Prediction of transfer passengers

Author: Szilárd Székely,

Supervisor: Maarten Soomer

Vrije University Amsterdam Faculty of Science

March-August, 2005

CONTENTS

Abstract

This master thesis describes a problem of the Dutch airline company KLM, which I've been working on as graduation for my mathematics study. Airlines timetable is changing all the time. When the new timetable contains major changes KLM interested the estimation of transfer flow. The problem is basically to forecast the number of transfer passengers from each arriving flight to each departing flight. It is essential because of setting the workloads and handling the critical rush periods. We are going to use linear programming model in order to get estimation for the transfers.

1. Problem definition

Introduction

Thousands of KLM passengers arrive in Amsterdam Airport Schiphol day by day. The vast majority of KLM passengers only fly to Amsterdam to transfer to another flight that brings them to their final destination. The transfer process is complex and requires many resources, especially during peak hours. The baggage process is a good example. To better plan staff and resources, a good prediction of transfer passengers is essential. To be more precise, the total number of transfer passengers per day can be considered as a known quantity, but the way that transfers are spread during the day is less trivial, especially when a new schedule contains major changes. Since this happens quite often, KLM wants to have a good prediction of the number of transfer passengers per time interval (e.g. 15 minutes).
For the baggage process, it is also important to predict the number of short connections at any time of the day. A Short connection is a transfer, where the arrival and the departure are very close to each other (30-60 minutes).

An "Incoming wave" is a period of the day, in which the number of arriving flights is relative high compared to other periods of the day.
At the end of an "incoming wave" the number of short connections increases, and consequently the workload for the baggage staff.

In this light, KLM is very interested in an algorithm capable of generating transfers at a detailed level, in such a way that the aggregate flows will be realistic. The transfers are never used at the very low level, but can be summed by various dimensions, providing valuable profiles to help predicting workload and availability of resources. This prediction is essential. It is senseless to have unnecessary overestimated machines and employees because of the cost of it. On the other hand the transfer flow must be handled.

Wouter Couzy and Bart van Asten from the Decision Support department of KLM presented the problem and gave feedback during the research.

The exact mathematical formulation is given later. A characterization of the transfer flow is given below.

Input

Categories

KLM observed that, when introducing a new timetable, the total number of transfer passengers is a known quantity in advance (or it can be estimated), but the way that transfers are spread during the day, is. The flights can be divided into several categories, depending on the airline, destination (European or not) etc. KLM also observed that the total percentage of transfer passengers on all flight in a certain category on a day does not change with a new timetable. Also the percentage of the transfer passengers that arrive from within Europe and connect to a flight to outside Europe does not change. The same holds for the other possibilities of connections. For these connection types also the average percentage of transfer passengers that have a certain transfer time is not changing.

The flights can be gathered by categories.
Let $C(i), i = 1..M$ be, a set of categories. Each category contains airlines that share the same transfer characteristics. KLM distinguishes the following categories:

i	*C(i)*	Description
1	KLM-Eur	All KLM flights from/to Europe
2	KLM-Ica	All KLM flights from/to ICA (intercontinental flight), except Paramaribo and Caribbean

3	Partner-Eur	All flights operated by partner airlines from/to Europe, including Air France
4	Partner-Ica-High	All flights operated by partner airlines from/to ICA, with high transfer rate
5	Partner-Ica-Low	All flights operated by partner airlines from/to ICA, with relative low transfer rate + KLM Caribbean
6	HV-Charter	All Transavia chartered flights
7	HV-Scheduled	All Transavia scheduled flights
8	MP-Charter	All MartinAir chartered flights
9	MP-Scheduled	All MartinAir scheduled flights
10	Rest-Eur	All flights operated by other carriers from/to Europe
11	Rest-Ica	All flights operated by other carriers from/to Ica + KLM-Paramaribo

Table 1: Categories of the flights

Arrivals

Let *A(i), i=1..M,* be mutually disjoined sets of arrivals belonging to a category *C(i).* Each arrival is characterised by a flight number (FLTNBR), scheduled time of arrival (STA), the number of passengers (pax) on board, and a EUR/ICA indicator.

Departures

Let *D(i), i=1..M*, be mutually disjoined sets of departures belonging to a category *C(i).* Each departure is characterised by a flight number, scheduled time of departure (STD), the number of passengers on board, and an EUR/ICA indicator.

Schedule

The set of arrivals, $A = \cup A(i)$, together with the set of departures, $D = \cup D(i)$, form the schedule of one day at Amsterdam Airport Schiphol.

ARRIVALS					DEPARTURES				
index	Scheduled time of arrival	# of pax on the board	Category	ICA/ EUR	index	Scheduled time of departure	# of pax on the board	Category	ICA/ EUR
1	8:45	120	2	ICA	1	8:45	247	1	EUR
2	8:45	200	5	ICA	2	8:45	91	9	ICA
3	9:00	80	1	ICA	3	8:45	122	3	EUR
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
n_a	23:20	95	6	EUR	n_d	24:15	120	4	EUR

Table 1: example of schedule

Average percentage of passengers with a transfer

Let $T(i)$, $i=1..M$, be the average percentage of transfer passengers for both $A(i)$ and $D(i)$. These percentages can be determined using historical data. It is assumed that those percentages will not change when a new timetable is introduced.

Example, $i=9$: The average number of passengers with a transfer is 20% for arriving and departing chartered flights from MartinAir.

Average percentage of EUR to EUR, EUR to ICA, ICA to ICA, ICA to EUR transfers.

Let *P(EUR to EUR)* be the probability that a transfer passenger arriving on a flight from within Europe, continues his/her journey on a flight with a destination in Europe.
Let *P(EUR to ICA)* be the probability that a transfer passenger arriving on a flight from within Europe, continues his/her journey on a flight with a destination outside Europe

P(ICA to ICA and *P(ICA to EUR)* are defined likewise.

Example:
P(EUR to EUR)=0.31, P(EUR to ICA)=1-P(EUR to EUR).
P(ICA to EUR)=0.89, P(ICA to ICA)=1-P(ICA to EUR).

These percentages can also be determined using historical data and it is again assumed that those percentages will not change when a new timetable is introduced

Minimum connecting times

There is a natural lower bound for the connection times, since the baggage process and the plane changing take some time.
The minimum connecting time (MCT) at Amsterdam Airport Schiphol is 40 minutes.

Distribution of transfer times

Let *K(EUR to EUR), K(EUR to ICA), K(ICA to EUR)* and *K(ICA to ICA)* be the distribution of transfer times (=STD-STA). These are given as probabilities

that the transfer time is in a certain interval. These distributions are again determined using historical data and are assumed to remain unchanged.

* Confidential *

Figure 1:connection time distribution (Eur-Eur)

Problem

Determine, within a few seconds of PC computation time, the number of transfers between each arrival(element from *A*), and each departure(element from *D*) such that the following conditions are satisfied exactly[1]:

- For each arrival, the number of transfer pax is less than or equal to the number of passengers.
- For each departure, the number of transfer pax is less than or equal to the number of passengers.
- No connections shorter than MCT.
- For each set of arrivals *A(i)*, the total number of transfer pax is equal to the total number of passengers in that set multiplied by the average percentage of transfer passengers in the set (*T(i)*).

- For each set of departures *D(i),* the total number of transfer pax is equal to the total number of passengers in that set multiplied by the average percentage of transfer passengers in the set (*T(i)*).
- The percentage of transfer passengers arriving on a flight from within Europe that are connecting to a flight with a destination in Europe equals *P(EUR to EUR)*.
- The percentage of transfer passengers arriving on a flight from outside Europe that are connecting to a flight with a destination in Europe equals *P(ICA to EUR).*

Whereas we look for the best fit to the distributions of transfer times

[1] For some particular schedules the constraints have to be relaxed.

2. Mathematical model
LP-approach

Linear programming

Any time when we face a constrained optimization problem we can think about linear programming. We are looking for values $x_1, x_2, x_3 ... x_n$, such that the following constraints hold:

$$a_{1,1}x_1 + a_{1,2}x_2 + a_{1,3}x_3 + ... + a_{1,n}x_n \leq b_1$$

$$a_{2,1}x_1 + a_{2,2}x_2 + a_{2,3}x_3 + ... + a_{2,n}x_n \leq b_2$$

…

$$a_{m,1}x_1 + a_{m,2}x_2 + a_{m,3}x_3 + ... + a_{m,n}x_n \leq b_m$$

And we would like to optimize

$$\min_{x_1, x_2 .. x_n} c_1x_1 + c_2x_2 + c_3x_3 + ... + c_nx_n$$

This function is called the objective functions. The variables $x_1, x_2, x_3 ... x_n$ are called decision variables.

We are not going to go very detailed in linear programming. We are using a model in order to gain a solution for a real problem. To examine the theory deeper, see Robert J. Vanderbei, Linear programming: Foundations and extensions, 2001.

Simplex method

This is the most commonly used method to solve linear programming problems. This method has already proved in practice to be quite efficient. Actually it starts in an edge of the solution set and it moves to the direction of

better solution along the edges. An edge is a part of the solution set where inequality(ies) is (are) granted by equality(ies).

Next example presents it in 2D clearly. Let's suppose we have the following constraints:

$5 \geq x \geq 0$

$5 \geq y \geq 0$

$x + y \leq 9$

Then the solution set

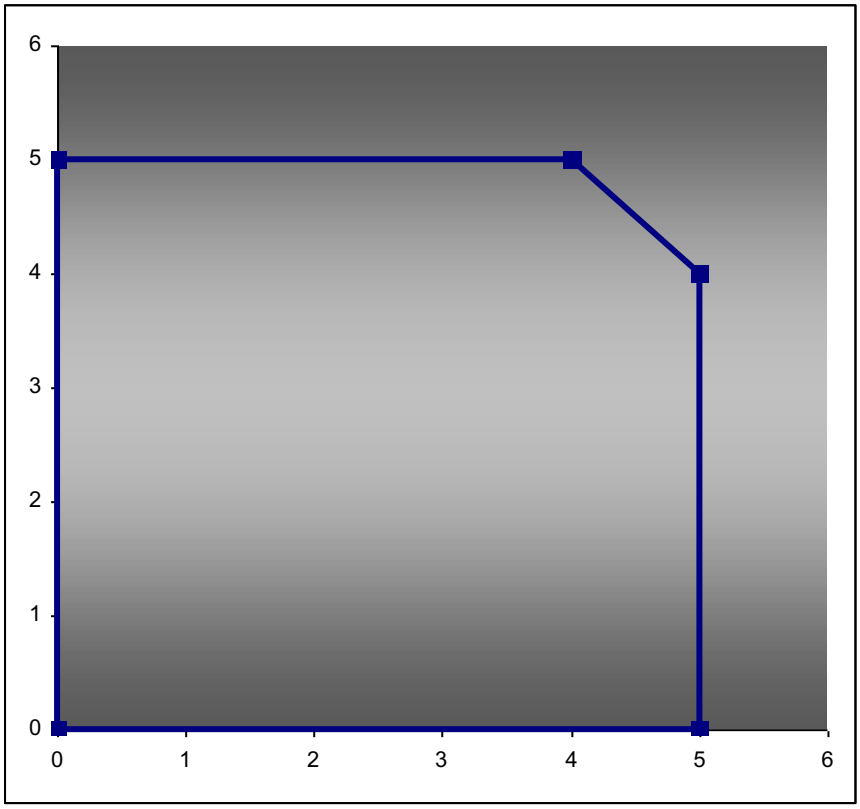

Figure 2: example in 2D

The blue line shows the edges. Simplex method starts in a corner and moves along the edges. It leads to the optimal solution because there is a theorem that says if there is optimal solution in the solution set then there is optimal solution among the corners as well. Optimal solution can appear inside the solution set but simplex method will prefer the optimal solution in the corner. Since we have constraints $x_i \geq 0$ it means many zero elements in the solution if it is possible. This phenomenon can be important in the case we have more than one solution. In this case the solver is going to choose the solution with

many zero elements. Later we talk about it when problem appears because of it.

Application of the LP on the problem

Notation

n_a: number of arriving flights

n_d: number of departing flights

T_i^a: scheduled time of the *i-th* arriving flight

T_j^d: scheduled time of the *j-th* departing flight

C_i^a: category of the *i-th* arriving flight

C_j^d: category of the *j-th* departing flight

M: number of the categories (in our case $M = 11$)

N_i^a: number of passengers on the board of the *i-th* arriving flight

N_j^d: number of passangers on the board of the *j-th* departing flight

I_i^a: 0 if *i-th* flight is an ICA arrival
1 if *i-th* flight is an EUR arrival

I_j^d: 0 if *j-th* flight is an ICA departure

1 if *j-th* flight is an EUR departure

MCT : minimum connection time

$T(i)$: average percentage of transfer passengers for both arrivals from *i-th* category and departures from *i-th* category.

$P_{EUR-EUR}, P_{ICA-EUR}$: the probability that a transfer passenger arriving on a flight from A-EUR or A-ICA, continues his/her journey on a flight from D-EUR.

$K_{EUR-EUR}, K_{EUR-ICA}..etc..$: distributions of connection time using EUR/ICA arrival flight and EUR/ICA departure flight. It means four distributions.

Decision variables

We are looking for the expected number of transfer passengers from each flight to each flight. Let these values be our decision variables.

t_{ij} **:** number of transfer pax between *i-th* arrival flight and j.th departure flight. It is defined just in the case:

$$T_i^a + MCT \leq T_j^d$$

Constraints

$t_{ij} \geq 0 \quad \forall i, j$ (i)

This is a natural assumption. The estimated number of transfer passengers must be positive. At this stage we do not insist on getting an integer solution. Of course the (non-integral) solution can be rounded

$$\sum_{j=1}^{n_d} t_{ij} \le N_i^a \quad \forall i \quad i = 1..n_a \qquad \text{(ii)}$$

For each arrival, the number of transfer pax is less than or equal to the number of passengers.

$$\sum_{i=1}^{n_a} t_{ij} \le N_j^d \quad \forall j \quad j = 1..n_d \qquad \text{(iii)}$$

For each departure, the number of transfer pax is less than or equal to the number of passengers.

$$\frac{\sum_{i \in K}\sum_{j} t_{ij}}{\sum_{i \in K} N_i^a} \le T(k) + \varepsilon_1 \text{ and } \frac{\sum_{i \in K}\sum_{j} t_{ij}}{\sum_{i \in K} N_i^a} \ge T(k) - \varepsilon_1 \text{ where } K = \left\{ i : C_i^a = k \right\} \text{ for all}$$

$$k = 1..11 \qquad \text{(iv)}$$

For each category, the total number of transfer pax at arrivals should be approximately equal to the total number of passengers in that set multiplied by *T(i)*.

$$\frac{\sum_{i}\sum_{j \in L} t_{ij}}{\sum_{j \in L} N_j^d} \le T(k) + \varepsilon_2 \text{ and } \frac{\sum_{i}\sum_{j \in L} t_{ij}}{\sum_{j \in L} N_j^d} \ge T(k) - \varepsilon_2 \text{ where } L = \left\{ j : C_j^d = k \right\} \text{ for all}$$

$$\forall k = 1..11 \qquad \text{(v)}$$

For each category, the total number of transfer pax at departures should be approximately equal to the total number of passengers in that set multiplied by *T(i)*.

Constraint (iv) ensures that $\sum_{i\in K}\sum_{j} t_{ij} \approx \sum_{i\in K}\left(N_i^a T(C_i^a)\right)$ and Constraint (v) ensures that $\sum_{i}\sum_{j\in K} t_{ij} \approx \sum_{j\in K}\left(N_{ji}^d T(C_j^d)\right)$, where *K* is any set of categories.

This can be used when formulating a constraint that ensures that the percentage of transfer passengers arriving on a flight from within Europe that are connecting to a flight with a destination in Europe equals *P(EUR to EUR)*.

Let *C* be the set of categories containing flights having an origin or destination in Europe. Than we can formulate the constraint as

$$\frac{\sum_{i\in C}\sum_{j\in C} t_{ij}}{\sum_{i\in C}\sum_{j} t_{ij}} = P_{EUR-EUR}$$

However this constraint is not linear in the decision variables, but we can use the above to replace the denominator. Now, the constraints (after adding epsilon) become:

$P_{EUR-EUR} - \varepsilon_3 \le \frac{\sum_{i\in K}\sum_{j\in L} t_{ij}}{\sum_{c\in C} T(c)\times(\sum_{i\in M(c)} N_i^a)} \le P_{EUR-EUR} + \varepsilon_3$ where $K = \{i : I_i^a = EUR\}$,

$L = \{j : I_j^d = EUR\}$, C is the set of the categories

from EUR (KLM-EUR, Rest-EUR..etc.) and $M(c) = \{i : C_i^a = c\}$ (vi)

This thus states that the percentage of transfer passengers arriving on a flight from within Europe that are connecting to a flight with a destination in Europe should be approximately equal to *P(EUR to EUR)*.

$$P_{ICA-EUR} - \varepsilon_4 \le \frac{\sum_{i \in K} \sum_{j \in L} t_{ij}}{\sum_{c \in C} T(c) \times (\sum_{i \in M} N_i^a)} \le P_{ICA-EUR} + \varepsilon_4 \quad \text{where } K = \{ i : I_i^a = ICA \},$$

$L = \{ j : I_j^d = EUR \}$, C is the set of the categories from ICA (KLM-ICA, Rest-ICA...etc.) and $M(c) = \{ i : C_i^a = c \}$ (vii)

This constraint is obtained in the same manner as constraint (vi) and states that the percentage of transfer passengers arriving on a flight from outside Europe that are connecting to a flight with a destination in Europe should be approximately equal to *P(ICA to EUR)*.

Goal function

In our case the goal function will measure the difference between the connection time distributions in a solution and the given connection time distributions (Figure 1).

<u>* Confidential *</u>

Table 2: connection time distributions in the reality

Every possible solution from the solution set determines an empirical distribution:

	P(Ttime<60)	P(Ttime<90)	P(Ttime<120)	P(Ttime<240)	P(Ttime<480)	P(Ttime<720)	P(Ttime<1440)
EUR-EUR	X_{11}%	X_{12}%	X_{13}%	X_{14}%	X_{15}%	X_{16}%	100%
EUR-ICA	X_{21}%	X_{22}%	X_{23}%	X_{24}%	X_{25}%	X_{26}%	100%
ICA-EUR	X_{31}%	X_{32}%	X_{33}%	X_{34}%	X_{35}%	X_{36}%	100%
ICA-ICA	X_{41}%	X_{42}%	X_{43}%	X_{44}%	X_{45}%	X_{46}%	100%

Table 3: empirical distribution in a solution from the solution set

We have to choose that solution which best fit the real distribution.
In order to reach this solution we define a function:

$F(t) = \sum_{k,l} \left| \; X_{k,l} - X'_{k,l} \; \right|$ If the real value is $X'_{k,l}$.

It is possible to create more complicated function to simulate the difference between the distribution functions but the above is easy to calculate and is linear because $X_{k,l}$ can be determined as a linear function of t_{ij}.

Let $f(l) : \{1,2,3,4,5,6,7\} \to \{60,90,120,240,480,720,1440\}$ a function.

$f(1) = 60, f(2) = 90, ..etc..$

$X_{1,l} = \dfrac{\sum t_{ij}}{\sum_{i \in K, j \in L} t_{ij}}$ where $K = \{i : I_i^a = EUR \;\}$ and $L = \{j : I_j^d = EUR \;\}$

$$X_{2,l} = \frac{\sum_{i,j:\left| T_i^a - T_j^d \right| \le f(l), i \in K, j \in L} t_{ij}}{\sum_{i \in K, j \in L} t_{ij}} \text{ where } K = \left\{ i : I_i^a = EUR \right\} \text{ and } L = \left\{ j : I_j^d = ICA \right\}$$

$$X_{3,l} = \frac{\sum_{i,j:\left| T_i^a - T_j^d \right| \le f(l), i \in K, j \in L} t_{ij}}{\sum_{i \in K, j \in L} t_{ij}} \text{ where } K = \left\{ i : I_i^a = ICA \right\} \text{ and } L = \left\{ j : I_j^d = EUR \right\}$$

$$X_{4,l} = \frac{\sum_{i,j:\left| T_i^a - T_j^d \right| \le f(l), i \in K, j \in L} t_{ij}}{\sum_{i \in K, j \in L} t_{ij}} \text{ where } K = \left\{ i : I_i^a = EUR \right\} \text{ and } L = \left\{ j : I_j^d = ICA \right\}$$

It is still non-linear but we can introduce the following constraints which ensure that we finally get linear goal function.

In order to obtain $X_{1,l}$:

$\sum_{i \in K, j \in L} t_{ij} = (\sum_{c \in C} T(c) \times (\sum_{i \in M(c)} N_i^a)) \times P_{EUR-EUR}$ where $K = \left\{ i : I_i^a = EUR \right\}$,

$L = \left\{ j : I_j^d = EUR \right\}$

It was Constraint (vi) in the ideal case ($\varepsilon_3 = 0$).

C is the set of the categories from EUR (KLM-EUR, Rest-EUR...etc.) and $M(c) = \left\{ i : C_i^a = c \right\}$

Now, $X_{1,l} = \frac{\sum_{i,j:\left| T_i^a - T_j^d \right| \le f(l), i \in K, j \in L} t_{ij}}{(\sum_{c \in C} T(c) \times (\sum_{i \in M(c)} N_i^a)) \times P_{EUR-EUR}}$ where $K = \left\{ i : I_i^a = EUR \right\}$,

$L = \left\{ j : I_j^d = EUR \right\}$

the denominator is constant.

Similarly the others,

in order to obtain $X_{2,l}$:

$$\sum_{i\in K, j\in L} t_{ij} = (\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{EUR-ICA} \text{ where } K = \{i : I_i^a = EUR\},$$

$$L = \{j : I_j^d = ICA\} \quad \text{(ix)}$$

C is the set of the categories from EUR (KLM-EUR, Rest-EUR...etc.) and $M(c) = \{i : C_i^a = c\}$

Hence:

$$X_{2,l} = \frac{\sum_{i,j:|\ T_i^a - T_j^d| \le f(l), i\in K, j\in L} t_{ij}}{(\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{EUR-ICA}}$$

where $K = \{i : I_i^a = EUR\}$, $L = \{j : I_j^d = ICA\}$

In order to obtain $X_{3,l}$:

$$\sum_{i\in K, j\in L} t_{ij} = (\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{ICA-EUR} \text{ where } K = \{i : I_i^a = ICA\},$$

$$L = \{j : I_j^d = EUR\} \quad \text{(x)}$$

C is the set of the categories from ICA (KLM-ICA, Rest-ICA..Etc.) and $M(c) = \{i : C_i^a = c\}$.

$$X_{3,l} = \frac{\sum_{i,j:|\ T_i^a - T_j^d| \le f(l), i\in K, j\in L} t_{ij}}{(\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{ICA-EUR}}$$

In order to obtain $X_{4,l}$:

$$\sum_{i\in K, j\in L} t_{ij} = (\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{ICA-ICA} \text{ where } K = \{i : I_i^a = ICA\},$$

$$L = \{j : I_j^d = ICA\} \quad \text{(xi)}$$

C is the set of the categories from ICA (KLM-ICA, Rest-ICA...Etc.) and $M(c) = \{i : C_i^a = c\}$.

$$X_{4,l} = \frac{\sum_{i,j:|\ T_i^a - T_j^d| \le f(l), i\in K, j\in L} t_{ij}}{(\sum_{c\in C} T(c) \times (\sum_{i\in M(c)} N_i^a)) \times P_{ICA-ICA}} \text{ where } K = \{i : I_i^a = ICA\},$$

$$L = \{j : I_j^d = ICA\}$$

Now, we have a function: $F(t_{ij}) = \sum_k \sum_l |X_{k,l} - X'_{k,l}|$ where $X_{k,l} - X'_{k,l}$ is a linear function of t_{ij} variables for all *k,l.*

There is a simple trick how to convert this form to linear form.

Let $d_{k,l}$ be new variables. Later we these extra variables.

If we make some extra constrains:

$$d_{k,l} \ge X_{k,l} - X'_{k,l},\ \forall k,l \quad \text{(xii)}$$

And

$$d_{k,l} \ge -X_{k,l} + X'_{k,l},\ \forall k,l \quad \text{(xiii)}$$

We have to minimize : $F(t_{ij}) = \sum_k \sum_l |X_{k,l} - X'_{k,l}|$

This is equivalent to minimize $\sum_{k,l} d_{k,l}$

This way we look for the best fit for the various transfer time distributions. (Table 2)

Complexity

This part is a brief examination about the size of the problem. Let *m, n* denote the number of the arrivals and departures respectively.

The number of the variables – the possible connections – is approximately equal to $\frac{m \times n}{2}$.

It comes from the definition of the variables. The number of the extra variables (24) is negligible compared to this number since $m \approx 500, n \approx 500$ in our data set.

The number of the constraints is approximately equal to $\frac{m \times n}{2} + m + n + 2M$ where *M* is the number of the categories (here *M* =11).

3. Results of implementation of the first LP-model

Input data

The original data file contains data of 29 days. This is historical data. We have exact data about the scheduled departure and arrival times and the number of passengers on board each flight. The categories of the flights are known as well.
KLM also provided the characterization of the categories and the various probabilities that we mentioned in the first chapter. Our purpose is to get an estimation about transfer passengers per time interval (15 minutes) from each arriving flight to each departing fight for each given day.

There are considerable differences between individual days. For instance just in the number of arrivals and departures. On the first day there were 530 arrivals. On the third day there were only 478 and on the 29*th* there were 508. Even if every other flight has the same details on different days there are completely different flights in different schedules This means all days have to be evaluated individually, and will probably have a completely different transfer flow.

Constraints

It would be ideal to use $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = 0$. Unfortunately, this doesn't provide a feasible solution for all schedules (see table 4). That is why we had to set these epsilons bigger than zero. Later it is written detailed (Table 4) how we changed epsilons exactly.
With these changes the problem is solvable. For some schedules it works with relative small epsilons. ($\varepsilon = 0.03$). In these cases we are very close to the

original constraints. Unfortunately there are some schedules with them the solver finds this problem infeasible if $\varepsilon \leq 0.1 - 0.2$. It is far from the original constraints. The consequence of it is a transfer flow with features that are not typical of Schipol.

With smaller or bigger epsilons every schedule (as input) is feasible. The connection time distributions fit well.
There is a question that is realistic solution or we get an extreme solution with too many $t_{ij} = 0$ and with very seldom $t_{ij}>0$. But these values are extremely high.

Implementation

The biggest difficulty of this project was the implementation. Although it is not too difficult to write a program for this LP-problem, the computation time might be unrealistic huge on a personal computer.

The original input was the schedule of the current day. It means around 500 arrival and around 500 departure flights. The data of these flights can be stored in vectors in the same size. It is not a problem. On the other hand we have to build the vector of the variables (in size ~500x250) and the constant matrix (in size ~ (500x250) x (500x250)). This matrix is full of zero elements but if we store it like ordinary constant matrix its size is few gigabytes on the hard drive.

The implementation was done using C++ and the Coin LP solver.
We tested the program on eight different days. Those were choosen random. All instances are solved within a few minutes. The results of four of them can be seen detailed below.

Tests proved that the optimal solution is not unique. It means there are solutions with different values for decision variables and the same objective

value. If we take a look at table nr.3 (last 3 rows) we can see there is a solution with an upper bound (This is the artificial limit for the individual transfers (t_{ij})) that is obviously also a feasible solution to the problem without upper bound. This solution is also optimal for the problem without upper bound, because it has the same objective value as the optimal solution found by the solver for this problem. However if we look at the ratio of non-zero's in the solutions, we observe that the solutions are not the same.

date	upper bound	ε_1 Constraint (iv)	ε_2 Constraint (v)	ε_3 Constraint (vi)	ε_4 Constraint (vii)	Ratio of non-zero elements	Objective value
1.	-	0.03	0.03	0.1	0.1	INFEASIBLE	
2.	-	0.03	0.03	0.1	0.1	0.3%	0
2.	10	0.03	0.03	0.1	0.1	1.9%	0
2.	8	0.03	0.03	0.1	0.1	2.4%	0
2.	8	0.01	0.01	0.08	0.08	2.4%	0.035
3.	-	0.03	0.03	0.03	0.03	0.3%	0.12
3.	8	0.03	0.03	0.03	0.03	2.5%	0.12
3.	10	0.03	0.03	0.03	0.03	2%	0.12
9.	-	0.01	0.01	0.01	0.01	0.3%	0.048
9.	10	0.01	0.01	0.01	0.01	1.7%	0.048
9.	8	0.01	0.01	0.01	0.01	2.2%	0.048

Table 4: test results

In the first column of table 4 we can see which day's schedule was used as input. The second one informs about the upper bound. We add it to the model like an extra constraint. The next four columns show us how strict we were in the constraints (iv)-(vii).

The last two columns are the issue of the test. The objective values measure the difference between the desirable connection time distributions and the empirical distributions in the optimal solution.

Usually the optimal solution has very good (very small differences in the connection time distribution) objective value. (~0.1) This is a summation of 24 differences hence average ~0.005 difference is between the real and created probabilities.

The non-zero elements in the optimal solution

The first tests show that different day has quite different solution. Another problem is the following:
According to these data, around 2.2 percentages of the all transfer opportunities are used (by at least 1 passenger). Without using an upper bound, this ratio is very low in our solution, namely around 0.3 percent. Although the non-zero elements of the transfer opportunities were not involved in the characterization of the transfer flow we should find a solution where this ratio is more realistic.

On the other hand the solution contains very high elements. It means if there is a transfer opportunity used by somebody it is used by many passengers (~100). This is not realistic either. As an extra constraint we limited the variables ($t_{ij} \leq c$) (upper bound). If *c* is 8-10 we get the proportion of non-zero elements about 2%.
(This constraint is improvable, if there are some very popular transfer there can be used higher c_{ij}

The results in table 4 suggest that might be a lot of different solutions with the same objective value for an instance. In this case we might improve the model with some extra constraints (such as the upper bounds), in order to avoid unrealistic solutions, without increasing the objective value.

Validation graphs

First let us take a look at figure 2. This graph was created by historical data. Our goal was not to reach this very detailed connection time distributions but comparing this one to the solution can be good for validation of the results. X-axis of all graphs is arrival time in the rest of the thesis.

* Confidential *

Figure 2: historical data for validation

On the figure 3 we can see the same distributions in the optimal solution from our model.

Figure 3: same connection time distribution in the solution

We can compare the two graphs. It also should be kept in mind that the first graph was created by data of many days, that is a kind of average. Small differences are acceptable for an individual day.

So far we have paid attention just for the proportion of the short connections (or any other interval e.g. 120-240, etc.) and all transfers. Optimal solution from the point of view of absolute numbers of transfers can be seen on figure 4 (every graph corresponds to our optimal solution).
We can compare it to the reality (Figure 4b: historical data).

* Confidential *

Figure 4

The same graph according to historical data:

* Confidential *

Figure 4b: historical data according to the results on the figure 4. (avarage of many days)

Figure 5: short connections at Schipol during the day

These graphs are very significant. These transfers have great importance because those have to be handled in short period. On figure 5 we can see the estimated peaks from the point of short connections . In the reality these peaks exist as well. Unfortunately we do not have detailed enough historical data to show the corresponding graphs in reality.

On the figure 6 shows all the EUR-EUR transfers in absolute values. That is also important how spread the transfers anyway, even if those ones are not short connections.

* Confidential *

Figure 6: connection time distributions in absolute values (in the optimal solution)

It is worth to see together figure 6 and 6b. Former belongs to our solution, latter comes from historical data.

* Confidential *

Figure 6b: historical data according to the results on the figure 6. (avarage of many days)

Figure 7: transfers at Schipol during the day (in the optimal solution)

Unfortunately we do not have detailed enough historical data to show the corresponding graphs in reality.

Conclusion

Every instance can be solved for some values of $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$. For some instances some of these values become too big, which results in unrealistic solutions.

The solution is not unique. The problem seems very underdetermined. We should use some independent extra constraints.
The solver gives us an extreme solution. It is called extreme because of two reasons.

First, the number of transfer possibilities which are unused is too high. Second, if some possibility is used there are unrealistic many transfer passengers. In practice, it means, for instance, there is an arriving flight from Stockholm and there are transfer passengers just to Moscow. Nowhere else but 54 to Moscow. Although all the transfers together satisfy the desirable connection time distributions and all the constraints, the solution is not realistic enough.

We should limit the number of the solutions by skipping the unrealistic ones. There are basically two ways how to do this in LP. Either we add extra constraints since we get a smaller, hopefully better solution set or we use a more complicated goal function to pick a more realistic solution.

In order to control the ratio of zero elements in the optimal solution we used upper bound as an extra constraint. This causes longer computation time. This one needs 4 minutes excess, making the total computation time (at most) 5 minutes.

4. Changes in the model

Problems

As concluded in the last chapter, the current solution is not realistic enough. So far we have realized basically two problems.

1. The problem of infeasibility.
2. Even if a schedule is feasible the group size distribution is quite different to the reality.

Transfer group means a group of passengers taking the same transfer. In our model it is one element of the solution. We can count how many group contains 1, 2, 3 etc.. passengers. Hence we get the group size distribution. The transfer group size distribution in the reality can be seen on table 5:

* Confidential *

Table 5: group size distribution in the reality

Without upper bound the solution is different to this:

* Confidential *

Table 6: group size distribution in the solution

Unfortunately, even if we use upper bound the solution is not similar to the real distribution.

* Confidential *

Table 7: group size distribution in the solution (with upper bound)

How to improve the model?

There are many ways to improve the model. The following changing is just a possible next step.

Our main purpose with these improvements is to solve the problem of infeasibility. Since we introduce more constraints by the extra variables we hope the group size distribution is going to be more realistic as well. If this problem remains we have to change the model again.

Instead of (vi) and (vii) constraints let's create more difficult goal function.
(i)- (v) Constraints can remain the same.
In the goal function:

Instead of $\sum_{k=1}^{4}\sum_{l=1}^{6} d_{k,l}$:

$$\sum_{k=1}^{4}\sum_{l=1}^{6} d_{k,l} + w_1\left|P'(EU-EU) - P(EU-EU)\right| + w_2\left|P'(ICA-EU) - P(ICA-EU)\right|$$

In this case we solve this problem:

Determine, within a few seconds of PC computation time, the number of transfers between each arrival, element from *A*, and each departure, element from *D* such that the following conditions are satisfied exactly[2]:

- For each arrival, the number of transfer pax is less than or equal to the number of passengers.
- For each departure, the number of transfer pax is less than or equal to the number of passengers.
- No connections shorter than MCT.
- For each set of arrivals A(i), the total number of transfer pax is equal to the total number of passengers in that set multiplied by T(i).
- For each set of departures D (i), the total number of transfer pax is equal to the total number of passengers in that set multiplied by T (i).
- (Extra: number of every individual transfer less then a constant)

Whereas we look for a best fit to the distributions of transfer times and a best fit for:

- The sum of all transfer passengers from Europe continuing her or his journey to Europe, divided by the total number of transfer pax arriving from Europe satisfying P (EUR to EUR).
- The sum of all transfer passengers outside from Europe continuing her or his journey to Europe, divided by the total number of transfer pax outside from Europe satisfying P (ICA to EUR).

This one is a little bit different then the original problem definition but maybe it leads to more realistic solution.

Mathematical description of these changes

[2] For some particular schedules the constraints have to be relaxed.

Instead of the following constraints there is a new extra part in the goal function.

$$\frac{\sum_{i\in K}\sum_{j\in L} t_{ij}}{\sum_{c\in C} T(c)\times(\sum_{i\in M} N_i^a)} \le P_{EUR-EUR} + \varepsilon_3 \text{ And } \frac{\sum_{i\in K}\sum_{j\in L} t_{ij}}{\sum_{c\in C} T(c)\times(\sum_{i\in M} N_i^a)} \ge P_{EUR-EUR} - \varepsilon_3$$

where $K = \{i : I_i^a = EUR\}$, $L = \{j : I_j^d = EUR\}$, C is the set of the categories from EUR (vi)

(KLM-EUR, Rest-EUR...etc.) and $M = \{i : C_i^a = c\}$

This thus states that the percentage of transfer passengers arriving on a flight from within Europe that are connecting to a flight with a destination in Europe should be approximately equal to *P(EUR to EUR)*.

$$\frac{\sum_{i\in K}\sum_{j\in L} t_{ij}}{\sum_{c\in C} T(c)\times(\sum_{i\in M} N_i^a)} \le P_{ICA-EUR} + \varepsilon_4 \text{ And } \frac{\sum_{i\in K}\sum_{j\in L} t_{ij}}{\sum_{c\in C} T(c)\times(\sum_{i\in M} N_i^a)} \ge P_{ICA-EUR} - \varepsilon_4$$

Where $K = \{i : I_i^a = ICA\}$, $L = \{j : I_j^d = EUR\}$, C is the set of the categories from ICA (vii)

(KLM-ICA, Rest-ICA...etc.) And $M = \{i : C_i^a = c\}$

This constraint is obtained in the same manner as constraint (vi) and stated that the percentage of transfer passengers arriving on a flight from outside Europe that are connecting to a flight with a destination in Europe should be approximately equal to *P(ICA to EUR)*.

The extra part in the goal function:

$$w\left|P_{eu-eu} - P'_{eu-eu}\right| + w\left|P_{ic-eu} - P'_{ic-eu}\right|$$

Actually, we minimize $\varepsilon_3, \varepsilon_4$ those were in the Constraint (vi)-(vii) in the previous version.

If the solution is far from the real probabilities there is a kind of penalty.

The importance of these probabilities can be set by changing the weights (w).

Mathematically, this extra part implies few extra inequalities (because of the absolute values, we apply the same trick like before):

We need two more extra variables called d_{eu-eu} and d_{ica-eu}.

Then

$$d_{eu-eu} \geq \frac{\sum t_{ij}}{E(eur)} - P_{eu-eu} \text{ And } d_{eu-eu} \geq P_{eu-eu} - \frac{\sum t_{ij}}{E(eur)} \quad \text{(xiv)}$$

The intercontinental case:

$$d_{ica-eu} \geq \frac{\sum t_{ij}}{E(ica)} - P_{ica-eu} \text{ And } d_{ica-eu} \geq P_{ica-eu} - \frac{\sum t_{ij}}{E(ica)} \quad \text{(xv)}$$

It counts two more rows in the matrix form.

The goal function has changed. Our purpose:

Minimize $\sum_{k=1}^{4}\sum_{l=1}^{6} d_{k,l} + w_1 d_{eu-eu} + w_2 d_{ica-eu}$

Other improvement

The vast majority of the flights belong to KLM-Eur or KLM-Ica category. By these categories we require relative big accuracy. The constraints related to these categories are strict. We use $\varepsilon_{1,klm} = 0.01$. By other cases we have bigger freedom. Here $\varepsilon_{1,all} = 0.03$.

It is a bit changing but it also expands the solution set.

Test results of the new version

The original goal (to fit the connection time distributions) is granted at least as much good as in the previous version. The first aim of the changes is to avoid the infeasibility.
By changing the weights we can set this penalty according to which element is important for us in the goal function.

The first day was infeasible with the previous method. Now we get the following results:

Upper bound	Ratio of non-zero elements	$\lvert P_{eu-eu} - P'_{eu-eu} \rvert$	$\lvert P_{ic-eu} - P'_{ic-eu} \rvert$	weights	Objective value
50	0,005	0,11	0,032	6,6	1,008
50	0,005	0,11	0,032	12,12	1,874
50	0,005	0,13	0,0036	18,18	2,728
15	0,014	0,11	0,032	6,6	1,008
15	0,014	0,11	0,032	12,12	1,874
15	0,014	0,13	0,005	18,18	2,728
8	0,024	0,11	0,032	6,6	1,008
8	0,024	0,11	0,032	12,12	1,874
8	0,024	0,122	0,019	18,18	2,728
50	0,005	0	0,169	36,6	1,643
50	0,005	0	0,169	24,12	2,634
50	0,005	0,066	0,087	18,12	2,404
50	0,005	0,11	0,032	14,12	2,097
50	0,005	0,066	0,087	16,12	2,27
50	0,005	0,066	0,087	15,12	2,201
50	0,005	0,066	0,087	14.5,12	2,153

Table 8: test results (first day)

In the second part of the table we can see how to influence the probabilities by changing the weights.

There were other two days under examination with the first method. In those cases the second method works as well.

Upper bound	Ratio of the non-zero elements	$\left\|P_{eu-eu} - P'_{eu-eu}\right\|$	$\left\|P_{ic-eu} - P'_{ic-eu}\right\|$	weights	Objective value
50	0,005	0	0	6,6	0
50	0,005	0	0	12,12	0
50	0,005	0	0	18,18	0
15	0,012	0	0	6,6	0
15	0,012	0	0	12,12	0
15	0,012	0	0	18,18	0
8	0,022	0	0	6,6	0
8	0,022	0	0	12,12	0
8	0,022	0	0	18,18	0

Table 9: test results (9. day)

Here we could fit everything very accurate. The only problem is the distribution of transfer groups. Our intention in the future is to make the solution closer to that distribution.

Upper bound	Ratio of the non-zero elements	$\left\|P_{eu-eu} - P'_{eu-eu}\right\|$	$\left\|P_{ic-eu} - P'_{ic-eu}\right\|$	weights	Objective value
50	0,005	0,002	0	6,6	0,0662
50	0,005	0,002	0	12,12	0,0787
50	0,005	0,002	0	18,18	0,0912
15	0,013	0,002	0	6,6	0,0662

15	0,013	0,002	0	12,12	0,0787
15	0,013	0,002	0	18,18	0,0912
8	0,022	0,002	0	6,6	0,0662
8	0,022	0,002	0	12,12	0,0787
8	0,022	0,002	0	18,18	0,0912

Table 10: test results (23. day)

The objective values are repeating themselves. It was quite suspicious. Maybe the solutions are the same for all weights.

To find out what changes if we change the weights we measure the difference between solutions:

$$\sum_{i,j} \left| t_{ij} - t'_{ij} \right|$$

Another kind of measure of the difference is that how many transfers just differ in two solutions. The former is better because that pays attention for the absolute value of the difference of the transfers as well. For instance if the Budapest-Amsterdam (7:30) → Amsterdam-Madrid (9:15) transfer is used by 15 passengers in the first solution and used by 1 passenger in the second then it means 1 difference according to the latter measure and 14 difference to the former. If this transfer opportunity is used by 1 passenger in the first solution and by 2 passangers by the second one, it means 1 difference according to both of measues.
If we do not change the upper bound (50) then about 600 elements are non-zero in the different solutions. Usually about 400 transfers are different in different solutions. The typical number for $\sum_{i,j} \left| t_{ij} - t'_{ij} \right|$ is about 2000.

We can conclude the solutions are different when using different weights.
If we change the upper bound we get quite different solutions but it is not surprise at all. (Feature of the simplex method)

The computation time doesn’t change compered to the previous version.
We can evaluate the solutions also using similar graphs as in chapter 3.

* Confidential *

Figure 10: connection time distributions in absolute values

* Confidential *

Figure 11: short connections

* Confidential *

Figure 12: connection time distributions (eu-eu transfers)

* Confidential *

Figure 13: connection time distributions (different weights in the goal function)

Figure 14: connection time distributions (different weights in the goal function)

All these graphs belong to solutions. The graphs about the historical data can be found in the chapter 3. The connection time distribution were quite similar to each other using by different weights in the goal function. The point of using different weights in the goal function was setting the desireable $d_{eu\text{-}eu}$ and $d_{ica\text{-}eu.}$

We are not satisfied enough with the solution. Now we divide the day by two parts. We try to satisfy the connection time distributions in both parts. Hopefully this will lead to a more realistic solution.

5. Splitted version

The model

Our purpose is to put more exact connection time distribution to the goal function. So far we have tried to fit these distributions:

* Confidential *

Table 2: connection time distributions in the reality

Now we try to fit this for the transfer passengers with the arrival time before noon and for the transfer passengers with the arrival time after noon as well.

Mathematically it means:

We need
- X^{am}_{eu-eu} : expected number of transfer pax Eur-Eur with the arrival before noon
- X^{am}_{eu-ica} : expected number of transfer pax Eur-Ica with the arrival before noon
- X^{am}_{ica-eu} : expected number of transfer pax Ica-Eur with the arrival before noon
- $X^{am}_{ica-ica}$: expected number of transfer pax Ica-Ica with the arrival before noon
- X^{pm}_{eu-eu} : expected number of transfer pax Eur-Eur with the arrival after noon
- X^{pm}_{eu-ica} : expected number of transfer pax Eur-Ica with the arrival after noon
- X^{am}_{ica-eu} : expected number of transfer pax Ica-Eur with the arrival after noon

- $X^{am}_{ica-ica}$: expected number of transfer pax Ica-Ica with the arrival after noon
- extra 24 variables for the before noon transfer time distributions $d^{am}_{k,l}$ *k=1..4, l=1..6*
- extra 24 variables for the after noon transfer time distributions $d^{pm}_{k,l}$ *k=1..4, l=1..6*

  These variables will measure the difference between the real distribution and the solution.
- the followings inequalities:

$$\frac{\sum_{\substack{EUR-EUR \\ waiting\_time<60 \\ arrival<12:00}} t_{ij}}{X^{am}_{eu-eu}} - X'^{am}_{eu-eu} \le d^{am}_{1,1} \text{ and } X'^{am}_{eu-eu} - \frac{\sum_{\substack{EUR-EUR \\ waiting\_time<60 \\ arrival<12:00}} t_{ij}}{X^{am}_{1,1}} \le d^{am}_{1,,1}$$

This implies:

$$\sum_{\substack{EUR-EUR \\ waiting\_time<60 \\ arrival<12:00}} t_{ij} - d^{am}_{1,1} \cdot X^{am}_{eu-eu} \le X'^{am}_{eu-eu} * X^{am}_{eu-eu}$$

and

$$\sum_{\substack{EUR-EUR \\ waiting\_time<60 \\ arrival<12:00}} t_{ij} - d^{am}_{1,1} \cdot X'^{am}_{eu-eu} \le (-0.17) * X'^{am}_{eu-eu}$$

Similar inequalities for the other extra variables.

The new goal:

$$\text{Minimize } \sum_{k=1}^{4}\sum_{l=1}^{6} d^{am}_{k,l} + \sum_{k=1}^{4}\sum_{l=1}^{6} d^{pm}_{k,l} + w_1 d_{eu-eu} + w_2 d_{eu-ica}$$

Test results

We expect this makes the solution less extreme (more non-zero elements, less high value)
In table 11 the transfer groups distribution in the previous solution (9. day, upper bound=8) are listed. The ratio of the non-zero elements was 2.2%.

* Confidential *

Table 11: group size distributions in the previous version

In table 12 the results of the splitted version are listed. The ratio of the non-zero elements did not change.

* Confidential *

Table 12: group size distributions in the splitted version

The computation time did not change as well compared to the previous version.

The solutions are very similar to the previous version. There is no significant difference in the transfer groups distributions (little bit better then before) between the two solutions. The historic group size distributions can be found in table 5.

Graphs for validation

* Confidential *

Figure 14

It was the basic validation graph. It also looks acceptable compared to the real one (chapter 3) like in the previous versions but we improved the model because of other problems.

* Confidential *

Figure 15

* Confidential *

Figur16

Unfortunately this changing did not help. Although the validation graphs are a little bit closer to the reality, the group size distribution problem still remains. It is not that big surprise because we have not done anything directly for that. That is a bad luck that simplex chooses solution from the edges. Finally we try to add an extra part into the goal function to find a solution with better group size distribution.

6. Final version of the LP-formulation for the problem of transfer passengers

This chapter is going to talk about how we solved the last obstacle to get realistic solution. Namely none of the previous versions provided realistic solution considering the group size distribution. The only difference is an extra part in the goal function. To remind what we use from the former models here is the complete description of the final version of the LP-formulation.

Constraints

- $\forall i, j \quad t_{ij} \geq 0$ (i)

-For each arrival, the number of transfer pax is less than or equal to the number of passengers. (ii)

-For each departure, the number of transfer pax is less than or equal to the number of passengers. (iii)

-For each set of arrivals A(i), the total number of transfer pax is approximately equal to the total number of passengers in that set multiplied by T(i). (iv)

-For each set of departures D(i), the total number of transfer pax is approximately equal to the total number of passengers in that set multiplied by T(i). (v)

Goal function

It is a summation of three parts.

First part (our main goal):

The difference between the real connection time distributions and the connection time distributions in the solution.

$F(t) = \sum_{k,l} \left| \; X_{k,l} - X'_{k,l} \; \right|$ If the real value is $X'_{k,l}$.

The X_{ij} values are the corresponding values from the solution. According to our notation in the previous chapters:

$\sum_{k=1}^{4} \sum_{l=1}^{6} d_{k,l}$

Second part

Originally it was in constrain form. We put it in the goal function because of the infeasibility. Namely some particular schedule was not solvable. After these corrections every schedule became feasible.

The difference between the real P(eur-eur) probability and the same probability in the solution.
The difference between the real P(ica-eur) probability and the same probability in the solution.

$w_1 d_{eu-eu} + w_2 d_{ica-eu}$

We were able to find solutions those are really close the desirable solution in sense of connection time distributions and P(eur-eur), P(ica-eur) values. Other features of the solutions were not realistic enough (number of nonzero elements, transfer group distributions)

That is why we put the extra constraint $\forall i, j \quad t_{ij} \leq 50$ and the third part of the goal function.

Third part

This part is going to be in order to get a solution with better group size distribution. It is to be regretted that is difficult to write this goal in linear form. There is no linear form to measure the difference between the distribution of the solution and a desirable distribution. It is not linear function of the variables. That is why we have to do it intuitive way.
It is easier to see if the decision variables have just one index. Let's convert t_{ij} into x_i then the additional part:

$w_3 \sum_j \left| \sum_{i=0}^{10} x_{j\times 20+i} - c_l \right|$ Where the weight is very small (~10^{-6}) because we just want to pick more realistic solution, we do not want to destroy the previous remarks.

The idea of this formula is to abuse the property of simplex method. For instance, if we take some penalty for the difference between the sum of 10 transfer and c_1=1, simplex-method will prefer solution where one of the 10 elements is equal to 1 and others are 0. (If this solution is possible. We suspect that there are many solutions including this kind of solution) this argument is not necessarily true. We hope it is going to be a good influence on the group size distribution.

We don't involve all the variables in this formula. Randomly we choose 10 elements from every 20 elements. It gives a kind of degree of freedom and it is also preferable from the point of view of the computation time.

C_l are suitable constants. According to real group size distribution we set c_i.

* Confidential *

Table 5: group size distribution in the reality

In the third par to of the goal function the constant are set the following:

* Confidential *

We do not want to reach completely the same distribution. Our goal is to get more realistic solution.

There is another problem with this part. Any absolute value in the goal function means extra variables because we have to convert it in linear form. New variables make the problem turn into higher dimension. Hence the computation time grows. The number of the necessary new variables in this formula is proportional to the number of the transfer opportunities. Sometimes it leads to extremely long computation time.

Results of the final version

Group size distribution in the solution

Although we had to pay for the improvements in computation time (after optimization of the program it takes approximately 40 minutes) our intuitive guess was successful. The group size distributions become much more realistic. Although there are still too many big transfer group compared to the reality the vast majority of the transfers is small transfer similarly to the real distribution.

Don’t forget that this group size distribution problem was our last obstacle. We had reached all our main goals before. Because of the small weights we still have the good remarks related to the connection time distributions and validation graphs.

The groups size distribution in the latest version:

* Confidential *

Table 13: group size distribution in the final version

Integers in the solution

This estimation of transfer passengers is a part of a bigger project called SPECS. There is a specific engine that requires integer estimations. Our solver doesn't provide absolutely integer solution. It doesn't theoretically. Because of the constraints and the goal function we got almost fully integer solution.
Around 90% of the elements of the solution are integer!

* Confidential *

Table 14: integers in the final solution

There are non-integer elements in the solution. We can replace these elements with the closest integer value.

Variations in the solutions

There is randomness in the third part of the goal function. We can choose other 10 elements in every 20 elements for the penalty. That's why can get different solutions. These solutions are more or less the same in our main goals (connection time distributions, ratio of non-zero elements, group size distributions). Despite of this fact these solutions belong to different validation graphs. Here we present two different versions. Finally we can choose the solution with better validation graphs.

Validation graphs by different solutions

1. Version

* Confidential *

Figure 17

2. Version

* Confidential *

Figure 18

We can compare these graphs to the reality again.

Historical data:

* Confidential *

Figure 18b: historical data for validation

The same in absolute value

1. Version

* Confidential *

Figure 19

There are differences between the versions but the transfer peaks appear in both cases in the morning and in the evening. One of the main purposes of this project is to predict these peaks. In the reality these peaks exist and now we got estimations about them.

2. Version

* Confidential *

Figure 20

Short connections

1. Version

* Confidential *

Figure 21

The importance of these graphs about the short connections is essential. These transfers have priority in the baggage process. Since all the other parameters of the transfer flow are fulfilled hopefully these estimations are correct.

2. Version

* Confidential *

Figure 22

7. Conclusions

The original problem was to forecast the number of transfer passengers at Schipol airport. In order to get estimation for the transfers we used Linear Programming model.

Since the first version of the model did not provide realistic solution we had to think the problem definition over. All the time we have been trying to find a proper problem definition that implies realistic solution. The model is very flexible. It is easy to add or erase constraints or extra parts to the goal function.
If there are too many or too strict criteria we can make the problem over determined and we can face the problem of infeasibility. If there are not enough constraints thousands of solutions can appear. This way the problem is underdetermined. We had to find the balance between these two problems. That is also possible we simply forget about important features of the transfer flow. Then the solution is not going to be realistic.
Whenever we add new constraints or extra variables we have to keep in mind we have to pay for it in the computation time.

Taking all into consideration we created basically four different versions. The first version was unusable. Many cases we found schedules infeasible. Though it is really fast where it works. The computation time is less then 1 minute. The tests were done on a laptop. (Pentium IV)
In the second version we could solve the infeasibility but the individual transfers had not realistic distribution. It could happen because we had never defined anything before that was related to the group size distribution. Here we converted few constraints to extra part in the goal function.
Neither the splitted version solved this problem. This version had more detailed goal function at the desirable connection time distributions.

These versions satisfied the original problem definition and were still quite fast (less then 5 minutes computation time) but we were looking for better solution.

The final version was found the best of all although it is not good enough. The computation time is very high. It is about 40 minutes to get a solution in an average personal computer.
There was another extra part applied in the goal function. It solved the group size distribution problem. The original problem seems underdetermined so the solution is not unique. The estimation for whole transfer flow seems realistic by historical data. Finally we have an algorithm that provides more or less realistic forecasts for the peaks, – the critical periods of the day – the short connections and the traffic of the entire day.

Recommendations for future research

1. The third part of the goal function is quite arbitrary.
   It can be changed.

2. In order to get smaller problem it can be useful to write an algorithm to skip connections those are not used.

3. If we can reduce the computation time (by the 2. point) we can provide solution as combination of different solutions.