2 level after the online prices have been pre-processed and, thus, the relevance of this data source is established. Then, the paper presents a model for explaining and predicting mean hotel occupancy rates by destination based on these prices. The results are very promising, the fit is excellent and the predictions are also good. In summary, prices have moved from reflecting the expected demand to reflecting the actual demand and occupancy rate.

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Keywords

Average daily rate, occupancy rate, dynamic prices, hotel performance metrics, virtual channel closures

Introduction

Average daily rate (ADR), occupancy rate by room, and revenue per available room (RevPAR) are the three most important indicators in hotel performance metrics. ADR represents the average rental income per paid occupied room in a given time period and is calculated by dividing rooms' revenue earned by number of rooms sold in a given period. The occupancy rate by room is the ratio, as a percentage, between the average daily number of rooms occupied in a given period and the total number of rooms available. These metrics are an important source of information for hotel establishments to evaluate their pricing policies. Our group works on developing methods for calculating tourism metrics for a destination[1] based on digital footprint with the objective of offering figures to complement official statistics. Official statistics are based on surveys and, thus, incur a great delay. Estimating tourism metrics based on digital footprint is much faster and means hotels can compare their performance to the destination's average performance the first day of the following month, weeks in advance of the release of the official statistics.

This paper makes two contributions. First, it establishes the validity of hotel room prices offered online through an Internet distribution channel (IDC) as a data source for modelling tourist accommodation metrics. Although both the prices collected from the IDC and ADR refer to hotel room prices, the exercise of analysing the relationship between them is not inconsequential as the IDC prices refer to prices offered, but maybe not realised, and ADR is calculated based on actual revenue earned. Additionally, IDC prices represent a single type of price whereas official ADR statistics are calculated using the prices applied to different types of client. In the context of hotel room price data collected from an IDC, the term *channel closure* refers to the decision made by a hotel not to offer its rooms through the IDC for a date or series of dates. Similarly, *virtual channel closure* (VCC) refers to the action of offering a price so high that it is effectively a channel closure. This research has uncovered that raw prices must be processed to remove VCCs in order to obtain a good model fit and good ADR prediction accuracy. This paper presents the method developed to remove VCCs and estimate mean ADR by destination.

The second contribution is to present an alternative method to estimate the occupancy rate by destination based on these prices. In the past, hotels and tourist accommodation had fixed prices for the low and high seasons. The low season was when the demand was expected to be lower and, so, the price was lower, too. The high season was when the demand was expected to be higher, so the price was also higher. Looking at room prices offered through Internet distribution channels at the present time, it can be seen that they are dynamic and change according to hotel policy and external factors (Oses et al, 2015). Furthermore, looking at the price distribution by hotel, in some cases the prices appear in clusters. This is reminiscent of the times when there were prices for low and high seasons but, instead of setting the price based on the demand expected for the full season, they are now set based on the demand expected in the short term. This was the inspiration for the second contribution: hotel prices for a destination reflect the destination's occupancy rate. Proving this hypothesis is a step forward towards understanding how new technologies, such as the Internet, can facilitate new tourism accommodation pricing behaviours and understanding that these behaviours are not random but reflect economic factors such as demand. Additionally, modelling occupancy rates based on online prices is important because it establishes a link between the Big Data universe and traditional occupancy rate statistics.

Official statistics data provided by the Spanish National Statistics Institute (INE) has been used to fit the models and assess the prediction accuracy. INE publishes the results of the Hotel Occupancy Survey (HOS) monthly (available from (Instituto Nacional de Estadística, n.d.)). According to the INE methodology (Instituto Nacional de Estadística, 2013a), this survey offers information on the two aspects considered when analysing tourist trends: with regard to demand, there is information on guests, overnight stays and average stay, distributed by country of residence of the guests and category of the establishment they are staying in, or by Autonomous Community (i.e. NUTS2) of origin in the case of Spanish guests. As regards supply, the information includes the estimated number of establishments open for the season, the estimated number of bedplaces, the occupancy rate, and the information on employment in the sector, in terms of the category of the establishment. The HOS questionnaire also requests the ADR for accommodation services, excluding taxes and any other service, applied to different types of client, for a double room with a bathroom. These fees are broken down according to the type of client to which they have been applied: traditional tour operators, traditional travel agencies (including hotel vouchers and checks), companies, individuals, groups, direct hiring on the hotel website and/or the hotel chain website, online tour operators, online travel agencies, and others (Instituto Nacional de Estadística, 2013b). This information is offered monthly, on national, Autonomous Community, provincial (i.e. NUTS3), and tourist area levels. This research makes use of the data published by INE regarding ADR and occupancy rate for 2013 and 2014 by Autonomous Community. The models are trained with historical data of online prices for all hotels in Spain and official hotel performance statistics for 2013 at the Autonomous Community level. The prediction performance of the models has been evaluated using 2014 data. The goodness-of-fit of the models and the prediction results are very good.

Related work

The System of Tourism Statistics (STS), a part of the National Statistical System, aims to provide the user with reliable, consistent, and appropriate statistical information on the socio-economic structure and development of the tourism phenomenon. It can, in turn, be integrated with all the other economic and social statistics at different territorial levels (state, infra-state, and international) (Massieu, 2001). The new International Recommendations for Tourism Statistics (IRTS) (UNWTO, 2008a) and Tourism Satellite Account: Recommended Methodological Framework (TSA:RMF) (UNWTO, 2008b) constitute the updated reference framework for the STS for harmonization, coordination, and integration of available tourism statistical information.

The Internet has become part of society's basic infrastructure. More and more activities are taking place through the Internet, leaving behind digital footprints, which can be detected and measured in real time by robots (Heerschap *et al*, 2014). This also applies to online distribution of tourist accommodation, opening new possibilities for exploiting digital footprint data to obtain valuable information and complement official statistics through tools such as data mining algorithms (Magnini *et al*, 2007).

There has been significant progress in the development and improvement of official statistics on tourism. In the case of Ireland, subnational tourism statistics combine the analysis of supply and demand with spatial representation, using the destination scorecard as a framework for destination management (Wall and McFeely, 2012). In the case of Spain, there have been significant advances in the development of new statistical indicators by INE, recently adding profitability, Average Daily Rate (ADR), and Revenue Per Available Room (RevPAR) to their reports (Instituto Nacional de Estadística, 2008).

To the best of our knowledge, profitability indicators such as ADR are used to reflect past changes in rates, but not to predict future rates. Forecasting, however, is an essential tool to stakeholders (Chen, 2011; Haensel and Koole, 2011; Chu, 2009; Song and Li, 2008; Song *et al*, 2003; Yüksel, 2007; Witt and Witt, 1995; Athanasopoulos and Hyndman, 2008). Some companies have worked to fill this void and now provide commercial products for forecasting indicators of interest to the accommodation industry (PKF Hospitality Research, 2010; PwC Hospitality Directions Europe, 2010). The forecasts provided by these consulting firms are predictions of what the performance indicators will be in the future. The predictions provided by the model presented in this paper, however, are predictions of the performance indicators of past months. The objective of the former is to provide an insight into the future, whereas the objective of the latter is to provide alternative means to estimate an indicator that can inform decision-making in the interval before the official statistics are released.

From an academic point of view, several econometric models have been tested for forecasting time series (Shen *et al*, 2011). Forecasting articles in the field of tourism have been reviewed by Song and Li (2008). In particular, leading indicators are used by Yap and Allen (2011) and Kulendran and Witt (2003). In the tourism demand literature, it has been widely reported that income and tourism prices are the leading demand determinants in tourism demand analysis (Yap and Allen, 2011). The results presented in this paper provide more evidence of the relationship between prices and demand.

Price data collection methodology

The online travel agency (OTA) *booking.com* was chosen as the Internet distribution channel (IDC) from which to collect the data, as this is probably the largest and most popular hotel room booking website in Europe. To achieve this task, a data scraping bot customised for *booking.com* (Roman, 2012) is used. The bot collects prices offered for a single overnight stay in a 'Double or Twin room' for two adults for all hotels in Spain available through the IDC at the time of the request. The chosen room type for which prices are collected is 'Double or Twin Room', as this is the standard product for hotels and it is often used to study the accommodation market (Abrate *et al*, 2012). Hotel room prices change in real-time, so data must be gathered periodically. The large number of data available means that sampling is necessary. The method followed to collect the prices is based on that of Abrate *et al* (2012). They collected fares for hotel room reservations 1, 2, 4, 7, 15, 22, 30, 45, 60, and 90 days in advance of the date of the overnight stay (i.e. the target date). Extending this approach, the bot used collects prices 0, 1,..., 27, 45, 60, and 90 days in advance of the target date. The variables collected are: the hotel's name, address, and star rating, the collection date (i.e. the date when the price is collected form the website), the target date (i.e. the date of the overnight stay), and the price. The price is a real number (in Euro) if the scraping bot finds the hotel on the collection date, for the given target date, and there is a price for the double standard or twin room. However, the price can be missing if the bot cannot find a price for this type of room for a hotel returned by the IDC.

Modelling and prediction of average daily rate by destination

The hypothesis that this section aims to prove is that mean ADR by destination can be derived from hotel prices offered online. The method proposed to achieve this is explained in the following subsections. To prove that the first step, VCC removal, is necessary, the results will be presented with and without this step. As official ADR statistics are tax-free, 10% VAT tax has been deducted from the prices collected from the IDC.

The data for Islas Baleares has been used to illustrate this section. There are 2663289 non-missing observations for Islas Baleares for 2013 for a total of 710 hotels, with target dates from 2013-01-01 to 2013-12-31 (365 in total) and collection dates from 2012-10-11 to 2013-12-31 (424 in total). There are 2613829 non-missing observations for Islas Baleares for 2014 for a total of 712 hotels, with target dates from 2014-01-01 to 2014-12-31 (365 in total) and collection dates from 2013-12-12 to 2014-12-31 (378 in total).

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Virtual channel closure removal

Virtual channel closures wrongly inflate the mean price. Therefore, a method to detect and remove VCCs is proposed. The method can be summarised mathematically as follows. For a given hotel, let Π be the set of all the prices offered online by the hotel and collected by the crawler. Let X represent the sequence of sorted, unique prices in Π. Then, any price in Π greater than or equal to the following threshold is a virtual channel closure:

where is the n-th percentile of Π and and are thresholds. Thus, this method looks for the minimum of the prices above a certain percentile for which the absolute increment or proportional change with respect to the previous price is greater than or equal to the corresponding threshold. To obtain a general method, it is necessary to look for VCCs above a certain percentile as some hotels offer occasional bargains that would make normal prices exceed the proportional change threshold. The default values for these parameters used in this research are the mode (or , the 50-th percentile), equal to 200€, and equal to 2.

Price aggregation

The results of the HOS published by INE provide monthly mean ADR by Autonomous Community (AC). Therefore, the price data collected from the IDC must be aggregated to obtain the monthly mean price by AC. First, the price series are aggregated by hotel and target date using a geometric mean. Then, the monthly means are obtained using the arithmetic mean by month and AC. Table shows aggregate prices using the raw data and the VCC-free data for Islas Baleares.

Linear regression for ADR estimation

In order to use a linear model, the existence of a linear relationship between the dependent and independent variables must be ascertained first. The pairwise relationships between the aggregate prices calculated in the previous section (*IDC.PRICE*) and the official ADR data (*INE.ADR*) are shown graphically with aggregate prices obtained from raw data in Figure and aggregate prices obtained from VCC-free data on Figure . These plots clearly show that, when VCCs are removed, there is a linear relationship between the aggregate prices and the official ADR statistics. In fact, in this case Pearson's linear correlation coefficient goes from 0.528 for the raw data to 0.949 for the VCC- free data.

The fitted linear regression model is

where *IDC.PRICE* is the aggregate price by month and AC and *INE.ADR* are the official ADR statistics released by INE monthly. Table shows the coefficients fitted and the model fit assessment statistics both for raw data (first row) and VCC-free data (second row). These statistics clearly show that the model fit using VCC-free data provides a better fit. The R-squared value shows that the model fitted with VCC-free data explains the variance in the data very well, 90% variance explained with VCC-free data as opposed to 28% with raw data. The result of the F-test shows that the proposed relationship between the response variable and the predictor is statistically reliable for

VCC-free data and thus, the R-squared is, also, reliable. Finally, the residual standard error or RMSE is the standard deviation of the unexplained variance. The range of the response variable in this sample is 54.46–95.22. Thus, the RMSE, 4.25 in the VCC-free case, is low relative to this range. From these three measures, it can be concluded that the model fitted with VCC-free data is a good fit for the data. The fitted values can be compared to the response variable in Figure for the model fitted with raw data and Figure for the model fitted with VCC-free data. The plots confirm the results seen in Table . The model trained with VCC-free data results in a better fit with residuals considerably smaller at certain points (February, July, and August, in particular).

Prediction

Ultimately, the objective of this research is to produce models that allow predicting various hotel performance metrics. The previous section concluded with a model that fits the data well. This section shows that this model can be used for predicting mean hotel room ADR by destination.

Data for 2013 (the training set) has been used to fit the model. The resulting model has been applied to data for 2014 (the test set). Table and Table show the resulting predictions when the aggregate prices (IDC.PRICE) have been calculated with raw and VCC-free data, respectively. Column 3 of these tables shows the aggregate price obtained from our data. Column 4 shows the official ADR figures (INE.ADR). Column 5 shows the estimates for ADR obtained with the fitted model (ADR.EST). Finally, column 6 shows the residual error of the ADR estimate. The RMSE for the ADR estimate is 7.982 using raw data and 5.262 using VCC-free data. The prediction accuracy can often be improved by using all previous available data. For example, to predict the ADR for March 2014 the models would be fitted using data from January 2013 to February 2014. The RMSE for the ADR estimates calculated with updated models is 7.234 using raw data and 5.07 using VCC-free data. The RMSE is lower with VCC-free data in both cases (predicting with fixed and updated models), providing more evidence that it is important to detect VCCs and deal with them appropriately in each case.

Results

This section shows that the methodology described in the previous sections is relevant and can be applied to data from other destinations, not only Islas Baleares, with very good prediction accuracy. The results also show that VCC removal is advisable for the prediction of mean hotel room ADR by destination.

Table and Table show Pearson's linear correlation coefficient for raw and VCC- free data by Autonomous Community for 2013 and 2014, respectively. The '+' signs indicate those cases in which the linear correlation has improved when removing VCCs, which is the case for 11, in 2013, and for 10, in 2014, out of the 17 ACs. Additionally, the mean magnitude of the difference is greater when the correlation improves than when it worsens (0.0672 versus -0.024 in 2013 and 0.134 versus -0.0251 in 2014).

The linear models fitted to the 2013 data for each Autonomous Community (AC) in Spain and their corresponding assessment statistics are shown in Table and Table , fitted with raw and VCC-free data, respectively. Mean RMSE is 3.04 for raw data and 2.44 for VCC-free data.

Table and Table show the residuals for the ADR updated-model predictions with raw and VCC-free data, respectively. The most dramatic improvement occurs for Cataluña and Andalucía when the data is pre-processed to remove VCCs. In some cases, raw data results in a slightly better RMSE. But the magnitude of the improvement when the VCC-free model achieves a better result can be much larger compared to the magnitude of the difference when the raw data model outperforms the VCC-free model. The mean RMSE is 9.87 when the predictions are calculated with a fixed model fitted with raw data, 9.92 when the predictions are calculated with updated models fitted with raw data, 3.82 when the predictions are calculated with a fixed model fitted with VCC-free data, and 3.38 when the predictions are calculated with updated models fitted with VCC-free data. Thus, the updated, VCC-free model is the overall best performer.

Modelling and prediction of average occupancy rates by destination

Now that the validity of the data source has been established, the hypothesis that this section aims to prove is that mean occupancy rates for a destination can be derived from hotel prices. The inspiration for this hypothesis comes from the clusters that can be observed in some hotels' price distributions and the fact that, not so long ago, hotels used to have low and high season prices according to the expected demand for the year. Thus, the first step of this method is to segment the prices and label them accordingly and, then, build a linear regression model. The data used for this section is the full data set including VCCs.

This section is illustrated using data for the Basque Country Autonomous Community. There are 1984149 non-missing observations for the Basque Country for 2013, with target dates from 2013-01-01 to 2013-12-31 (365 in total) and collection dates from 2012-10-08 to 2013-12-31 (450 in total). There are 1776256 non-missing observations for the Basque Country for 2014, with target dates from 2014-01-01 to 2014-11-30 (334 in total) and collection dates from 2013-12-12 to 2014-11-30 (354 in total).

Segmentation of prices

The focus of this section is to model and predict hotel occupancy rates. We propose a pre-processing step based on the segmentation of these prices. For each hotel, three percentiles of its prices (P_{30}, P_{70}, and P_{99}) have been used as the bounds for the four regions, which are defined as follows: the first region is named Seg1 and covers the prices in the $[0,P_{30})$ interval; the second region is Seg2 and its interval is $[P_{30}, P_{70}]$; the third Seg3 and the interval $(P_{70}, P_{99}]$; and Seg4 with interval $(P_{99}, +\infty)$. Each price has been labelled with the name of the region it belongs to. These segments can be interpreted in the following manner. Seg1 corresponds to the low prices. Seg2 is the segment of ”normal” or middle prices. Finally, Seg3 and Seg4 correspond to high and unusually high prices, respectively.

The collection method collects 31 prices per hotel for each target date and, as the number of days in each month is known, the expected number of collected prices can be calculated based on the number of active hotels operating online each month. With this expected number of prices, the percentage of prices that fall into each segment can be calculated. However, the number of collected prices often does not match the expected number as hotels sometimes decide not to offer their rooms through the OTA, e.g. when they have no vacancies, are close to full capacity, or are closed. Therefore, a last variable, named NoShow, provides the percentage of missing prices. The result of the segmentation of the prices offered online by hotels in the Basque Country during 2013 can be seen in Table . The last column in this table, named ”INE”, provides the official occupancy rate by room extracted from the INE’s Hotel Occupancy Survey (Instituto Nacional de Estadística, n.d.).

Linear regression model fitting

Seg1, Seg2, Seg3, Seg4, and NoShow are collinear as they represent the percentage of prices in each range (Seg1 to Seg4) and the percentage of missing prices (NoShow) with respect to the number of expected prices. Thus, to fit a linear model, at most four out of these five variables must be selected.

The first step is to explore whether there is a linear relationship between the target measurement (or dependent variable), ”INE”, and the segmented price measurements (or independent variables), Seg1 to Seg4 and NoShow. The linear correlation coefficients can be seen in Table and Figure shows these pairwise relationships graphically. Both Seg2 and NoShow have a weak linear correlation coefficient with ”INE” and the plot shows the relationship is not linear. Thus, they are dropped and the linear model is fitted using Seg1, Seg3, and Seg4, which have strong linear correlations with the response variable.

A linear regression model of ”INE” given Seg1, Seg3, and Seg4 is fitted. Table shows the estimated model parameters. The fitted values can be compared to the response variable in Table and Figure . The adjusted R-squared value is 0.96351, which shows that this model explains the variance in the data very well. The F-statistic is 97.8 with a p-value below 0.00001, which shows that the proposed relationship between the response variable and the set of predictors is statistically reliable. Thus, the R-squared is, also, reliable. Finally, the residual standard error or RMSE is 2.35. The theoretical range of the response variable, the occupancy rate, is $0 - 100$, and, in this sample, $34.78 - 75.62$. Thus, the RMSE is low relative to this range. From these three measures, it can be concluded that this model is a good fit for the data.

Prediction

The previous section concluded with a model that fits the data well. This section studies whether this model is suitable for predicting occupancy rates for new data. Data for 2013 has been used to fit the model. Now, the segmentation bounds obtained for the 2013 data (the training set) and the fitted model are applied to data for 2014 (the test set). Columns 2 to 6 of Table show the resulting segmentation percentages, or independent variables, for 2014. Column 7 shows the official occupancy rates published by INE (at the time of writing, the official occupancy rates for November and December 2014 are not available). The predictions provided by this model can be seen in column 8 and Figure . Column 9 shows the residuals.

Results

This section shows that the methodology described in the previous section is relevant and can be applied to data from other destinations, not only the Basque Country. The linear models fitted to the 2013 data for each Autonomous Community (AC) in Spain and their corresponding assessment statistics are shown in Table . For Canarias and Castilla-La Mancha, and to some extent for Madrid, the null hypothesis of the F-test cannot be rejected and, thus, the fitted model might not be performing better than the mean model. For Canarias, for example, the mean model is 77. The RMSE of the fitted model is 4.9 and that of the mean model is 5.3 for 2013. For the 2014 predictions, the fitted model has an RMSE of 4.8 and the RMSE for the mean model is 6.4. Thus, both models are quite close. The fit of the models for the remaining Autonomous Communities is very good.

The models fitted in the previous section have been used to predict the occupancy rates for 2014. Table shows the predictions and Table the corresponding residuals. For a model to produce good predictions, it is necessary that the conditions in which the model was fitted remain stable. In this case, the predictions are good if the demand and the hotels in a destination are behaving similarly in 2013 and 2014. However, the residuals show that the predictions for some destinations are very inaccurate.

The accuracy of the predictions can be improved by re-fitting the model using all of the available training data. This way the segmentation boundaries are recalculated and account for changes in the price distributions. Table shows the predictions calculated with the updated models and Table the corresponding residuals (due to the processing time required this test has only been performed for the first sixteen Autonomous Communities). The residuals show that these predictions are much better than the predictions of the fixed model in all cases but the Basque Country’s (País Vasco) case.

Discussion and future work

The main conclusion to be drawn from these results is that monthly mean prices by destination calculated from prices offered online have a strong positive linear correlation with official ADR figures when VCCs are removed. This is a significant finding because price data collected from the IDC refers to **offered** prices whereas the ADR reported by hoteliers is based on **charged** prices. Furthermore, the data used for this research are prices collected from a single IDC, but the HOS questionnaire requests the ADR for accommodation services broken down according to the type of client to which they have been applied: traditional tour operators, traditional travel agencies, companies, individuals, groups, direct hiring on the hotel website and/or the hotel chain website, online tour operators, online travel agencies, and others. The strong positive correlation between aggregate IDC prices and official ADR figures suggests that IDC prices set the trend for the prices charged by the hotels to the other types of client.

For Castilla-La Mancha, Cataluña, and Aragón, the linear correlation coefficient between these two measures is unexpectedly low. It would be interesting to further examine the reasons behind this as it is not due to VCCs and, even though the nature of the prices used to calculate IDC aggregate prices and official ADR figures is different (offered versus charged), it is surprising to find such a large disconnect.

The importance of detecting and dealing appropriately with virtual channel closures has been highlighted. VCCs can be found under different circumstances; when a hotel is closed for a period but still wants to appear in the searches, or when it is full or close to full, for example. The results show that the presence of VCCs has an adverse effect on the linear correlation between online prices and official ADR. The removal of VCCs improved the results dramatically for Cataluña and Andalucía, two of the most popular tourism destinations in Spain. The large residual for Cataluña in November 2014 can be due to the celebration of the independence vote, when presumably many reporters travelled to Cataluña to cover this rare event and gave rise to an increase in VCCs.

Additionally, the results show that the methodology described here for the modelling and prediction of mean hotel room ADR by destination is a promising research line. The model allows predicting the mean ADR of a month by destination the first day of the following month, far in advance of the release of the official statistics. The accuracy of the estimates provided by the model varies for the different Autonomous Communities. Further research is necessary to determine whether customising the VCC removal algorithm's parameters for the different Autonomous Communities would improve the prediction accuracy.

Regarding occupancy rate, the results presented in this paper show that the methodology described here for the modelling and prediction of a destination's average hotel occupancy rate is also a promising research line. The proposed model provides a direct and clear explanation of occupancy rates in terms of hotel room prices offered online. The key is the dynamic nature of the prices.

As mentioned earlier, models are expected to achieve good prediction accuracy when the environment is stable. In this case, Spain is going through a major economic crisis that is affecting the different Autonomous Communities differently. Hotel room prices have generally risen in 2014 compared to 2013, but the rate of change has been different for the different destinations. Thus, accommodation prices and demand have not been stable during this period. Under these unstable circumstances, it can be said that the model's performance is satisfactory.

Again, the accuracy of the estimates provided by the model varies for the different Autonomous Communities. There are several factors that can be causing this. The most immediate factor is the different uses of segment four (Seg4) prices. When a hotel has received many bookings for a certain date, which would lead to reporting a high occupancy rate, the hotel might decide to disappear from the OTA for that date (reflected in the NoShow variable) or offer an unusually high price on the OTA (reflected in the Seg4 variable). Additionally, some hotels are closed for a period (seasonal hotels). Some of these hotels choose to disappear from the OTA for that period. However, others choose to show on the OTA with unusually high prices for the closing period. To improve the fit and prediction accuracy of the model, it is essential to detect these behaviours and label unusually high prices correctly in two different classes. Finally, the case of the Basque Country is a special one insofar as the updated model achieves a worse performance than the fixed model. The reason for this unexpected behaviour must be investigated.

The number of active hotels and the size of the sample are different for different destinations. Within the same destination, these vary for the different months and years. These variations are greater for some destinations. As the model is based on percentage of prices in each segment, the smaller the sample the bigger the impact that variations have on the percentages. The accuracy level could also be related to the level of technological engagement of the hotel industry in a destination. Some hotels have very dynamic online prices, whereas others have constant or nearly constant prices that do not provide any information regarding the occupancy rate. In relation to these points, it would be interesting to study what impact customising the price segmentation method for each destination has on the model fit and accuracy. Also, the segmentation bounds could be recalculated taking into account all the available data before applying to model to produce a prediction, as the segmentation step relies only on the available IDC price data and not on data related to occupancy rates.

The testing of these hypotheses with the objective of reducing the prediction error is an interesting line of future work. Other lines of future work include investigating whether it is possible to modify the model to predict ADR and occupancy rates before the month is over.

Conclusions

The technological developments in the past couple of decades mean that the world now moves faster than ever. Much of the speedup is due to the Internet, which has had a great impact on tourism practices. Traditional tourism measurement methods are low-tech, slow, and costly. New measurement methods appropriate for the current level of technology are necessary. Additionally, the vast amount of available data means it is now possible to develop predictive models to inform decision-making processes.

This paper makes two important contributions. First, it validates hotel room prices offered online as a data source. Second, it presents two models for modelling and prediction of average ADR and occupancy rate by destination, respectively.

The proposed method for modelling and predicting monthly mean hotel room ADR by AC based on said prices achieves good results. Additionally, the importance of detecting and dealing appropriately with virtual channel closures has been highlighted. Further research must be conducted to determine the factors that make the model have better prediction accuracy in some destinations.

This paper also presents a model for explaining and predicting a destination's average hotel occupancy rate based on online room prices. The proposed model provides a direct and clear explanation of occupancy rates in terms of online hotel room dynamic prices. The model is important because it provides an alternative, fast method to estimate a destination's average occupancy rate before the official figures are released. The good results obtained show that prices have moved from reflecting the expected demand to reflecting the actual demand and occupancy rate.

RevPAR can be calculated by multiplying ADR by the occupancy rate. As this paper presents models to estimate the two latter metrics, it also provides an alternative method to estimate RevPAR.

In summary, this paper is a step forward towards understanding how new technologies, such as the Internet, can facilitate new tourism accommodation pricing behaviours and understanding that these behaviours are not random but reflect economic factors such as demand. It also establishes a link between online hotel prices and traditional ADR and occupancy rate statistics, i.e. a link between the big data universe and traditional tourism statistics. Further research must be conducted to determine the factors that make the models have better prediction accuracy in some destinations.

References

Abrate, G., Fraquelli, G., and Viglia, G. (2012), 'Dynamic pricing strategies: Evidence from european hotels', *International Journal of Hospitality Management*, Vol 31, No 1, pp. 160–168.

Athanasopoulos, G. and Hyndman, R. J. (2008), 'Modelling and forecasting australian domestic tourism', *Tourism Management*, Vol 29, No 1, pp. 19 - 31.

Chen, K.-Y. (2011), 'Combining linear and nonlinear model in forecasting tourism demand', *Expert Systems with Applications*, Vol 38, No 8, pp. 10368 - 10376.

Chu, F.-L. (2009), 'Forecasting tourism demand with arma-based methods', Tourism Management, Vol 30, No 5, pp. 740 - 751.

Haensel, A. and Koole, G. (2011), 'Booking horizon forecasting with dynamic updating: A case study of hotel reservation data', *International Journal of Forecasting*, Vol 27, No 3, pp. 942 - 960.

Heerschap, N., Ortega, S., Priem, A., and Offermans, M. (2014), 'Innovation of tourism statistics through the use of new big data sources', In Global Forum of Tourism Statistics, Prague.

Instituto Nacional de Estadística. (n.d.), 'Hotel occupancy survey', http://www.ine.es/jaxi/menu.do?L=1&type=pcaxis&path=%2Ft11/e162eoh&file=inebase (Last accessed: 12th December, 2014)

Instituto Nacional de Estadística. (2008), 'Indicators on the profitability of the hotel sector', http://www.ine.es/jaxiT3/Tabla.htm?t=2057 (Last accessed: 10th February, 2015)

Instituto Nacional de Estadística. (2013a), 'Hotel occupancy survey 2013: methodology', http://www.ine.es/en/daco/daco42/ocuphotel/notaeoh_13_en.pdf (Last accessed: 12th December, 2014)

Instituto Nacional de Estadística. (2013b), 'Hotel price index (hpi). base 2008 (as of january 2014) methodological note', http://www.ine.es/en/daco/daco42/prechote/meto_iph_base2008_en.pdf (Last accessed: 12th March, 2015)

Kulendran, N. and Witt, S. F. (2003), 'Leading indicator tourism forecasts', Tourism Management, Vol 24, No 5, pp. 503 - 510.

Magnini, V., Honeycutt, E., and Hodge, S. (2007), 'Hotel management and operations', In D. Rutherford & M. O'Fallon (Eds.), (4th ed., p. 399- 414). New York: Van Nostrand Rheinhold.

Massieu, A. (2001), *Tourism statistics: international perspectives and current issues*, In J. Lennon (Ed.), (p. 3-13). London: Continuum.

Oses, N., Gerrikagoitia, J., and Alzua, A. (2015), 'Dynamic pricing patterns on an internet distribution channel: the case study of Bilbao's hotels in 2013', In Enter Conference on Information and Communication Technologies in Tourism.

PKF Hospitality Research (2010), 'Assessing accuracy: Hotel horizons forecasts', http://www.pkfc.com/en/pkfhome/nism/accuracyassessment/PKFAccuracy0610.pdf (Last accessed: 15-12- 2010)

PwC Hospitality Directions Europe (2010), 'UK hotels forecast 2011 and 2012: How big a party for hotels in 2012?' http://www.pwc.co.uk/hospitalitydirections (Last accessed: 15-12-2010)

Roman, I. (2012), *Measuring the influence of events in hotel room prices,* Unpublished Master's thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea.

Shen, S., Li, G., and Song, H. (2011), 'Combination forecasts of international tourism demand', *Annals of Tourism Research*, Vol 38, No 1, pp. 72 - 89.

Song, H. and Li, G. (2008), 'Tourism demand modelling and forecasting - a review of recent research', *Tourism Management*, Vol 29, No 2, pp. 203 - 220.

Song, H., Witt, S. F., and Jensen, T. C. (2003), 'Tourism forecasting: accuracy of alternative econometric models', *International Journal of Forecasting*, Vol 19, No 1, pp. 123 - 141.

UNWTO. (2008a), 'The conceptual framework for tourism statistics - international recommendations for tourism statistics 2008 (IRTS 2008)', http://statistics.unwto.org/content/irts-2008 (Last accessed: 2nd February 2015)

UNWTO. (2008b), 'The conceptual framework for TSA - tourism satellite account: Recommended methodological framework 2008 (TSA:RMF 2008)', http://statistics.unwto.org/content/tsarmf-2008 (Last accessed: 2nd February 2015)

Wall, C. and McFeely, S. (2012), 'Ireland case study: measuring and analysing regional tourism', In UNWTO/InRoute 1st Seminar on Regional Tourism: setting the focus.

Witt, S. F. and Witt, C. A. (1995), 'Forecasting tourism demand: A review of empirical research', *International Journal of Forecasting*, Vol 11, No 3, pp. 447 - 475.

Yap, G. and Allen, D. (2011), 'Investigating other leading indicators influencing Australian domestic tourism demand', *Mathematics and Computers in Simulation*, Vol 81, No 7, pp. 1365 - 1374.

Yüksel, S. (2007), 'An integrated forecasting approach to hotel demand', *Mathematical and Computer Modelling*, Vol 46, No 7–8, pp. 1063 - 1070.

Table Aggregate prices for Islas Baleares for 2013

AC	Month	Raw mean price	VCC-free mean price
Islas Baleares	2013-Jan	111.39	72.08
Islas Baleares	2013-Feb	146.74	83.36
Islas Baleares	2013-Mar	111.06	84.35
Islas Baleares	2013-Apr	87.36	86.53
Islas Baleares	2013-May	86.41	84.43
Islas Baleares	2013-Jun	107.82	106.23
Islas Baleares	2013-Jul	131.43	128.72
Islas Baleares	2013-Aug	145.09	141.11
Islas Baleares	2013-Sep	129.08	119.42
Islas Baleares	2013-Oct	91.44	86.15
Islas Baleares	2013-Nov	111.93	100.68
Islas Baleares	2013-Dec	104.65	97.87

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

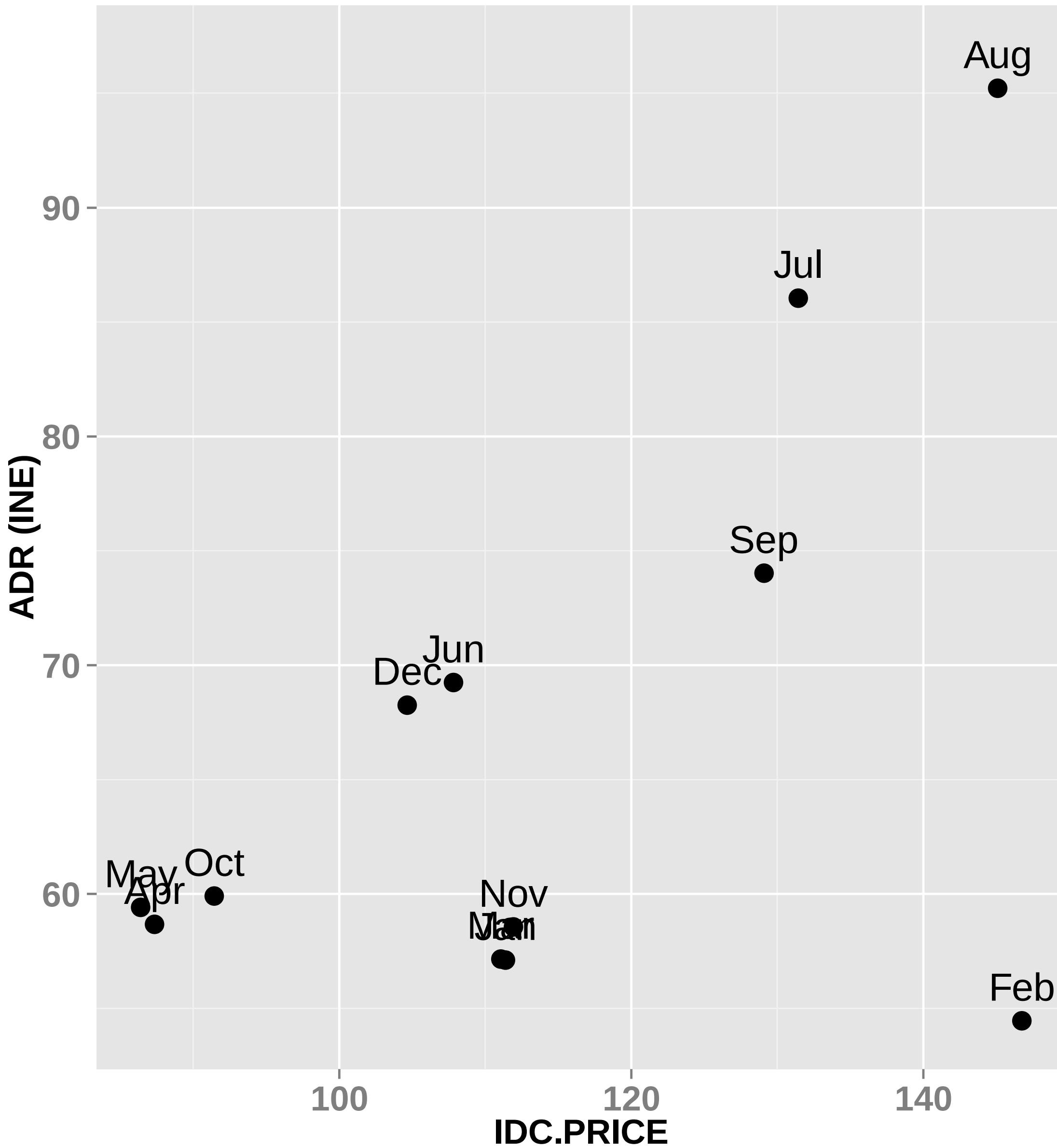
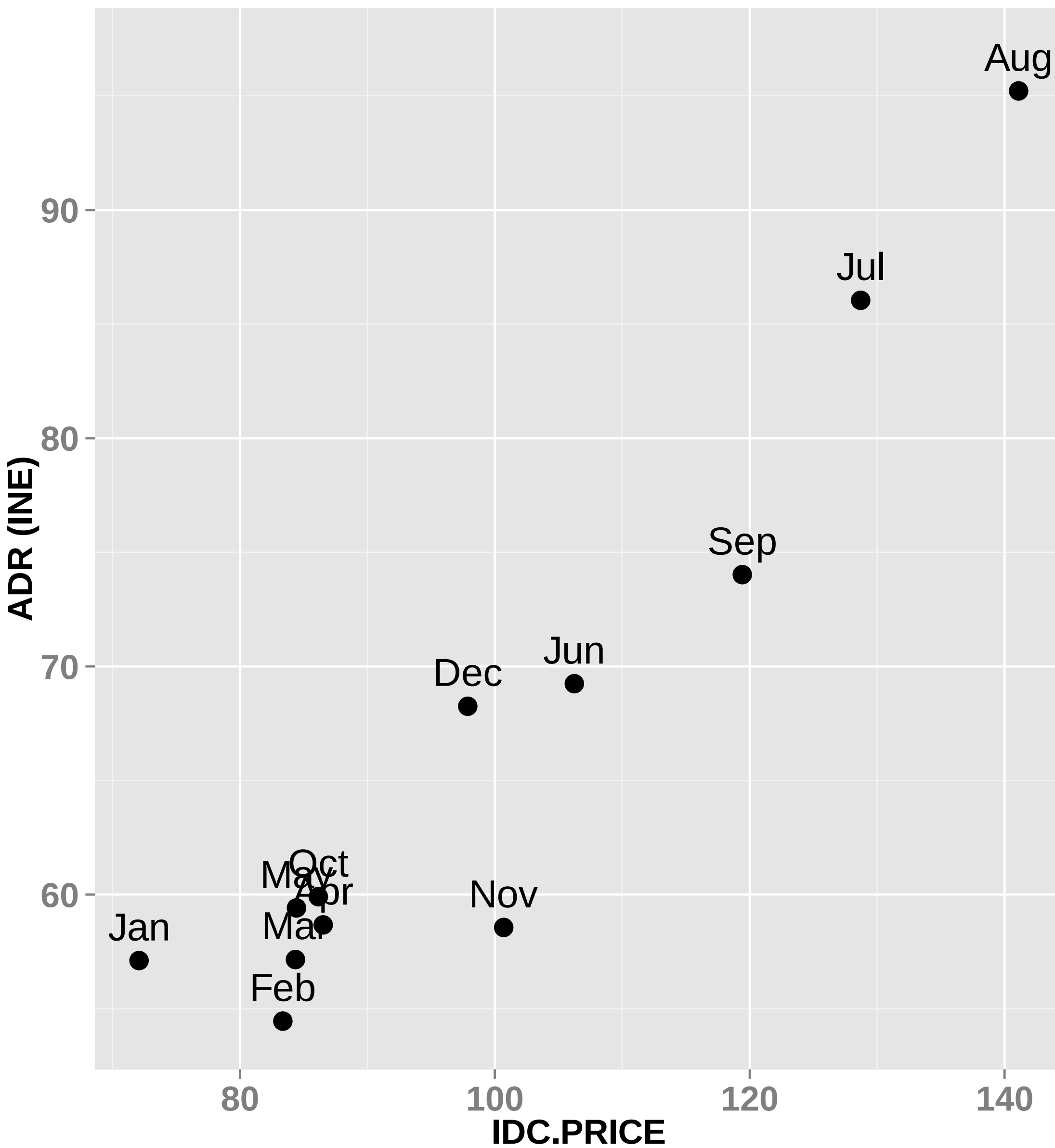


Figure Mean price-ADR pair plot for raw data


Aug
Jul
Sep
Jun
Dec
Oct
May
Apr
Nov
Jan
Mar
Feb
90
80
70
60
80
100
120
140
ADR (INE)
IDC.PRICE

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Figure Mean price-ADR pair plot for VCC-free data

Table ADR model fit for Islas Baleares for 2013

Data	Alpha	Beta	RMSE	R-Squared	F-Test	p-value
Raw	29.20	0.33	11.43	0.28	3.87	0.08
VCC-free	8.92	0.58	4.25	0.90	90.19	0.00

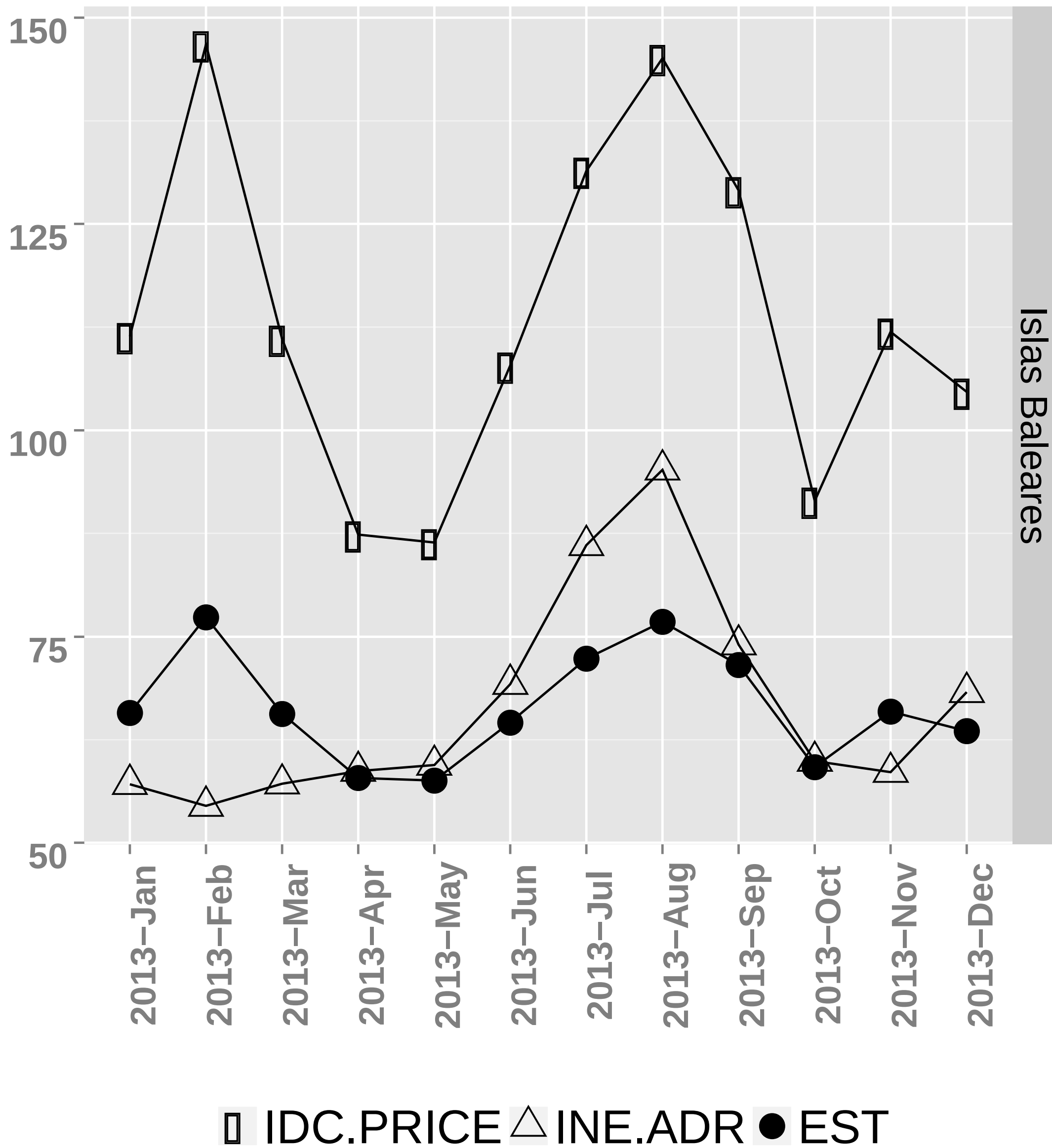


Figure ADR model fit plot for Islas Baleares for 2013 with raw data

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

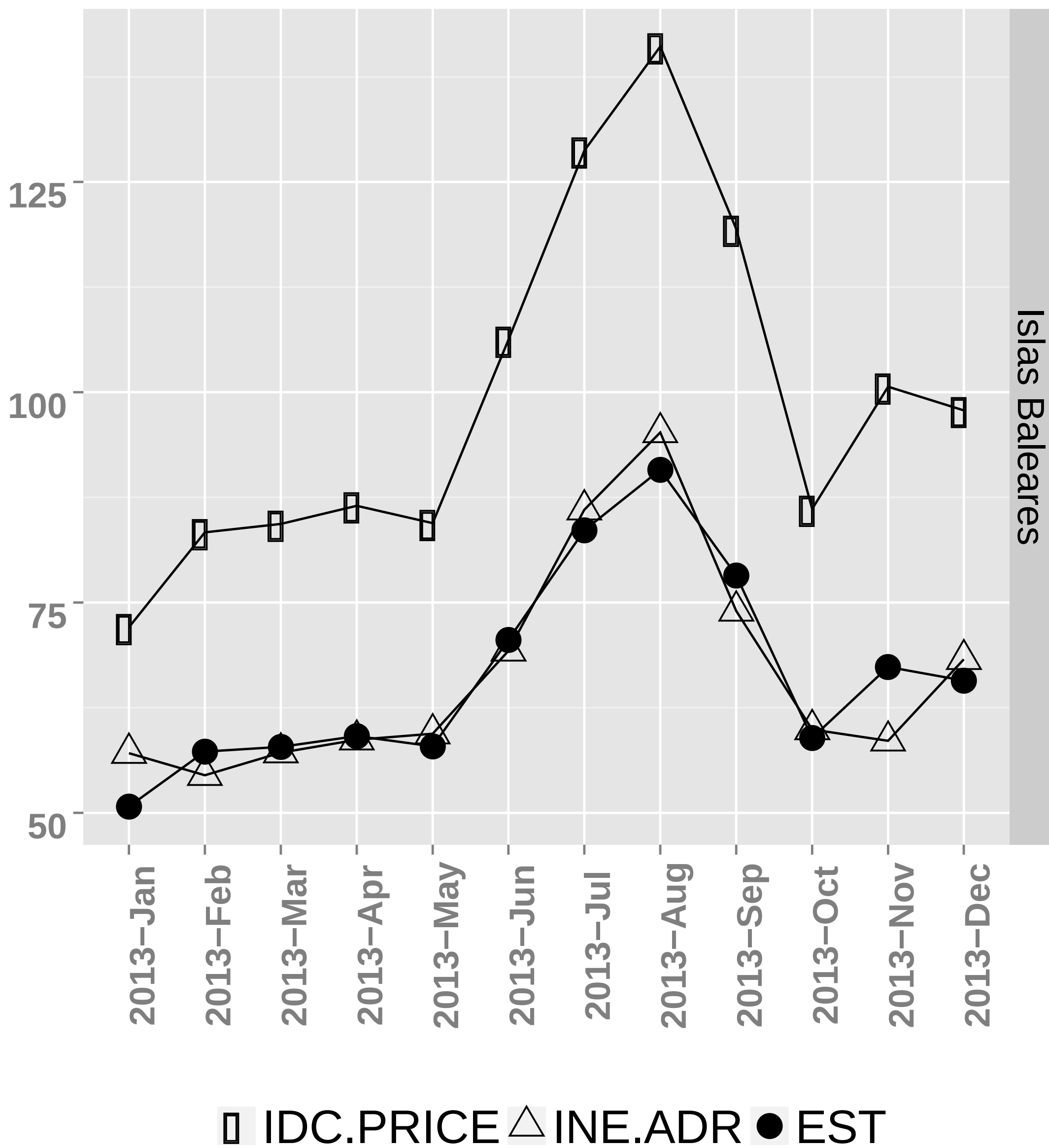


Figure ADR model fit plot for Islas Baleares for 2013 with VCC-free data

Table ADR predictions for Islas Baleares for 2014 with raw data and fixed model

AC	Month	IDC.PRICE	INE.ADR	ADR.EST	Residual
Islas Baleares	2014-Jan	81.66	60.51	55.99	4.52
Islas Baleares	2014-Feb	90.54	55.99	58.91	-2.92
Islas Baleares	2014-Mar	96.63	56.47	60.91	-4.44
Islas Baleares	2014-Apr	106.36	61.09	64.10	-3.01
Islas Baleares	2014-May	97.82	60.94	61.30	-0.36
Islas Baleares	2014-Jun	122.73	72.63	69.47	3.16
Islas Baleares	2014-Jul	142.89	89.66	76.09	13.57
Islas Baleares	2014-Aug	166.32	99.49	83.77	15.72
Islas Baleares	2014-Sep	138.46	81.91	74.63	7.28
Islas Baleares	2014-Oct	103.54	62.81	63.17	-0.36
Islas Baleares	2014-Nov	116.54	66.07	67.44	-1.37
Islas Baleares	2014-Dec	102.72	77.42	62.91	14.51

Table ADR predictions for Islas Baleares for 2014 with VCC-free data and fixed model

AC	Month	IDC.PRICE	INE.ADR	ADR.EST	Residual
Islas Baleares	2014-Jan	77.64	60.51	53.97	6.54
Islas Baleares	2014-Feb	87.25	55.99	59.55	-3.56
Islas Baleares	2014-Mar	93.86	56.47	63.38	-6.91
Islas Baleares	2014-Apr	101.52	61.09	67.83	-6.74
Islas Baleares	2014-May	96.24	60.94	64.76	-3.82
Islas Baleares	2014-Jun	119.08	72.63	78.02	-5.39
Islas Baleares	2014-Jul	140.95	89.66	90.71	-1.05
Islas Baleares	2014-Aug	162.20	99.49	103.04	-3.55
Islas Baleares	2014-Sep	134.35	81.91	86.88	-4.97

Islas Baleares	2014-Oct	97.30	62.81	65.38	-2.57
Islas Baleares	2014-Nov	105.37	66.07	70.06	-3.99
Islas Baleares	2014-Dec	102.72	77.42	68.52	8.90

Table Pearson's linear correlation coefficient for IDC.PRICE and INE.ADR for raw and VCC-free data by Spanish Autonomous Community (AC) for 2013

AC	Raw data	VCC-free data	Difference	
Andalucía	0.867	0.876	0.009	+
Aragón	0.699	0.806	0.107	+
Canarias	0.663	0.784	0.120	+
Cantabria	0.970	0.978	0.009	+
Castilla-La Mancha	0.306	0.315	0.009	+
Castilla y León	0.813	0.806	-0.008	
Cataluña	0.344	0.268	-0.076	
Comunidad de Madrid	0.974	0.977	0.003	+
Comunidad Foral de Navarra	0.955	0.971	0.016	+
Comunidad Valenciana	0.980	0.997	0.017	+
Extremadura	0.819	0.804	-0.015	
Galicia	0.982	0.980	-0.002	
Islas Baleares	0.528	0.949	0.421	+
La Rioja	0.726	0.693	-0.033	
País Vasco	0.988	0.988	0.000	+
Principado de Asturias	0.974	0.964	-0.010	
Región de Murcia	0.708	0.736	0.028	+

Table Pearson's linear correlation coefficient for IDC.PRICE and INE.ADR for raw and VCC-free data by Spanish Autonomous Community (AC) for 2014

AC	Raw data	VCC-free data	Difference	
Andalucía	0.765	0.842	0.078	+
Aragón	0.355	0.230	-0.125	
Canarias	0.719	0.720	0.001	+

Cantabria	0.768	0.979	0.211	+
Castilla-La Mancha	0.683	0.728	0.045	+
Castilla y León	0.815	0.806	-0.009	
Cataluña	-0.022	0.434	0.456	+
Comunidad de Madrid	0.929	0.928	-0.001	
Comunidad Foral de Navarra	0.944	0.925	-0.019	
Comunidad Valenciana	0.987	0.987	-0.001	
Extremadura	0.769	0.767	-0.002	
Galicia	0.448	0.964	0.517	+
Islas Baleares	0.918	0.939	0.020	+
La Rioja	0.640	0.622	-0.019	
País Vasco	0.992	0.994	0.001	+
Principado de Asturias	0.959	0.966	0.007	+
Región de Murcia	0.776	0.777	0.001	+

Table ADR models for Spanish Autonomous Communities (AC) for 2013 fitted with raw data

AC	Alpha	Beta	RMSE	R-squared	F-Test	p-value
Andalucía	-55.56	1.99	6.44	0.75	30.27	0.00
Aragón	-31.74	1.55	2.26	0.49	9.54	0.01
Canarias	60.05	0.23	3.34	0.44	7.85	0.02
Cantabria	9.16	0.84	1.84	0.94	157.69	0.00
Castilla-La Mancha	36.19	0.34	1.25	0.09	1.03	0.33
Castilla y León	14.02	0.71	0.90	0.66	19.53	0.00
Cataluña	57.70	0.27	5.23	0.12	1.34	0.27
Comunidad de Madrid	31.57	0.56	0.90	0.95	182.34	0.00
Comunidad Foral de Navarra	13.49	0.71	1.77	0.91	104.90	0.00
Comunidad Valenciana	-18.55	1.25	2.23	0.96	236.88	0.00
Extremadura	-3.82	1.08	1.42	0.67	20.33	0.00

Galicia	9.77	0.83	0.81	0.96	271.71	0.00
Islas Baleares	29.20	0.33	11.43	0.28	3.87	0.08
La Rioja	35.00	0.38	1.68	0.53	11.13	0.01
País Vasco	4.49	0.93	1.22	0.98	402.52	0.00
Principado de Asturias	3.15	0.98	1.44	0.95	184.60	0.00
Región de Murcia	-73.65	2.55	7.48	0.50	10.08	0.01

Table ADR models for Spanish Autonomous Communities (AC) for 2013 fitted with VCC-free data

AC	Alpha	Beta	RMSE	R-squared	F-Test	p-value
Andalucía	-44.92	1.87	6.24	0.77	32.97	0.00
Aragón	-47.07	1.86	1.87	0.65	18.50	0.00
Canarias	50.32	0.36	2.77	0.61	15.90	0.00
Cantabria	7.71	0.88	1.56	0.96	222.55	0.00
Castilla-La Mancha	35.13	0.36	1.24	0.10	1.10	0.32
Castilla y León	13.65	0.72	0.92	0.65	18.48	0.00
Cataluña	66.27	0.18	5.37	0.07	0.77	0.40
Comunidad de Madrid	30.91	0.57	0.84	0.95	206.36	0.00
Comunidad Foral de Navarra	-2.49	0.97	1.43	0.94	166.12	0.00
Comunidad Valenciana	-24.37	1.38	0.88	0.99	1582.84	0.00
Extremadura	-2.39	1.05	1.47	0.65	18.23	0.00
Galicia	9.90	0.83	0.85	0.96	244.35	0.00
Islas Baleares	8.92	0.58	4.25	0.90	90.19	0.00
La Rioja	28.73	0.49	1.76	0.48	9.22	0.01
País Vasco	4.53	0.94	1.21	0.98	412.78	0.00
Principado de Asturias	6.58	0.92	1.68	0.93	132.48	0.00
Región de Murcia	-80.65	2.70	7.17	0.54	11.82	0.01

Table ADR residuals for Spanish Autonomous Communities (AC) for 2014 with raw data and updated-model predictions

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Andalucía	3.3	0.7	-3.6	-13.9	-12.8	-12.3	13.9	-64.9	-23.2	-12.6	-10.6	-5.4
Aragón	3.2	-3.6	-1.6	-3.3	-3.6	-3.4	-1.5	1.2	-1.9	-0.3	-3.8	6.5
Canarias	0.7	1.0	1.6	4.3	-1.6	-3.8	3.3	3.8	-0.0	-0.6	0.6	7.7
Cantabria	-0.1	-0.4	-1.0	-3.2	-1.6	-0.6	1.3	-0.9	1.6	-0.3	-3.5	-19.3
Castilla-La Mancha	-3.0	-2.0	-3.0	2.8	-1.6	-2.1	0.5	1.4	1.7	-0.7	-2.4	-0.1
Castilla y León	-0.8	-0.1	-1.4	-1.1	-2.4	-0.4	0.3	0.5	-2.4	-0.3	0.0	1.5
Cataluña	-1.6	19.2	-2.5	-4.4	-4.0	-5.2	2.5	8.6	-0.1	-0.8	-325.8	-0.6
Madrid	-0.8	-1.3	-0.7	-0.6	-0.4	-0.3	0.3	-0.5	-7.2	0.2	0.3	-5.7
Navarra	-0.4	-2.1	-3.5	-2.5	-4.2	-3.8	4.7	-2.5	-3.4	-2.4	0.4	-2.9
C.Valenciana	3.1	-0.4	-3.2	-0.8	-1.4	-0.7	0.4	0.7	-0.9	-0.5	-1.2	-4.4
Extremadura	0.7	1.2	-2.5	-2.8	-4.3	-0.5	-1.0	0.8	-2.2	-1.4	-0.9	-2.4
Galicia	-1.1	-0.7	-0.3	-1.3	-2.1	-1.2	0.8	-1.7	-0.1	-4.4	-23.2	-2.7
Islas Baleares	4.5	-3.8	-4.7	-2.8	-0.0	3.5	13.5	11.6	2.9	-0.5	-3.2	14.6
La Rioja	-0.2	-0.0	0.2	2.2	-1.9	0.0	-1.4	0.1	-1.2	0.4	0.9	4.3
País Vasco	-0.6	-0.7	-2.5	-1.2	-2.6	-1.2	-1.3	0.2	-2.3	0.9	0.5	1.2
Asturias	1.0	-1.1	-1.7	-0.4	-0.8	-0.4	-3.5	-1.5	-1.7	-2.8	-4.8	3.0
Murcia	-2.7	0.1	-8.5	-7.1	-8.5	-2.8	3.0	1.8	2.3	-5.7	-1.3	20.8

Table ADR residuals for Spanish Autonomous Communities (AC) for 2014 with VCC-free data and updated-model predictions

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Andalucía	5.0	1.5	-4.6	-13.2	-11.2	1.1	11.2	9.3	-9.3	-7.3	-0.3	-13.1
Aragón	4.9	-0.3	-3.4	-5.3	-4.2	-4.0	-2.1	1.0	-2.3	-0.7	-3.5	5.4
Canarias	-1.4	-0.7	-0.5	3.0	-1.5	-3.7	3.0	3.2	-0.4	-1.7	-1.6	7.4
Cantabria	0.9	1.9	-1.5	-3.7	-2.2	-0.6	0.6	-0.4	1.2	0.9	-2.2	0.8

Castilla-La Mancha	-2.7	-1.4	-3.0	1.2	-2.3	-2.5	0.3	1.1	1.4	-1.0	-2.6	-0.3
Castilla y León	-0.8	-0.2	-1.4	-1.2	-2.4	-0.4	0.2	0.6	-2.0	-0.4	-0.0	1.8
Cataluña	-2.9	19.1	-3.4	-5.0	-4.0	-5.2	3.2	10.1	0.1	-1.0	2.5	0.2
Madrid	-0.9	-1.3	-0.8	-0.6	-0.4	-0.3	0.5	-0.5	-5.9	-0.0	0.6	-6.1
Navarra	-0.2	-2.6	-4.4	-4.4	-6.6	-5.0	3.3	-3.7	-4.1	-2.2	-0.0	-3.6
C.Valenciana	3.4	-0.9	-3.6	-2.6	-3.0	-2.9	-2.5	-1.8	-1.6	-0.5	-2.0	-5.1
Extremadura	0.9	1.1	-2.3	-2.8	-4.0	-0.6	-1.0	1.0	-2.1	-1.5	-1.0	-2.5
Galicia	-1.2	-0.1	-0.6	-0.8	-1.5	-0.9	0.5	-2.0	0.2	-1.2	-0.5	-4.1
Islas Baleares	6.5	-4.3	-7.2	-6.4	-3.2	-4.1	1.5	-0.9	-2.8	-1.4	-2.5	10.4
La Rioja	-0.1	-0.0	-0.2	1.7	-3.2	-0.4	-1.9	-0.5	-1.8	-0.0	0.6	4.0
País Vasco	-0.3	-0.5	-2.7	-1.4	-2.5	-0.9	-1.5	0.0	-2.2	0.7	0.2	0.7
Asturias	0.3	-1.8	-2.1	-0.9	-1.2	-0.8	-3.6	-1.0	-1.9	-3.1	-4.2	2.4
Murcia	-1.9	0.7	-8.3	-7.5	-8.6	-3.2	3.6	0.3	2.0	-5.1	-1.9	21.2

Table Percentage of prices in each segment and the corresponding official occupancy rates for the Basque Country for 2013

Month	Seg1	Seg2	Seg3	Seg 4	NoShow	INE
2013-Jan	25.74	51.02	6.58	0.07	16.59	34.78
2013-Feb	22.88	54.58	4.63	0.26	17.65	40.25
2013-Mar	13.38	56.52	10.99	0.43	18.66	49.76
2013-Apr	13.17	66.64	7.94	0.13	12.11	49.61
2013-May	6.45	66.18	10.21	0.30	16.86	56.49
2013-Jun	6.55	61.16	12.79	0.43	19.07	56.68
2013-Jul	3.16	36.61	28.97	0.75	30.51	66.92
2013-Aug	3.68	32.08	35.18	0.80	28.27	75.62
2013-Sep	3.38	51.85	26.22	0.67	17.88	66.38
2013-Oct	6.43	65.18	13.35	0.42	14.62	56.99
2013-Nov	16.43	58.03	4.93	0.18	20.43	48.13
2013-Dec	24.58	46.55	4.13	0.18	24.56	38.84

Table Linear correlation coefficients for the variables

	Seg1	Seg2	Seg3	Seg4	NoShow	INE
Seg1	1.0000	0.0899	-0.7749	-0.8060	-0.2249	-0.9403
Seg2	0.0899	1.0000	-0.6590	-0.5913	-0.8906	-0.3795
Seg3	-0.7749	-0.6590	1.0000	0.9457	0.5957	0.9175
Seg4	-0.8060	-0.5913	0.9457	1.0000	0.6082	0.9108
NoShow	-0.2249	-0.8906	0.5957	0.6082	1.0000	0.4442
INE	-0.9403	-0.3795	0.9175	0.9108	0.4442	1.0000

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

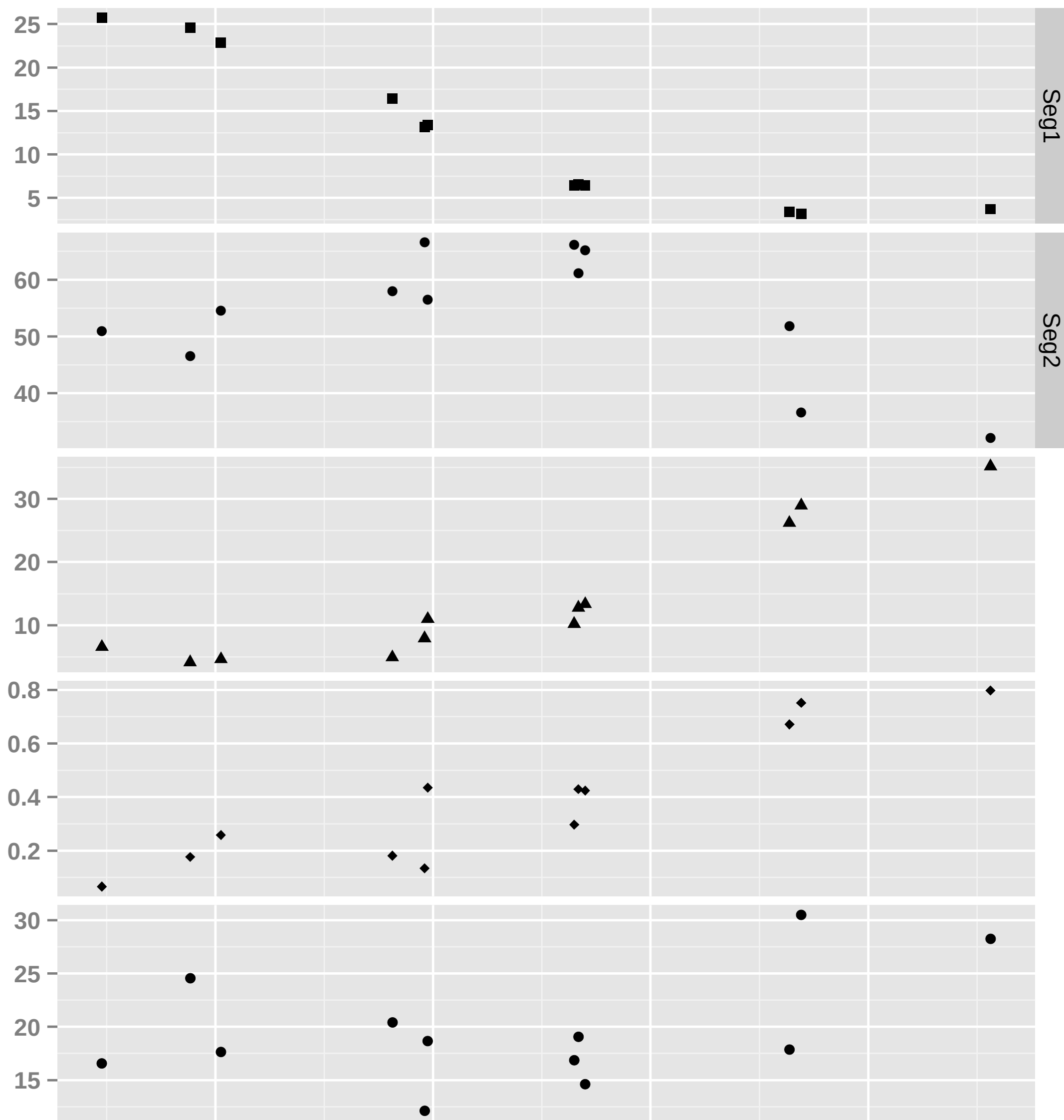


Figure Independent and dependent variable pair plots

Table Fitted linear regression model for occupancy rate estimation

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	55.6304	3.4695	16.03	0.0000
Seg1	-0.8243	0.1406	-5.86	0.0004
Seg3	0.5436	0.2087	2.61	0.0314
Seg4	0.6254	9.4754	0.07	0.9490

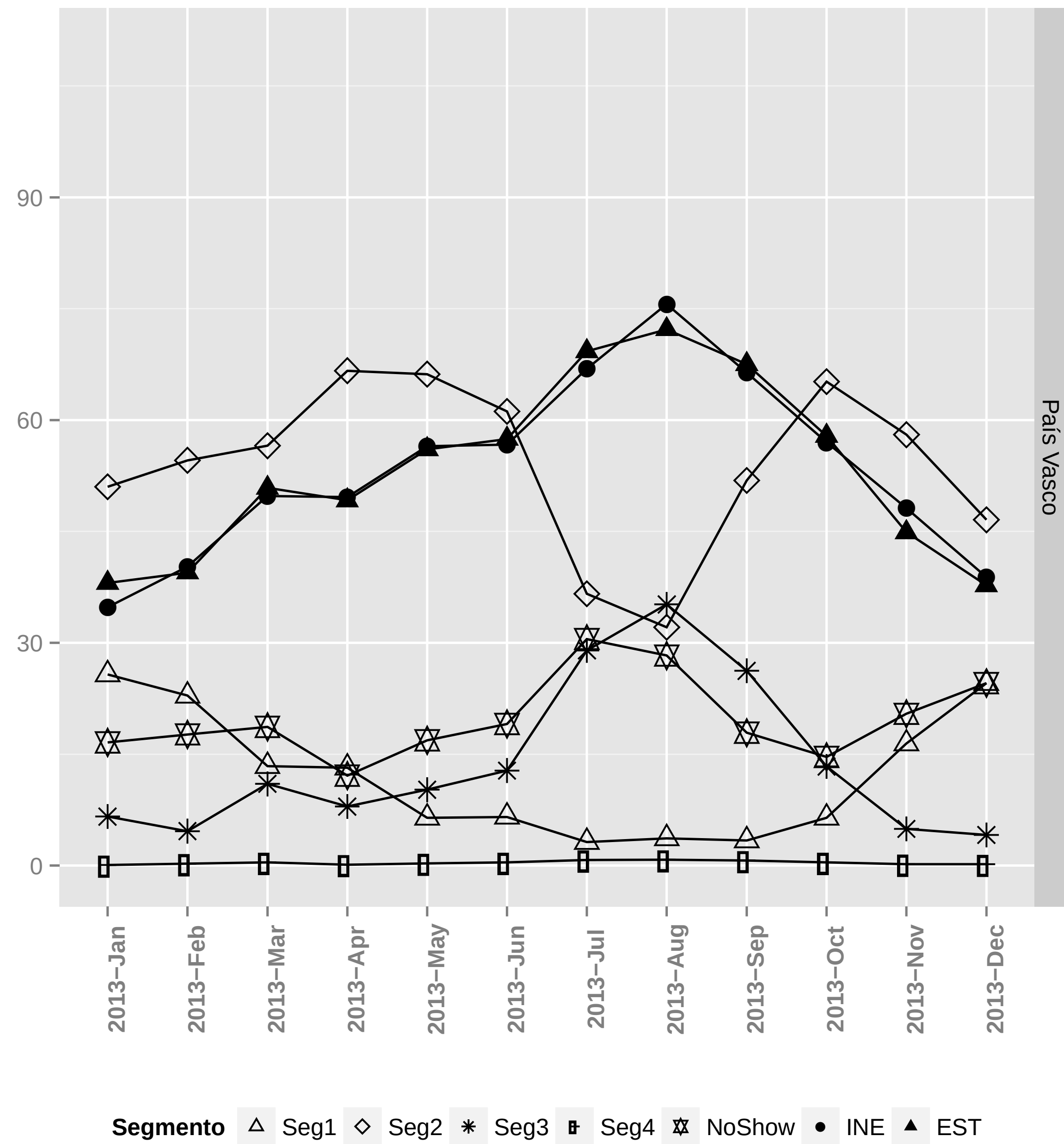


Figure Occupancy rate linear regression model fit

Table Linear regression model estimates for occupancy rate and residuals for the Basque Country for 2013

Month	INE	Mod.2 est.	Mod.2 res.
2013-Jan	34.78	38.04	-3.26
2013-Feb	40.25	39.45	0.80
2013-Mar	49.76	50.85	-1.09
2013-Apr	49.61	49.18	0.43
2013-May	56.49	56.05	0.44
2013-Jun	56.68	57.45	-0.77
2013-Jul	66.92	69.24	-2.32
2013-Aug	75.62	72.22	3.40
2013-Sep	66.38	67.52	-1.14
2013-Oct	56.99	57.85	-0.86
2013-Nov	48.13	44.88	3.25
2013-Dec	38.84	37.72	1.12

Table Independent variables and model 2 predictions for the Basque Country for 2014

Month	Seg1	Seg2	Seg3	Seg4	NoShow	INE	Est.	Res.
2014-Jan	26.70	39.77	6.89	0.70	25.95	34.93	37.80	-2.87
2014-Feb	29.99	38.20	8.66	1.65	21.50	38.89	36.65	2.24
2014-Mar	25.46	32.35	19.28	4.45	18.46	42.99	47.91	-4.92
2014-Apr	17.75	30.39	23.74	5.58	22.54	50.64	57.40	-6.76
2014-May	11.34	31.22	27.56	6.42	23.46	59.19	65.28	-6.09
2014-Jun	9.29	29.28	28.90	7.78	24.75	61.58	68.55	-6.97
2014-Jul	6.82	14.05	34.55	13.12	31.45	67.97	77.00	-9.03
2014-Aug	5.02	10.39	30.85	13.00	40.74	76.89	76.39	0.50
2014-Sep	5.96	18.28	34.00	10.14	31.62	70.75	75.54	-4.79
2014-Oct	10.01	27.08	23.59	5.35	33.98	61.73	63.55	-1.82

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

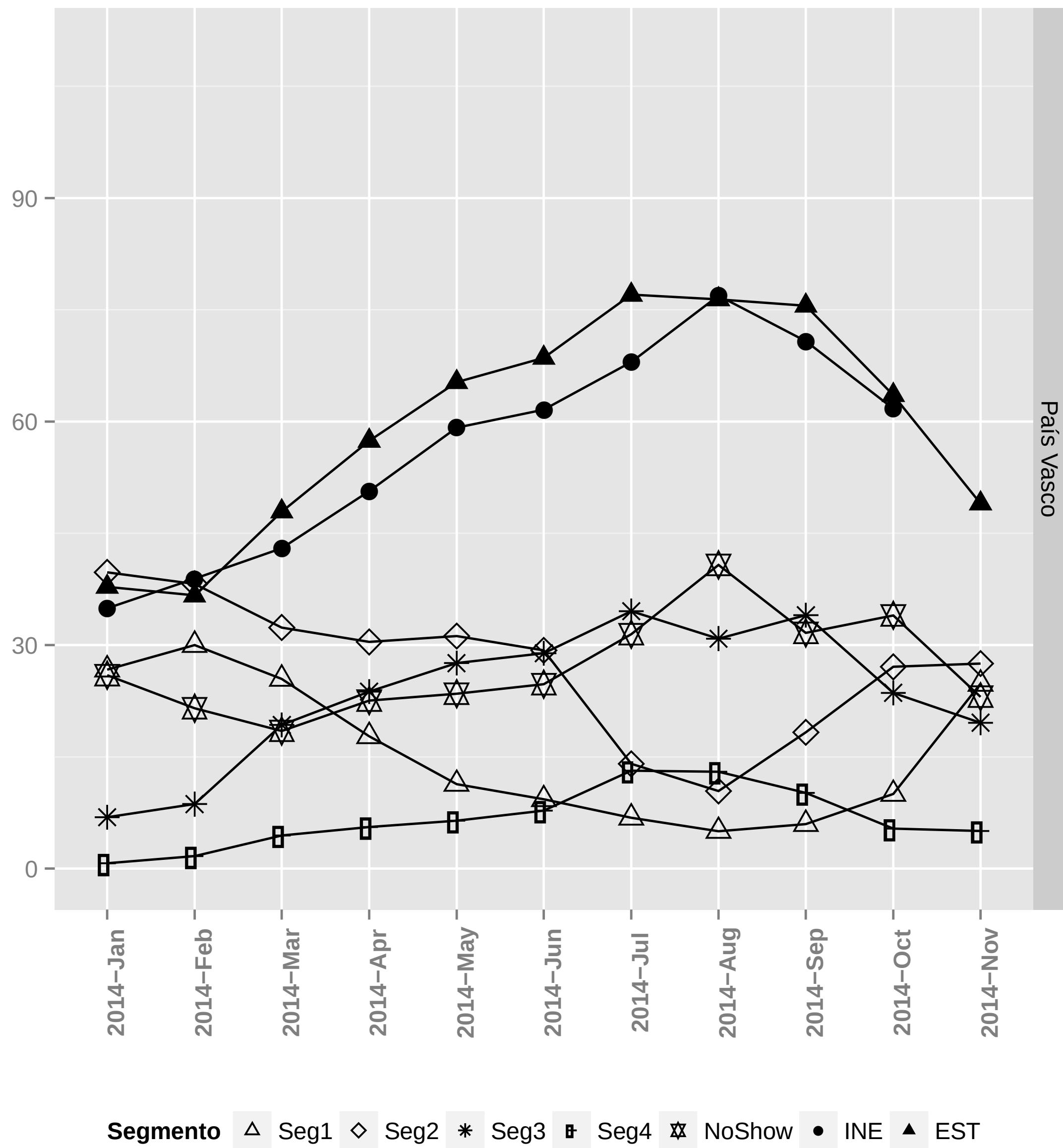


Figure Independent variables and model 2 predictions for the Basque Country for 2014

Table Linear regression models fitted and their assessment statistics for Spanish Autonomous Communities (AC) for 2013

AC	Intercept	B1	B2	B3	RMSE	Adj.R-sq.	FStat.	p-val.
La Rioja	25.78	-0.13	1.36	3.13	2.57	0.93	46.52	0.00
Navarra	28.34	-0.98	1.91	-7.98	3.07	0.89	30.50	0.00
Murcia	40.57	-1.24	0.88	7.01	4.29	0.64	7.40	0.01
Extremadura	29.81	-0.63	0.77	-6.15	2.63	0.78	13.98	0.00
Canarias	80.55	-0.31	0.00	1.57	4.89	0.22	2.02	0.19
Cantabria	49.23	-1.96	0.27	8.47	1.95	0.98	204.38	0.00
País Vasco	55.63	-0.82	0.54	0.63	2.35	0.96	97.82	0.00
Castilla-La Mancha	33.52	-0.70	-0.18	4.60	3.28	0.33	2.85	0.11
Aragón	38.09	-1.85	0.81	3.79	3.03	0.67	8.37	0.01
Asturias	34.61	-1.12	0.47	10.00	2.57	0.96	90.36	0.00
Islas Baleares	86.68	-1.69	0.32	-6.07	4.37	0.94	90.36	0.00
Castilla y León	24.42	-0.55	1.25	-3.37	1.86	0.92	44.15	0.00
Galicia	35.82	-0.93	0.13	14.10	1.92	0.96	90.66	0.00
Madrid	42.15	0.03	0.59	3.60	5.12	0.44	3.83	0.06
C.Valenciana	62.96	-1.31	0.46	1.10	2.17	0.97	105.27	0.00
Cataluña	52.59	-0.71	0.90	3.17	1.98	0.98	173.68	0.00
Andalucía	29.43	-0.22	1.91	-10.91	4.84	0.84	20.21	0.00

Table Predictions of the fitted linear regression models for Spanish Autonomous Communities (AC) for 2014

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
La Rioja	32.7	43.1	58.4	70.1	80.0	80.6	87.8	95.2	86.1	70.2
Navarra	22.2	25.3	48.4	43.1	37.1	35.3	-27.0	-2.5	18.4	23.6
Murcia	42.9	60.7	88.6	94.1	92.4	96.8	109.1	111.7	95.4	76.2
Extremadura	22.7	23.2	28.8	13.4	19.7	25.5	22.3	19.9	26.3	22.5
Canarias	83.2	87.3	93.0	83.2	76.3	75.8	77.3	80.6	81.4	83.0
Cantabria	16.5	7.9	27.9	55.4	48.6	63.0	130.4	161.0	92.9	49.4
País Vasco	37.8	36.6	47.9	57.4	65.3	68.5	77.0	76.4	75.5	63.6
Castilla-La Mancha	30.8	33.9	47.7	56.0	61.7	60.1	55.6	57.5	66.4	57.6
Aragón	29.6	39.7	52.8	59.3	53.5	51.9	55.2	63.1	57.5	56.2
Asturias	23.1	20.9	32.2	50.5	46.9	46.2	96.8	135.1	60.9	42.3
Islas Baleares	33.0	31.4	41.8	48.6	49.2	55.6	38.2	29.7	50.1	54.5
Castilla y León	20.5	22.4	33.8	38.2	46.3	43.6	44.3	39.2	39.1	37.7
Galicia	23.1	29.0	54.5	75.1	77.7	81.7	171.5	221.8	119.2	92.7

Madrid	52.7	54.6	58.2	61.8	65.2	61.4	53.9	48.9	80.3	70.6
C.Valenciana	34.3	42.4	50.8	57.8	58.7	66.5	75.7	78.0	69.6	60.3
Cataluña	39.5	51.9	63.3	71.7	77.9	81.5	91.6	95.1	86.1	72.1
Andalucía	30.7	31.2	38.7	39.3	39.6	50.2	27.1	-11.8	41.4	48.1

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Table Residuals for the predictions of the fitted linear regression models for Spanish Autonomous Communities (AC) for 2014

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
La Rioja	0.3	-2.3	-9.2	-16.1	-23.8	-28.8	-38.9	-35.2	-21.5	-14.2
Navarra	-0.2	4.2	-12.2	-0.6	12.1	10.1	73.0	59.8	33.8	21.2
Murcia	-10.5	-18.1	-43.5	-46.7	-46.0	-49.7	-55.5	-53.0	-40.6	-24.3
Extremadura	-4.2	-0.7	-0.9	25.0	18.1	7.8	9.8	15.2	9.5	12.1
Canarias	3.0	-1.0	-7.9	-9.5	-7.2	-1.1	1.0	3.4	0.6	-1.7
Cantabria	5.8	20.0	3.2	-18.5	-12.0	-21.5	-76.7	-91.0	-42.1	-5.0
País Vasco	-2.9	2.2	-4.9	-6.8	-6.1	-7.0	-9.0	0.5	-4.8	-1.8
Castilla-La Mancha	-11.4	-9.8	-21.5	-25.1	-29.0	-29.6	-27.4	-24.8	-32.2	-24.1
Aragón	-3.2	-6.0	-22.6	-27.8	-18.6	-18.0	-19.7	-18.0	-20.1	-20.4
Asturias	-1.2	3.1	-6.1	-16.6	-14.1	-10.2	-48.6	-67.7	-11.6	-9.1
Islas Baleares	2.3	15.6	15.9	18.3	14.0	26.5	48.2	62.8	37.0	6.1
Castilla y León	0.0	2.8	-5.3	-3.0	-9.7	-6.8	-5.3	8.3	5.2	2.6
Galicia	-2.5	-5.0	-29.5	-45.0	-43.3	-45.2	-129.5	-165.5	-76.6	-55.5
Madrid	-4.1	2.2	3.9	0.7	4.3	4.8	5.8	3.4	-7.0	1.9
C.Valenciana	3.1	5.6	0.6	0.7	0.3	-0.6	-6.8	0.5	1.4	0.8
Cataluña	-2.3	-3.7	-11.1	-14.9	-19.5	-16.5	-20.1	-13.1	-14.3	-8.8
Andalucía	3.2	11.0	7.5	15.5	16.4	9.0	35.4	85.1	27.2	11.3

Table Predictions of the updated linear regression models for Spanish Autonomous Communities (AC) for 2014

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
La Rioja	32.7	40.5	52.4	58.7	62.1	59.8	58.5	68.1	60.8	55.8
Navarra	22.2	24.7	55.1	44.7	53.8	53.4	45.8	47.7	49.1	37.8
Murcia	42.9	54.8	58.4	56.2	52.5	55.5	69.9	63.5	53.3	38.9
Extremadura	22.7	23.1	36.9	32.7	38.5	34.7	32.3	31.9	34.2	34.8
Canarias	83.2	86.6	90.1	82.8	76.9	76.3	80.2	80.3	80.0	80.8
Cantabria	16.5	9.6	28.6	49.8	38.4	46.9	72.6	88.3	51.8	38.5
País Vasco	37.8	43.5	57.3	72.4	63.3	84.2	113.5	90.4	77.1	60.5
Castilla-La Mancha	30.8	27.9	32.9	41.5	38.6	32.3	28.9	31.3	36.5	32.0
Aragón	29.6	36.4	46.8	55.4	41.5	34.8	48.8	56.2	36.8	36.0
Asturias	23.1	26.0	35.7	48.3	34.7	39.6	80.2	101.6	43.9	30.6
Islas Baleares	33.0	31.5	49.8	111.5	74.2	151.1	246.1	253.5	97.6	68.0
Castilla y León	20.5	25.2	38.0	39.6	42.6	36.1	36.0	38.4	36.3	38.1
Galicia	23.1	27.8	22.8	50.2	40.3	38.8	100.3	119.6	50.6	48.6

Tourism Economics Fast Track, DOI: http://dx.doi.org/10.5367/te.2015.0491

Madrid	52.7	54.6	57.8	61.2	64.7	60.7	55.0	52.6	82.8	68.6
Comunidad Valenciana	34.3	43.2	50.9	56.7	58.1	66.9	76.6	75.1	68.7	60.2
Cataluña	39.5	51.2	59.0	68.4	72.5	75.7	89.0	96.6	74.2	62.3

Table Residuals for the predictions of the updated linear regression models for Spanish Autonomous Communities (AC) for 2014

AC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
La Rioja	0.3	0.3	-3.2	-4.8	-5.9	-8.0	-9.6	-8.1	3.8	0.2
Navarra	-0.2	4.9	-18.9	-2.3	-4.6	-8.0	0.2	9.6	3.2	7.0
Murcia	-10.5	-12.2	-13.3	-8.8	-6.1	-8.4	-16.3	-4.8	1.5	13.0
Extremadura	-4.2	-0.6	-9.0	5.7	-0.8	-1.4	-0.1	3.1	1.6	-0.2
Canarias	3.0	-0.2	-5.0	-9.2	-7.8	-1.6	-2.0	3.7	2.0	0.5
Cantabria	5.8	18.3	2.4	-13.0	-1.9	-5.4	-18.9	-18.3	-1.0	5.9
País Vasco	-2.9	-4.6	-14.3	-21.7	-4.1	-22.6	-45.5	-13.5	-6.3	1.2
Castilla-La Mancha	-11.4	-3.8	-6.7	-10.6	-5.8	-1.9	-0.7	1.3	-2.2	1.5
Aragón	-3.2	-2.6	-16.5	-24.0	-6.6	-1.0	-13.3	-11.2	0.6	-0.2
Asturias	-1.2	-2.0	-9.7	-14.4	-1.9	-3.5	-32.0	-34.1	5.4	2.6
Islas Baleares	2.3	15.5	7.9	-44.7	-11.0	-69.0	-159.8	-161.0	-10.5	-7.4
Castilla y León	0.0	0.0	-9.5	-4.5	-6.1	0.7	3.0	9.0	8.0	2.2
Galicia	-2.5	-3.8	2.1	-20.1	-5.8	-2.3	-58.3	-63.3	-7.9	-11.4
Madrid	-4.1	2.2	4.3	1.3	4.7	5.5	4.7	-0.3	-9.5	3.9

Comunidad Valenciana	3.1	4.9	0.5	1.8	1.0	-1.0	-7.8	3.3	2.2	0.9
Cataluña	-2.3	-3.1	-6.7	-11.5	-14.0	-10.7	-17.5	-14.5	-2.4	1.0

Notes

1 In this paper, the term 'destination' refers to NUTS2 regions as defined in the Nomenclature of Units for Territorial Statistics.