

Forecasting of taxi times: the case of the Barcelona-El Prat airport

Oriol Lordan¹, Jose M Sallan¹, Marta Valenzuela-Arroyo¹

¹ Universitat Politècnica de Catalunya - BarcelonaTech. ETSEIAT, C/ Colom 11. 08222 Terrassa

1. Abstract

This study computes linear regression models to predict taxi in and out times at Barcelona-El Prat Airport using R software. Taxi time results are given for each stand as a function of the hour of the day. Also, it is analysed how many variables are influencing the time with their relevancy degree. The study has been carried out using airport's daily operational data for the months of June, July and August 2013.

2. Introduction

Taxi time is the most critical parameter when planning ground operations. A bad prediction of taxi times may have consequences both for airport operations and for airlines. From the point of view of the airport, a poorly computed taxi-time can drive to delays, schedules modifications and a low exploitation of airport facilities. Regarding to airlines, taxi time is directly proportional to aircraft fuel consumption and loses in efficiency, becoming more expensive the operation.

Nowadays, Barcelona-El Prat Airport is working with taxi times related to a group of stands but not with each stand directly. This is not desirable in order to avoid the inefficiencies just mentioned.

So, the aim of this study is to determine the taxi in and out time for each stand, classify this time as a function of the hour of the day and detect how many variables, called predictors, are affecting it with its relevancy degree.

In order to compute taxi times it has been used Linear Regression Models by implementing airport's daily operational data in R software. The methodology used includes four differentiated parts. First, at the airport facilities, it has been used operational software to inform when an aircraft was at determined positions. Then it has been filtered the information as not all of it was correct. After that, it has been computed taxi times and created binary variables to be introduced finally in R software which using the function BIGLM has computed taxi in and out models.

3. Theoretical Framework

3.1. Statistics

The basic statistical tool that this study is going to use is the Linear Regression. The aim of a linear regression is to determine the relationships between:

- A dependent variable y
- A set of independent variables called predictors x_1, x_2, \dots, x_p

In order to represent these relationships, it is used a Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

Therefore, the aim of this study is to find the estimators of the predictors ($\beta_0, \beta_1, \varepsilon_i \dots$). The most used estimators are called Least Squares. This method means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation. In other words:

$$\text{MIN} \sum_{i=1}^n e_i^2 \quad (2)$$

In a linear regression model it is important to determine the **adjusted coefficient of determination**:

$$R_{adj}^2 = 1 - \frac{\frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}}{\frac{\sum (y_i - \bar{y})^2}{n - 1}} = \frac{(n - 1)R^2 - p}{n - p - 1} \quad (3)$$

This coefficient determines the quality of the prediction taking into account the variables and the observations. A good prediction will take a value near 1 while a bad prediction will take numbers close to 0.

In order to introduce the variables in the model it has been used binary variables. With them, it can be evaluated the influence of the different observations classified over a group. For example, it can be evaluated the influence of each aircraft that is found on the airport.

It is important to remark that in this study exists a great possibility of having collinearity, which is a relationship between two variables. This is not desirable as a strong correlation leads to low parameter estimation precision provided that there are not enough information. In order to detect collinearity in least squares it is used the **Variance Inflation Factor (VIF)**. VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of this problem. As an standard it is accepted a predictor whose VIF test obtains a value lower than 10.

3.2. Methodology

First thing to do for obtaining the necessary data consists in setting up the operational software of the airport. Thus, it is created the controlled areas. For the case of the taxi-in at the exits of the runways and for the taxi-out at the runway headers.

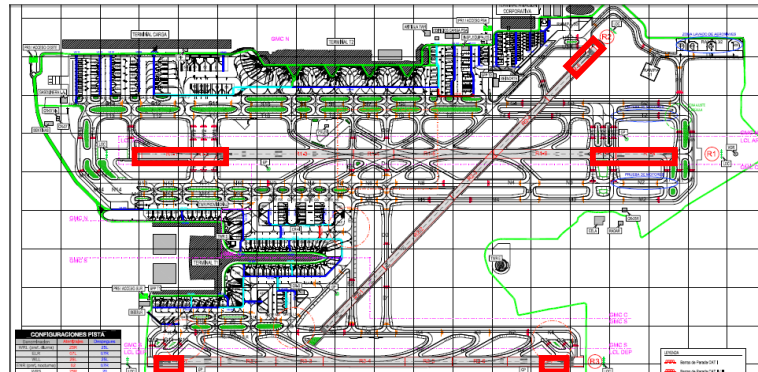


Figure 1. Areas for the runway headers at Barcelona-El Prat Airport

The received information. Then it is removed information which is not correct, for example, duplicated notifications of departures. After that, when all information is filtered it is created the binary variables and it is computed the linear regression model for the taxi time.

3.3. Predictors

The predictors that this study will take into account because they have affection on the taxi time are:

Predictor	Definition	Why does it have influence on taxi time?
Stand Number (S)	Variable that represents the different stands of the airport	The taxi time depends on the position of the stand as it is the start or the end of the movement
Arrival (A)	Variable that represents the rapid exits	In taxi-in time, it is the start point of the taxi time calculation
Departs (D)	Variable that represents the runway that the aircraft uses for departing	In taxi-out time, it is the final point for the taxi-out calculation
Stand/Arrival (SA)	Variable that represents a taxi-in movement	This binary variable is the union of a Stand and an Arrival. Therefore, it represents the whole taxi-in movement
Stand/Departure (DS)	Variable that represents a taxi-out movement	This binary variable is the union of a Stand and a Departure. Therefore, it represents the whole taxi-out movement

Time of the day	Variable to take into account the hour in slots of 3hours	Because not all hours are equally congested. The time is divided into periods of 3h
Runway configuration	Variable for the different runway configurations	Depending on the runway configuration, the distance travelled by the aircrafts is increased or decreased
Arrival/Departure	Variable for differentiating an arrival and a departure flight	Depending on the type of operation, the distance travelled by the aircrafts is increased or decreased
Airline	Variable for the airlines that operate at the airport	Airlines can influence time because aircrafts taxiing velocity are mainly decided by airlines
Aircraft	Variable for the different aircrafts that operate the airport	The different characteristics of the airplanes can influence on the time: capacity of manoeuvre, optimum speed for taxiing,...
Terminal	Variable to discern between the two terminals	Depending on the position of the terminal, the distance travelled by the aircrafts is increased or decreased

Table 1. List of predictors

The following graphic shows the relevancy of the predictors when used each one of them to build the model:

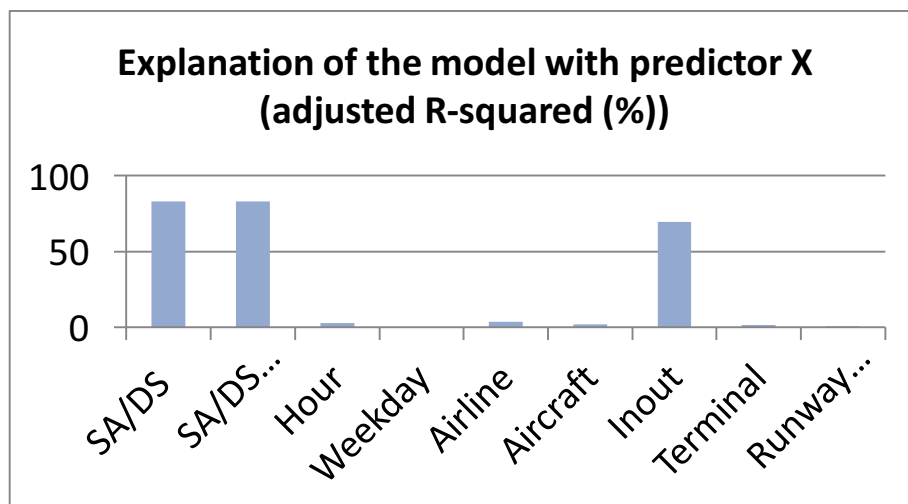


Figure 2. Predictor's influence

As can be seen, only the areas and the arrival/departure predictors have real relevance on taxi times. The reason to be of the areas importance is logical as it is the start and end point of taxi times. Regarding to the influence of the arrival/departure indicator, it comes to say that taxi time is high dependant on if it is an arrival or a departure

operation. For the rest of the predictors, none of them achieves and 4% explanation of the taxi time.

From this graphic it can be obtained another important conclusion. SA and DS predictors explain better the model than using the areas separately. It is better from all comparisons: the adjusted coefficient of determination, the mean error of the difference between the real value and the predicted value, the negative cases found and the % of error at 75, 85 and 95 percent of the cases.

However, due to collinearity not all the predictors can be taken into account into the model. The explanation of these collinearities is summarized on the following points:

- **Arrival/Departure predictor.** It is a linear combination of the Areas variable. Every time that it is had an S with an A it is added a 0 and, for the opposite case, with S and D it is added a 1. So the sum of elements of SA is the same than the sum of 0 as well as that the sum of SD is equal to 1. Therefore, the R software interprets that the variable is a combination and does not evaluate the influence of the indicator. This fact can be taken into account to make a different approach to the model. At the start, the idea was having a model in which with an indicator could discriminate a departure and an arrival flight. If this can't be done maybe it is better to use **two models** separately: one for taxi-in and one for taxi-out.
- **Runway configuration predictor.** The reason to be of this collinearity resides in the fact that there is a runway configuration preference for diurnal and nocturnal operations. Therefore it is highly related with the hour of the day.
- **Airline predictor.** The airline collinearity also has a reason to be as in airlines operation normally the stands are always the same, in other words, there is a connection between stand and airline so high that the coefficient of airline cannot be evaluated.
- **Aircraft predictor.** This predictor is multicollinear with other variables. It has a mixed influence with stand, airline and other variables.
- **Terminal predictor.** It is formed after the stand number, therefore, it is so related with it that the software interprets that they are a combination.

The weekday predictor was also discarded to be used in the model even though in this case there was not detected collinearity. The difference between using the weekday predictor or not did not exceed a 0.7% improvement to the model. Therefore, it is recommendable not to use variables on the model that are not useful.

Finally, it is had a model that uses the areas and the hour of the day. Unfortunately, when it was computed the VIF test with these variables it was obtained an unwanted result: the hour variable is highly correlated with the areas.

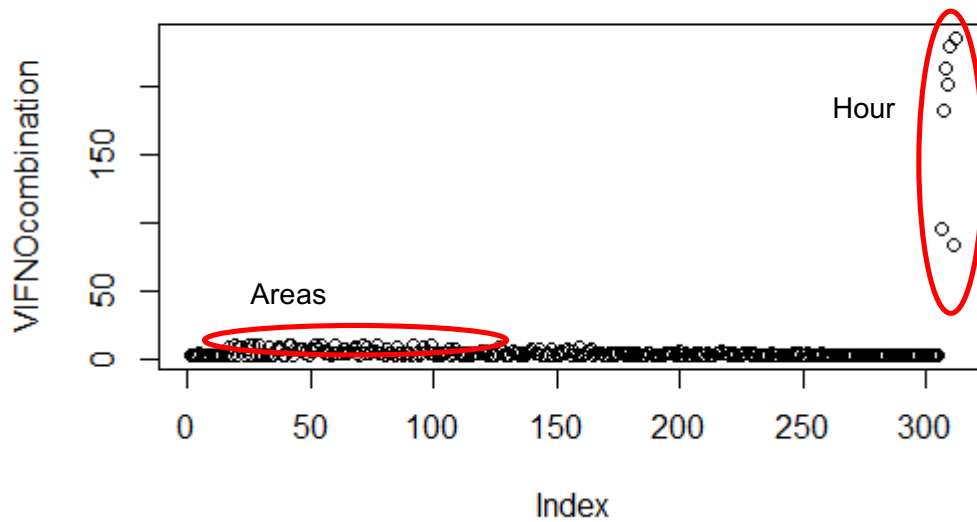


Figure 3. VIF Test on areas and hour of the day predictors

In order to fix this result, it has been created a new variable that combines both the area with the hour of operation.

4. Results

The final models built present the following characteristics:

	Taxi-in	Taxi-out
$R_{adjusted}^2$ (%)	76.2	44.93
Mean absolute error of residuals (min)	0.475	2.157657
Error at 75% of cases (min)	0.388	1.866761
Error at 85% of cases (min)	0.583	2.703804
Error at 95% of cases (min)	0.962	4.105564

Table 2. Final model characteristics

Starting with the taxi-in model, the results can be considered as excellent. The model predicts up to 76% of the taxi time and the mean error committed between the real value and the predicted value is only of 0.475 min, which is less than 30 seconds. Moreover in a 95% of the cases the error committed is lower than 1 min.

The results for the taxi-out case are not as satisfactory. The model predicts a 45% of the taxi time and the mean error is placed at 2.16 min. It is worth mentioning that in a 75% of the cases the error is lower than 2 min.

5. Conclusions

This study had two aims. On one hand, to obtain linear regression models to compute taxi times for arrivals and departures as a function of the hour of the day at Barcelona-El Prat Airport and on the other hand, to analyse the influence of all possible predictors that affect the time.

Starting with the influence of predictors, it is detected that only the areas predictor already explains the 82.89% of the taxi time. Then, a combination of the areas and the hour predictor, needed due to collinearity, increases this percentage to an 83.09%. From the rest of predictors only the indicator of arrival/departure obtains great influence on the model. This influence is already gathered as it is had one model for departures and one model for arrivals. All the other predictors only help explaining maximum a 3.63% of taxi time. In addition, they cannot be added to the model as they present collinearity with the areas/hour combinative predictor. For example, airlines usually operate at the same stands and runway configuration is high dependant on the hour or the day, as it has preferable configurations during diurnal and nocturnal operations.

Regarding to the linear regression models built to compute taxi times, the results vary depending on the type of operation. The model for taxi-in operations can be qualified as excellent as it has an adjusted coefficient of determination of 76% and only 0.48 min of mean absolute error on residuals. The taxi-out model is not as efficient provided that its coefficient is of 45% and its mean error of 2.16 min. The reason to be of these results on taxi-out times is the time aircrafts spend on holding areas and later on the runway headers. Both waiting include a factor of uncertainty to the model which makes vary taxi-out time results. Solutions to this problem are discussed in the following section of future lines.

5.1. Future Lines

It can be contemplated two options that could obtain some improvements in the taxi-out time model.

The first option would consist in measure both the waiting time of aircrafts in holding areas and in runway headers. It could be used some additional control areas at the airport facilities for this purpose. Then, the model for taxi-out times would be divided into three: a model for the time that the aircraft is moving, a model for determining the time the aircrafts are on holding areas and a time for determining the time aircrafts spend on the header of the runways. The final taxi-out time would be the sum of the three predicted values.

Another option would be changing the area positioning of the end of taxi-out times and place it at the entrance of each runway holding areas. Obviously, this time would not be exactly taxi-out time, but it would inform of the time of arriving to the runway to the airport authorities.