Airport Taxi Time Analysis

Group 6

POLITECNICO MILANO 1863 · JFK JOHN F. KENNEDY INTERNATIONAL AIRPORT · BOS · PHX PHOENIX SKY HARBOR INTERNATIONAL AIRPORT

0. Introduction - Data Creation

- **What is Taxi Time?**

Amount of time an aircraft spends on the surface of the airport while engine is running.

TotalTaxi = Taxi Out + Taxi In

- **Dataset creation**

Our dataset was handcrafted using different sources (4), each one containing different information we considered crucial for our "totalTaxi" analysis:

1 - Combined Flights 2019: 2019 USA domestic flights.
2 - Airport Weather 2019: USA airports weather conditions, with some corrections (ex. Phoenix).
3 - Aircraft dataset: aircrafts specs.
4 - Concurrent flights: meaning flights that arrived (arrivals) or took off (departures) from the same airport, the same day, at the same time block.

- **Airports selection**

We kept 3 origins: JFK, BOS, and PHX.
Why?
First, they have a similar domestic traffic;
Secondly, JFK has the highest TaxiTime;
Lastly, we selected one in the North and one in the South.

- **Final dataset**

125k observations, in which each "route" is one individual, with 24 features, including our new "totalTaxi".

Facts about consumption (B777)

APU	Ground	Flight
3.9Kg Fuel	23.3Kg Fuel	113.3Kg Fuel
12.3Kg CO2	73.4Kg CO2	356.9Kg CO2
per min	per min	per min

RESEARCH QUESTIONS:

- How do different airlines perform with respect to Taxi Time?
- What are the most impactful features that explain Taxi Time?
- Can we actually predict (accurately) Taxi Time?

1. Data Preparation

- Downsampling

Keep only 20k samples per airport to make our analysis faster.

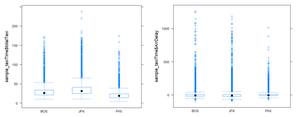
- Data conversion

Categorical variables conversion into factors and dates into datetime type.

```
$ routeID          : Factor w/ 238 levels "BOS-ATL-DL","BOS-ATL-WN",...: 23 22 10 8 53 10 6 10 26 33 ...
$ FlightDate       : Date[1:60000], format: "2019-01-21" "2019-03-28" "2019-04-10" "2019-09-26" ...
$ Origin           : Factor w/ 3 levels "BOS","JFK","PHX": 1 1 1 1 1 1 1 1 1 1 ...
$ DepTimeblk       : Factor w/ 19 levels "0001-0559","0600-0659",...: 9 4 12 10 4 2 1 14 6 15 ...
$ ArrTimeblk       : Factor w/ 19 levels "0001-0559","0600-0659",...: 10 7 14 11 7 4 3 16 9 19 ...
$ Dest             : Factor w/ 87 levels "ABQ","ANC","ATL",..: 37 36 14 11 85 14 5 14 39 45 ...
$ CRSDepTime       : num [1:60000] 1305 830 1609 1420 818 ...
$ DepDelay         : num [1:60000] 4 -2 -1 49 117 -8 -1 -3 -3 17 ...
$ Distance         : num [1:60000] 187 1010 728 369 1185 ...
$ DayOfWeek        : Factor w/ 7 levels "1","2","3","4",..: 1 4 3 4 7 3 6 5 4 7 ...
$ Operating_Airline: Factor w/ 4 levels "9E","AA","DL",..: 1 1 2 4 3 2 4 2 3 3 ...
$ ArrDelay         : num [1:60000] 12 -31 0 40 96 -30 -24 -8 -23 1 ...
$ AWND             : num [1:60000] 20.13 15.21 13.65 10.51 7.83 ...
$ PRCP             : num [1:60000] 0 0.01 0.01 0.27 0 0 0 0 0 0.03 ...
$ SNOW             : num [1:60000] 0 0 0 0 0 0 0 0 0 0 ...
$ TMAX             : num [1:60000] 10 54 47 84 85 33 83 79 27 70 ...
$ NUMBER_OF_SEATS  : num [1:60000] 76 76 160 143 160 187 143 187 180 157 ...
$ totalTaxi        : num [1:60000] 49 26 47 18 18 34 19 42 25 31 ...
$ age              : num [1:60000] 11 11 2 13 18 5 12 7 3 28 ...
$ month            : Factor w/ 12 levels "1","2","3","4",..: 1 1 3 4 9 8 11 9 9 12 4 ...
$ concurrent_TOT   : num [1:60000] 44 41 47 51 45 65 24 55 56 49 ...
```

- Outliers

taxiTime & ArrDelay outliers detected



We keep only rows s.t:
Q1-1.5*IQR<taxiTime<Q3+1.5*IQR
Q1-1.5*IQR<ArrDelay<Q3+1.5*IQR

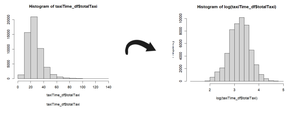
- Numerical feature scaling

To make the linear regression models more performant.

2. Feature Selection

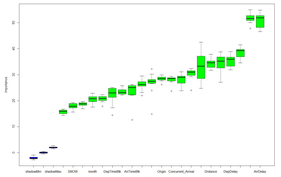
- Output log transformation

totalTaxi is left skewed -> normalization

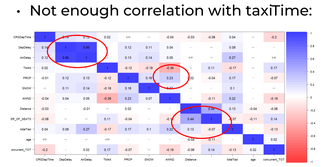


- Random forest selection

Get an overview of the most impactful features:



- Correlation plot
  - Too much interdependant
  - Not enough correlation with taxiTime:



- Backward linear selection
  - First linear regression model by including all the variables
  - Many of them are not significant enough --> backward selection to keep only the best predicates:

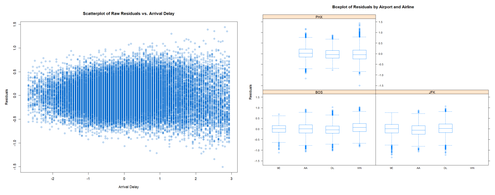
```
Coefficients:
                                      Value  Std.Error  t-value p-value
(Intercept)                       3.421770 0.004825565  709.0922       0
Operating_AirlineAA              -0.049728 0.004932575  -10.0809       0
Operating_AirlineDL              -0.127624 0.004821799  -26.4682       0
Operating_AirlineWN              -0.420559 0.003549401  -75.7846       0
ArrDelay                          0.173212 0.000361881   39.7104       0
Distance                          0.052108 0.001423907   36.5948       0
OriginJFK                         0.143768 0.003390412   40.0422       0
OriginPHX                        -0.311029 0.003448577  -90.1906       0
concurrent_TOT                    0.049317 0.001404585   35.1257       0
Operating_AirlineAA:ArrDelay     -0.034316 0.004446468   -7.7177       0
Operating_AirlineDL:ArrDelay     -0.041805 0.004113316  -10.1634       0
Operating_AirlineWN:ArrDelay     -0.076896 0.003539964  -13.8803       0
ArrDelay:OriginJFK                0.017751 0.003251152    5.4599       0
ArrDelay:OriginPHX               -0.022472 0.003631169   -6.1885       0
```

3. Variance Structure

Linear Model Formulation:

log(TaxiTime) ~ Airline + ArrDelay + Distance + OriginAirport + ConcurrentFlights + Airline : ArrDelay + OriginAirport : ArrDelay

- **Where do we see Heteroscedasticity?**



- Clear Heteroscedasticity of residuals vs. Arrival Delay.
- Possible Heteroscedasticity with respect to Airline and Origin Airport.
- Residuals vs. Other Variables appear to be Homoscedastic.

- **Variance Structure Selection:**

Step 1: Grouping Factor. Are Airlines and Airport significant grouping factors? Which one is better?

M1 : weights = varIdent(form = ~1 | Airline)
M2 : weights = varIdent(form = ~1 | OriginAirport)

MODEL	AIC	BIC	LOGLIK
LM	19786.86	19920.43	-9878.432
M1	19613.04	19773.31	-9788.518
M2	19709.25	19860.62	-9837.624

Step 3: Combining Variance Covariate and Grouping Factor:

M6 : weights = varConstPower(form = ~ ArrDelay|Airline)

MODEL	AIC	BIC	LOGLIK
LM	19786.86	19920.43	-9878.432
M4	18474.51	18625.89	-9220.257
M6	18188.05	18392.85	-9071.024

Step 2: Variance Function for Arrival Delay. VarConstPower expected to be better (values ~ 0):

M3 : weights = varPower (form = ~ ArrDelay)
M4 : weights = varConstPower (form = ~ ArrDelay)

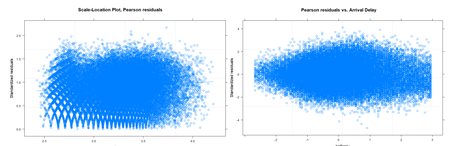
MODEL	AIC	BIC	LOGLIK
LM	19786.86	19920.43	-9878.432
M3	19143.76	19286.23	-9555.883
M4	18474.51	18625.89	-9220.257

Step 4: Selecting the Best Model: M6.

```
Coefficients:
                                      Value p-value
(Intercept)                       3.432843   0e+00
Operating_AirlineAA              -0.050968   0e+00
Operating_AirlineDL              -0.137998   0e+00
Operating_AirlineWN              -0.431013   0e+00
ArrDelay                          0.197696   0e+00
Distance                          0.051819   0e+00
OriginJFK                         0.145411   0e+00
OriginPHX                        -0.310152   0e+00
Concurrent_TOT                    0.048097   0e+00
Operating_AirlineAA:ArrDelay     -0.047080   0e+00
Operating_AirlineDL:ArrDelay     -0.060177   0e+00
Operating_AirlineWN:ArrDelay     -0.093560   0e+00
ArrDelay:OriginJFK                0.013778   1e-04
ArrDelay:OriginPHX               -0.021929   0e+00
```

- **Likelihood Ratio Test and Residuals:**

- All the predictors are tested again with Likelihood Ratio Test (using Maximum Likelihood to fit the models), to check their significance.
Low p-values confirm predictors' significance.



- Scale-Location Plot doesn't highlight any pattern.
- The scatterplot of Pearson residuals vs. ArrDelay shows improvements from the homoscedastic model.

4. Correlation Structure

- **Correlation Structure Selection:**

- Where can we find correlation?
There may be correlation between taxitime and the:
  - departure airport . Grouping factor : "Origin" .
  - day of the week or the month of the departure. Grouping factor: "DayOfWeek" or "Month" .
  - route of the plane . Grouping factor: "route ID" .
  - size or age of the plane. Grouping factor: "NUMBER_OF_SEATS" or "age" .
  - day of departure. Grouping factor: "FlightDate".

- Main Problem:
Many of those variables led to problems of memory allocation.

- Grouping Factor Selected:
  - "route ID" : 238 levels
  - "FlightDate" : 365 levels

- Why CorCompSymm:
Computational problem made not possible the use of CorSymm.
We introduced the grouping factor only, not having any position variabe to use.

- **Models:** weights = varConstPower(form = ~ ArrDelay|Airline)
M6 : no correlation structure
MCorr.1 : correlation = corCompSymm(form = ~ 1 | routeID)
MCorr.2 : correlation = corCompSymm(form = ~1 | flightDate)

MODEL	AIC	BIC	LOGLIK
M6	18188.05	18392.85	-9071.024
MCorr.1	10324.76	10529.56	-5139.381
MCorr.2	18796.23	19001.03	-9375.115

p-value : <.0001

The best model is MCorr.1:
- the routeID identifyes our "subjects".
- AIC improves significally.

Intervals for Rho (95%): Lower: 0.20, Est. : 0.25, Upper: 0.30
Zero is not included: we keep this correlation structure.

- **Plots of the residuals:**

Similar to those obtained with the use of the variance structure only, but we can notice a slight improvement

5. Random Effects - LMM

- **Random Intercept Selection:**

- Models:
We tested the variables earlier selected:

MRe.1 : random = ~ 1 | Origin
MRe.2 : random = ~ 1 | DayOfWeek
MRe.3 : random = ~ 1 | month
MRe.4 : random = ~ 1 | routeID
MRe.5 : random = ~ 1 | NUMBER_OF_SEATS
MRe.6 : random = ~ 1 | age
MRe.7 : random = ~ 1 | FlightDate

Formulation: log(TaxiTime) ~ Airline + ArrDelay + Distance + Origin + ConcurrentFlights + Airline:ArrDelay + Origin:ArrDelay

weights: varConstPower (form = ~ ArrDelay | Airline)

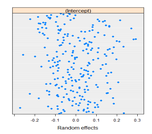
No correlation structure

- AIC test

```
       Model df      AIC      BIC    logLik
MRe.1      1 23 20213.50 20418.30 -10083.749
MRe.2      2 23 20187.39 20392.19 -10070.695
MRe.3      3 23 19837.29 20042.09  -9895.644
MRe.4      4 23 11292.06 11496.86  -5623.031
MRe.5      5 23 17813.65 18018.45  -8883.826
MRe.6      6 23 19651.94 19856.74  -9802.970
MRe.7      7 23 19464.32 19669.12  -9709.159
```

The best model with random intercept is MRe.4., which is the one associated to the lowest AIC (11.292) and uses as grouping factor routeID.

- Plot of the Random Intercepts (Mre.4)



From the plot we can see that all the random intercept are very close to 0.

- Random Intercept vs Correlation Structure

AIC MCorr.1 Vs. AIC MRe4: the best one is the one with the correlation structure (MCorr.1).

```
       Model df      AIC      BIC    logLik
MCorr1     1 23 10324.76 10529.56 -5139.381
MRe.4      2 23 11292.06 11496.86 -5623.031
```

- LMM
LMMCorr.1: variance structure, correlation structure and random intercept Vs.
MCorr.1: variance structure and correlation structure.
A smaller sample of the dataset was used to fit the models

```
             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
MCorr.1          1 23 2192.314 2358.122 -1073.157
LMMCorr.1        2 24 2181.540 2354.557 -1066.770 1 vs 2 12.77449   4e-04
```

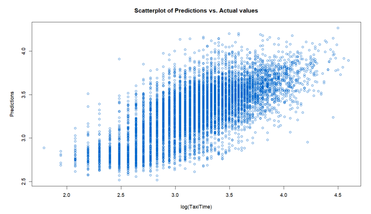
- the improven given by the LMM model is minimum
- the model is difficult to run while dealing with the complete dataset. We keep MCorr.1.

6. Predictions and Results

Model Formulation :
log(TaxiTime) ~ Airline + ArrDelay + Distance + Origin + ConcurrentFlights + Airline:ArrDelay + Origin:ArrDelay

Model :
weights = varConstPower (form = ~ ArrDelay | Airline),
correlation = corCompSymm (form = ~1 | routeID)

Test Set :
- 10k observations randomly selected from the initial dataset, balanced in Origin Airport (33.3% for each).
- ~8900 observations used for prediction.




```
        Welch Two Sample t-test

data:  test_setSresidual and model_df$residual
t = -0.034257, df = 12066, p-value = 0.9727
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006768341  0.006535832
sample estimates:
 mean of x  mean of y
-0.1054691 -0.1053529
```

RMSE 0.315

Results:

- **What are the most impactful features for Taxi Time?** Arrival Delay, Origin Airport and Airline are the strongest predictors, followed by Concurrent Flights and Flight Distance. Taxi Time therefore mainly depends on how internally efficient and well-organized Airports and Airlines are and on crowding of runways. Moreover, disruptions in schedules caused by Arrival Delays determine additional waiting times (ex. waiting for services such as buses for passengers, boarding sleeves, airport personnel, ...).

- **Can we accurately predict Taxi Time?** With the data available, Taxi Time can be generally predicted in its correct range. To improve accuracy, more in-depth data would be necessary. For example, adding information about International Flights departing and arriving to each Airport would likely increase the significance of the variable Concurrent Flights; airports' internal data would possibly help as well (ex: how much personnel and facilities are available/used during the day).

Airline	Average TaxiTime	Kg Fuel per Flight	Kg CO2 per Flight	Fuel Cost per Flight
Southwest Airlines	17.38 minutes	405 kg	1276 kg	430 €
American Airlines	28.40 minutes	662 kg	2085 kg	702 €
Delta Air Lines	30.30 minutes	706 kg	2224 kg	748 €
Endeavor Air	32.29 minutes	752 kg	2370 kg	797 €