Worldwide Passenger Flows Estimation

Rodrigo Acuna-Agost [1], Ezequiel Geremia[1], Thiago Gouveia[2], Serigne Gueye[2], Micheli Knechtel[2], and Philippe Michelon[2]

[1]Amadeus IT
[2]Université d'Avignon et des Pays de Vaucluse, LIA

ESGI 2016
Avignon, 23-27 mai 2016

1 Introduction

In 2013, 9.4 million flights have transported 842 million passengers and 13.4 million tons of freight and mail in Europe only. Air traffic has increased considerably in the last years and it is expected that this trend will continue: an annual increase of 2.5% in the number of flights is expected until 2021 and a total increase of 50% is estimated for 2035, bringing to a total of 14.4 million flights ([1, 2, 6]). It is therefore essential that airlines companies, service companies (as Amadeus) and national and international regulatory agencies acquire planning and control tools. Nevertheless, traffic growth and increased number of passenger induce problems of increasingly larger sizes but also new problems.

In this paper, we propose to study one of these new issues consisting in estimating the worldwide number of passengers by origin-destination pairs, on a monthly basis. If the number of passengers per flight may be estimated by statistical methods, they do not allow directly to deduce the number of travellers by origin-destination since several itineraries are generally available for a given pair OD, with the consequence that on a given flight, the passengers have different origins and destinations.

Estimating the number of passengers for each O-D pairs will allow to:

- analyse, over the time, the evolution of the demand for each pair O-D: an origin-destination route in growth may prompt an airline to open flight to serve at least one section, and, conversely, to close flights on routes significantly decreasing.
- estimate tourism flows entering or leaving a given city, constituting a significant economic indicator (again, it is possible to know the number of passengers arriving at a particular airport but not their origin, which is economically important).
- anticipate the spread of infectious diseases such as Ebola or Zika. It might be emphasized that the UNESCO is in relation with Amadeus for this purpose.

Besides obtaining these socio-economic indicators, a rapid resolution of the problem is a prerequisite for more advanced applications, including optimization of network airlines. The network optimization consists in evaluating the effect of the addition or removal of one, or more, flights on the profitability of different markets (a market being precisely defined by the amount of passengers on a O-D route, while profitability is directly related to the aircraft load factor).

2 Problem Definition and Model

Deducting an Origin-Destination matrix from partial data on segments doesn't constitute a new issue by itself. For example, one can cite the Bierlaire's works, which give a clear survey of the various existing approaches, [3, 4, 5]. It is however completely new in the field of aviation and there are two main reasons for this.

The size of the problem in the airline industry is absolutely huge and vastly superior to the inherent problems in other areas, making the resolution by traditional methods unattractive. The second reason is related to obtaining data on a global scale, covering all airlines and airports (whose number exceeds three thousand, constituting more than ten million potential O-D pairs), which are unavailable for most of the companies, except services providers who work with all the airlines companies, such as Amadeus.

Specifically, our problem can be stated as follows:
Knowing the flow of passengers leaving from each airport, the flow of passengers arriving at the airports, an estimated number of passengers on each flight, lower bounds (limit below which the flight is cancelled) and upper (capacity the plane) on the number of passengers that can be transported on the flight, the possible itineraries for each O-D pairs and the probability of using them (again estimated by statistical methods), find the number passenger for each O-D pairs.

Note that in this paper, we will use indifferently the word "flight" and the word "leg" to designate a trip between a take off and a landing. Let now:

- a_i be the total number of passengers arriving at airport i,
- s_i be the total number of passengers leaving airport i,
- α_{od}^l be the proportion of passengers using leg l for going from o to d,
- $\hat{P}_l$ be an estimation of the number of passengers on leg l,
- $\underline{P_l}$ and $\overline{P_l}$ be the lower and upper bounds on the number of passengers on leg l,
- A be the total number of airports,
- L be the total number of legs.

Let us define two sets of decision variables (although the second set of variables could be removed):

- X_{od}: flow of passengers from o to d.
- P_l: number of passengers on leg l.

The problem can then be modelled as follows:

$$
\begin{aligned}
\min \quad & \sum_{l=1}^{L} \beta_l (P_l - \hat{P}_l)^2 \\
\text{s.t.} \quad & \sum_{d=1}^{A} X_{od} = s_o & \forall o \in \{1, ..., A\} \quad & (1) \\
& \sum_{o=1}^{A} X_{od} = a_d & \forall d \in \{1, ..., A\} \quad & (2) \\
& P_l = \sum_{(o,d) \in \{1,...,A\}^2} \alpha_{od}^{l} X_{od} & \forall l \in \{1, ..., L\} \quad & (3) \\
& \underline{P_l} \leq P_l \leq \overline{P_l} & \forall l \in \{1, ..., L\} \quad & (4) \\
& X_{od} \geq 0 & \forall (o,d) \in \{1, ..., A\}^2 \quad &
\end{aligned}
$$

Constraints (1) and (2) ensure that the number of passengers arriving to and leaving the airports are equal to the corresponding data. Constraints (3) compute the number of passengers per leg as a function of the flows of passengers, while constraints (4) and (5) indicate the domains of the decision variables. The objective function minimizes a weighted quadratic error with respect to the expected number of passengers per flight.

As a consequence, the whole problem consists in minimizing a convex quadratic (and separable) objective function subject to linear constraints. Such a problem does not present any particular theoretical difficulties. Nevertheless, worldwide instances include more than 3300 airports, leading to more than ten millions of O-D pairs ! The challenge in solving this non-linear problem is thus to deal with its huge size.

3 Analysis

In a first attempt, we proposed a Linear Programming based approach for solving the problem. The basic idea is to substitute the quadratic error by absolute values:

$$
\begin{aligned}
\min \quad & \sum_{l=1}^{L} \beta_l |P_l - \hat{P}_l| & & \\
\text{s.t.} \quad & \sum_{d=1}^{A} X_{od} = s_o & \forall o \in \{1, ..., A\} & \qquad (1) \\
& \sum_{o=1}^{A} X_{od} = a_d & \forall d \in \{1, ..., A\} & \qquad (2) \\
& P_l = \sum_{(o,d) \in \{1,...,A\}^2} \alpha_{od}^l X_{od} & \forall l \in \{1, ..., L\} & \qquad (3) \\
& \underline{P_l} \leq P_l \leq \overline{P_l} & \forall l \in \{1, ..., L\} & \qquad (4) \\
& X_{od} \geq 0 & \forall (o,d) \in \{1, ..., A\}^2 &
\end{aligned}
$$

Then, by introducing new variables t_l and two constraints (per leg), we got:

$$
\begin{aligned}
\min \quad & \sum_{l=1}^{L} \beta_l t_l & & \\
\text{s.t.} \quad & t_l \geq P_l - \hat{P}_l & \forall l \in \{1, ..., L\} & \\
& t_l \geq -P_l + \hat{P}_l & \forall l \in \{1, ..., L\} & \\
& \sum_{d=1}^{A} X_{od} = s_o & \forall o \in \{1, ..., A\} & \qquad (1) \\
& \sum_{o=1}^{A} X_{od} = a_d & \forall d \in \{1, ..., A\} & \qquad (2) \\
& P_l = \sum_{(o,d) \in \{1,...,A\}^2} \alpha_{od}^l X_{od} & \forall l \in \{1, ..., L\} & \qquad (3) \\
& \underline{P_l} \leq P_l \leq \overline{P_l} & \forall l \in \{1, ..., L\} & \qquad (4) \\
& X_{od} \geq 0 & \forall (o,d) \in \{1, ..., A\}^2 & \\
& t_l \geq 0 & \forall l \in \{1, ..., L\} &
\end{aligned}
$$

Unfortunately, preliminary numerical results, obtained by solving the quadratic and linear models with a commercial solver on small instances, showed that, first, the linear model provides a solution whose quality is worse than the solutions furnished by the quadratic model and, second, the linear model is surprisingly slow in comparison with the quadratic one. It was therefore decided to not go further in this direction. We rather propose a Lagrangean Relaxation approach.

4 A Lagrangean Approach

A closer look to the problem shows that constraint (3) actually links the X_{od} and the P_l variables. Associating a dual variable λ_l to each of these linking constraints and sending then in the objective function leads to the definition of the Lagrangean function L(λ):

$$
\begin{array}{llll}
\mathrm{L}(\lambda) = \min & \sum_{l=1}^{L} \beta_l (P_l - \hat{P}_l)^2 - \lambda_l (P_l - \sum_{(o,d)\in\{1,...,A\}^2} \alpha_{od}^l X_{od}) & & \\
\text{s.t.} & \sum_{d=1}^{A} X_{od} = s_o & \forall o \in \{1,...,A\} & (1) \\
& \sum_{o=1}^{A} X_{od} = a_d & \forall d \in \{1,...,A\} & (2) \\
& \underline{P_l} \leq P_l \leq \overline{P_l} & \forall l \in \{1,...,L\} & (4) \\
& X_{od} \geq 0 & \forall (o,d) \in \{1,...,A\}^2 &
\end{array}
$$

which, in turn, gives:

$$
\begin{array}{llllll}
\mathrm{L}(\lambda) = \sum_{l=1}^{L} & \min & \beta_l (P_l - \hat{P}_l)^2 - \lambda_l P_l & + \min & \sum_{l=1}^{L} \sum_{(o,d)\in\{1,...,A\}^2} \lambda_l \alpha_{od}^l X_{od} & \\
& \text{s.t.} & \underline{P_l} \leq P_l \leq \overline{P_l} & \text{s.t.} & \sum_{d=1}^{A} X_{od} = s_o & \forall o \in \{1,...,A\} \\
& & & & \sum_{o=1}^{A} X_{od} = a_d & \forall d \in \{1,...,A\} \\
& & & & X_{od} \geq 0 & \forall (o,d) \in \{1,...,A\}^2
\end{array}
$$

since there is no more link between X_{od} and P_l variables, nor between two different P_l.

Hence, the Lagrangean Relaxation decomposes the initial problem into a huge number of one variable problems and a problem with a large number of variables . The last problem is the standard transportation problem for which very efficient algorithms have been developed for solving it.

It is well known that L is a concave but non-differentiable function and that $\forall \lambda \in R^L, \mathrm{L}(\lambda) \leq V(P)$ where $V(P)$ is the optimal value of the initial problem. In other word, whatever the multiplier is, $\mathrm{L}(\lambda)$ is a lower bound on the optimal value of the problem. Since, we are interested in finding the best lower bound, we need to solve the Dual problem:

$$\max\{\mathrm{L}(\lambda)/\lambda \in R^L\}$$

Thanks to the convexity of the initial problem, it is also known that the optimal value is actually equal to the optimal value of the original problem.

5 Conclusion

The ESGI week has been dedicated to the understanding of the problem and the available data, to design an instance generator (for preliminary testing) and we proposed a Lagrangean Relaxation approach well suited for parallelization.

References

[1] Eurocontrol. Challenges of growth, 2013 https://www.eurocontrol.int/sites/default/files/article//content/documents/official-documents/reports/201306-challenges-of-growth-2013-task-4.pdf.

[2] Eurocontrol. Seven-year forecast, 2015. https://www.eurocontrol.int/sites/default/files/content/documents/official-documents/forecasts/seven-year-flights-service-units-forecast-2015-2021-Feb2015.pdf.

[3] M. Bierlaire, Evaluation de la demande en trafic : quelques méthodes de distribution, rapport Technique 91/12, Facultés Universitaires Notre-Dame de la Paix de Namur, 1991.

[4] M. Bierlaire, Mathematical models for transportation demand analysis, PhD thesis, Facultés Universitaires Notre-Dame de la Paix de Namur, 1996.

[5] M. Bierlaire, F. Crittin, An efficient algorithm for real-time estimation and prediction of dynamic OD tables, Operations Research 52, 1, 116-127, 2004. year = 2004,

[6] IATA http://www.iata.org/pressroom/pr/Pages/2015-11-26-01.aspx