# QSI 2.0: MNL model

## Overview

A Fair Market Share model (FMS) is used to estimate the desirability of all the paths going from an airport A to an airport B, without taking into account the prices, but only the characteristics of the path. It can be very useful to compare desirability of the paths, to check if reality of market shares (the desirability of one path taking into account also the price) is far from fair market share and to use its results for other tools.

Air France-KLM already has a FMS model, called QSI, which has been calibrated on aging US data and presents some limitations (it works with a limited number of variables and categories, it does not give a clear overview about variables impact and it is impossible to include continuous variables). In order to gain knowledge on the FMS, a new model has been studied and implemented: the Multinomial Logistic regression model (MNL).

### The MNL model

Multinomial Logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

The idea behind MNL model, as in many other statistical classification techniques, is to construct a linear predictor function that constructs a score from a set of weights that are linearly combined with the explanatory variables (features) of a given observation using a dot product:

$$score(\boldsymbol{x_i}, k) = \beta_k \cdot \boldsymbol{x_i}$$

Where $\boldsymbol{x_i}$ is the vector of explanatory variables describing observation $i$ (the itinerary $i$), $\beta_k$ is a vector of weights (or regression coefficients) corresponding to outcome $k$, and $score(\boldsymbol{x_i}, k)$ is the score associated with assigning observation $i$ to category $k$.

Training the MNL model means finding the best vector of coefficients $\boldsymbol{\beta}$. After the training step, the probabilities (the Fair Market Share) of one itinerary $i$ belonging to a specific OnD (Origin and Destination) is computed as follows:

$$\Pr(y = i|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x_i})}{\sum_{j \in OnD} \exp(\boldsymbol{\beta} \cdot \boldsymbol{x_j})}$$

## MNL – operating level

Given an OnD (for instance PAR – NYC), the aim is to calculate the Fair Market Share of all the operated itineraries from the origin to the destination (Figure 1: FMS at operating level).

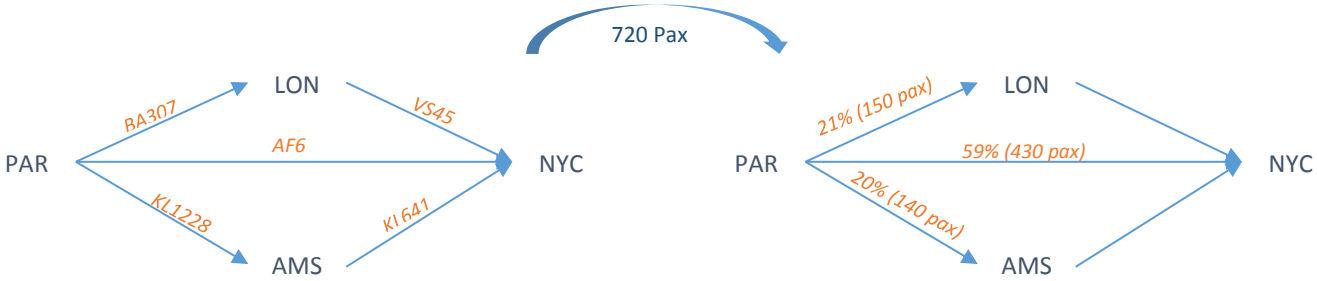


*Figure 1: FMS at operating level*

Features considered in the model are listed in Figure 2: features for MNL operating model.

| Characteristic | Description |
|---|---|
| Connection time | • Total connection time of the path |
| Elapsed time | • Total elapsed time of the path |
| Elapsed time ratio | • Ratio of elapse time and minimum elapse time of the path |
| Is direct/connecting itinerary | • Is the path direct? |
| Number of stops/connections | • Number of stopovers<br>• Number of connections |
| Frequency | • How often is the path operated during the week |
| Carrier reputation | • Skytrax rate |
| Time preference | • Computation of the preferred departure time depending on Local Elapse Time |
| Market presence | • Ratio of the number of paths of the company on the OD by total number of paths<br>• Percentage of flights from the company at origin<br>• Percentage of flights from the company at destination |
| Connection type | • Number of code shares<br>• Is the same airline operating all legs?<br>• Number of connections in the same city but different airports |
| Shortest leg ratio | • Ratio between longest and shortest leg |
| Double sided connectivity | • How many return paths ate offered by the operating company per week |
| Aircraft reputation | • Is the flight and A380 or B787 |
| Carrier clustering | • Percentage of the path operated by each type of company (LCC/legacy…) |

*Figure 2: features for MNL operating model*

All the OnDs have been divided in several groups, with respect to the origin and destination continents: this means that we computed the coefficients $\boldsymbol{\beta}$ for each group (one vector for Europe-Europe Onds, one for Europe-North America Onds…), using the following continents:

- EUR (Europe)
- AME (America)
- AMO (Africa & Middle East)
- APC (Asia and Pacific)
- COI

## Steps of the training code

For all the OnDs of each group and all days:

1. The start point is the pairwise dataset, with all possible pairs of paths of the same OnD:

| Paths | OnD | Day | Features of the first itinerary of the pair | | | | Features of the second itinerary of the pair | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Feature 1 | Feature 2 | … | Feature N | Feature 1 | Feature 2 | … | Feature N |
| **Path A vs path B** | **CDG-JFK** | **1** | $\varphi_{1,A}$ | $\varphi_{2,A}$ | … | $\varphi_{N,A}$ | $\varphi_{1,B}$ | $\varphi_{2,B}$ | … | $\varphi_{N,B}$ |
| **Path B vs path C** | **CDG-JFK** | **1** | $\varphi_{1,B}$ | $\varphi_{2,B}$ | … | $\varphi_{N,B}$ | $\varphi_{1,C}$ | $\varphi_{2,C}$ | … | $\varphi_{N,C}$ |
| **Path A vs path C** | **CDG-JFK** | **1** | $\varphi_{1,A}$ | $\varphi_{2,A}$ | … | $\varphi_{N,A}$ | $\varphi_{1,C}$ | $\varphi_{2,C}$ | … | $\varphi_{N,C}$ |
| **Path B vs path C** | **CDG-JFK** | **2** | $\varphi_{1,B}$ | $\varphi_{2,B}$ | … | $\varphi_{N,B}$ | $\varphi_{1,C}$ | $\varphi_{2,C}$ | … | $\varphi_{N,C}$ |
| **Path D vs path E** | **MAD-SFO** | **1** | $\varphi_{1,D}$ | $\varphi_{2,D}$ | … | $\varphi_{N,D}$ | $\varphi_{1,E}$ | $\varphi_{2,E}$ | … | $\varphi_{N,E}$ |
| **Path D vs path E** | **MAD-SFO** | **2** | $\varphi_{1,D}$ | $\varphi_{2,D}$ | … | $\varphi_{N,D}$ | $\varphi_{1,E}$ | $\varphi_{2,E}$ | … | $\varphi_{N,E}$ |
| **Path E vs path F** | **MAD-SFO** | **2** | $\varphi_{1,E}$ | $\varphi_{2,E}$ | … | $\varphi_{N,E}$ | $\varphi_{1,F}$ | $\varphi_{2,F}$ | … | $\varphi_{N,F}$ |
| **Path D vs path F** | **MAD-SFO** | **2** | $\varphi_{1,D}$ | $\varphi_{2,D}$ | … | $\varphi_{N,D}$ | $\varphi_{1,F}$ | $\varphi_{2,F}$ | … | $\varphi_{N,F}$ |

2. Creation of a pairwise logistic regression with the **difference** of each feature

| Paths | OD | Day | $\Delta\varphi_1$ | $\Delta\varphi_2$ | ... |
|---|---|---|---|---|---|
| **Path A vs path B** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,B}$ | $\varphi_{2,A} - \varphi_{2,B}$ | ... |
| **Path B vs path C** | **CDG-JFK** | **1** | $\varphi_{1,B} - \varphi_{1,C}$ | $\varphi_{2,B} - \varphi_{2,C}$ | ... |
| **Path A vs path C** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,C}$ | ... | |
| **Path B vs path C** | **CDG-JFK** | **2** | $\varphi_{1,B} - \varphi_{1,C}$ | ... | |
| **Path D vs path E** | **MAD-SFO** | **1** | $\varphi_{1,D} - \varphi_{1,E}$ | ... | |
| **Path D vs path E** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,E}$ | ... | |
| **Path E vs path F** | **MAD-SFO** | **2** | $\varphi_{1,E} - \varphi_{1,F}$ | ... | |
| **Path D vs path F** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,F}$ | ... | |

3. Duplication of the database with one more column with **TRUE** or **FALSE** (True = choice of the 1st itinerary, False = choice of the 2nd itinerary)

| Paths | OD | Day | $\Delta\varphi_1$ | $\Delta\varphi_2$ ... | Choice |
|---|---|---|---|---|---|
| **Path A vs path B** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,B}$ | $\varphi_{2,A} - \varphi_{2,B}$ | **TRUE** |
| **Path B vs path C** | **CDG-JFK** | **1** | $\varphi_{1,B} - \varphi_{1,C}$ | $\varphi_{2,B} - \varphi_{2,C}$ | **TRUE** |
| **Path A vs path C** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,C}$ | ... | **TRUE** |
| **Path B vs path C** | **CDG-JFK** | **2** | $\varphi_{1,B} - \varphi_{1,C}$ | ... | **TRUE** |
| **Path D vs path E** | **MAD-SFO** | **1** | $\varphi_{1,D} - \varphi_{1,E}$ | ... | **TRUE** |
| **Path D vs path E** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,E}$ | ... | **TRUE** |
| **Path E vs path F** | **MAD-SFO** | **2** | $\varphi_{1,E} - \varphi_{1,F}$ | ... | **TRUE** |

| | | | | | |
|---|---|---|---|---|---|
| **Path D vs path F** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,F}$ | **…** | **TRUE** |

| Paths | OD | Day | $\Delta\varphi_1$ | $\Delta\varphi_2$ | Choice |
|---|---|---|---|---|---|
| **Path A vs path B** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,B}$ | $\varphi_{2,A} - \varphi_{2,B}$ | **FALSE** |
| **Path B vs path C** | **CDG-JFK** | **1** | $\varphi_{1,B} - \varphi_{1,C}$ | $\varphi_{2,B} - \varphi_{2,C}$ | **FALSE** |
| **Path A vs path C** | **CDG-JFK** | **1** | $\varphi_{1,A} - \varphi_{1,C}$ | **…** | **FALSE** |
| **Path B vs path C** | **CDG-JFK** | **2** | $\varphi_{1,B} - \varphi_{1,C}$ | **…** | **FALSE** |
| **Path D vs path E** | **MAD-SFO** | **1** | $\varphi_{1,D} - \varphi_{1,E}$ | **…** | **FALSE** |
| **Path D vs path E** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,E}$ | **…** | **FALSE** |
| **Path E vs path F** | **MAD-SFO** | **2** | $\varphi_{1,E} - \varphi_{1,F}$ | **…** | **FALSE** |
| **Path D vs path F** | **MAD-SFO** | **2** | $\varphi_{1,D} - \varphi_{1,F}$ | **…** | **FALSE** |

4. Computation of the logistic regression (function glm) on the whole pairwise dataset (merge the True and the False dataset in an unique dataset)

   With the glm function on R we compute the coefficients $\boldsymbol{\beta}$ for each group of continents ($\boldsymbol{\beta}_{Eur-Eur}$, $\boldsymbol{\beta}_{Eur-Ame}$, $\boldsymbol{\beta}_{Ame,Coi}$ …):

   glm(Choice ~ $\Delta\boldsymbol{\varphi_1} + \Delta\boldsymbol{\varphi_2} + \cdots + \Delta\boldsymbol{\varphi_N}$ )

   From these coefficients, we have to obtain a probability for each itinerary, which corresponds to the FMS.

   From the theory, we know that, for any path K, the conditional probability to choose path A is:

$$P(path_A|path_K) = \frac{e^{\beta\cdot\varphi_A}}{1 + e^{\beta\cdot\varphi_A}}$$

   And the score of path A can be calculated as:

$$score(path_A) = \frac{P(path_A|path_K)}{1 - P(path_A|path_K)} = e^{\beta\cdot\varphi_A}$$

   And if we normalize this score, we obtain the FMS of path A:

$$FMS_{opr}(path_A) = \frac{e^{\beta \cdot \varphi_A}}{\sum_{K \in OnD} e^{\beta \cdot \varphi_K}}$$