Mark Irwin

Short-term Revenue Forecasting at KLM

Master thesis, defended on August 27, 2010

Thesis advisor: Dr. F.M. Spieksma

Mastertrack: Applied Mathematics

Mathematisch Instituut, Universiteit Leiden

Contents

Chapter 1

Introduction

What would we do if we could see into the future? Would we really want to know what is going to happen, or not? In our personal lives we would perhaps choose not to know, since this might affect the way we lead our daily life. But in business knowing future events can be very valuable. For instance, if a salesman would know that Holland will win the 2010 Football World Cup, he could start manufacturing all sorts of memorabilia, which he would then be able to sell once the moment arrives. Unfortunately, at this moment there is no way of predicting the future with absolute certainty. The aforementioned salesman could however take his chances and manufacture the memorabilia anyway. If however the unthinkable happens and Holland does not win the World Cup, he will be left with worthless junk.

Forecasting future events is common practice within every large firm and therefore also within KLM Royal Dutch Airlines. Just as in the example of the salesman KLM has to take strategic decisions based on predictions of the future. Accurate forecasts of customer demand, luggage demand, oil prices, fleet maintenance costs etcetera will all help to determine the optimal strategy for the company.

This thesis will focus on short-term revenue forecasting within KLM. Within the Revenue Management department of KLM two types of revenue forecasts are made, long-term and short-term. Every year at the end of March the budget for the upcoming fiscal year, which runs from April to the following March, is determined. This is the so-called *long-term revenue forecast*. During the year at the end of each month a revenue forecast is made for the following month, two months and three months. These are called *short-term forecasts*.

In the last few years the Decision Support team AMS/RP at KLM have developed tools that have made it possible to analyze the enormous amounts of data on a very detailed level. These tools have given rise to the main question of this thesis:

> **Given the very detailed data available is it possible to improve the accuracy of short-term forecasting within KLM?**

Improving accuracy is not the only goal though. Currently revenue forecasting at KLM is done by experts who base a large part of their forecast on experience and market knowledge. This makes it difficult for others to gain a good understanding of how a forecast is made. Also, this makes KLM very dependent on these experts. If for some reason an expert would not be able to make the forecasts anymore, it would be difficult for a newcomer to make the forecasts simply because he does not have the same experience and knowledge. Therefore the second question of this thesis is

> **Is it possible to develop a standard forecasting method that also gives a clear insight into how a forecast is built up?**

In order to answer these questions a basic knowledge of the structure of the KLM is needed. A short explanation of this structure is given in Chapter 2. Also the concept of Revenue Management and current methods used for short-term forecasting is treated in this chapter. As many new expressions will be introduced a list of definitions is given at the end of this thesis.

The art of forecasting is not something new and several methods are already available. Chapter 3 will discuss a couple of the existing methods that have been examined. Each method will be explained as well as the suitability for our project. Finally advantages and disadvantages of each method will be discussed.

Chapter 2

KLM

2.1 General information

With nearly 30.000 employees it is fair to say that KLM is a large commercial airline company. Although many people associate working for KLM with flying, the actual business of flying is done by a minority of these employees. Roughly speaking KLM can be divided into 6 divisions, namely Commercial, Inflight Services, Operations, Ground Services, Cargo and Engeneering&Maintenance. Of these divisions only Operations consists of the pilots and flight crews flying the planes.
Determining ticket prices and whom to sell these to is done within the Pricing&Revenue Management department, which is a part of the division Commercial.

2.1.1 Revenue Management

"Revenue management is the art and science of predicting real-time customer demand at the micromarket level and optimizing the price and availability of products"[1].

Revenue management deals with maximizing income when the number of products is fixed in advance. Perhaps without even recognizing it this is something we encounter on a regular basis in our daily life. High rates for hotel rooms during national holidays, weekend evening rates for car parking and discount rates of the National Railways during off-peak hours are just a few examples. In the airline industry the products on sale are seats. Airline companies try to maximize their income by selling similar seats to different types of customers for different prices. The main factor for enabling different prices for the same seat is by changing sales conditions. For example, low-priced tickets have to be bought long in advance, the traveler should stay a weekend at the destination, and the ticket cannot be changed. High-priced tickets have no such conditions.

The great difficulty in revenue management is predicting the demand for the product on sale and the price potential customers are willing to pay for that product. At KLM in general there are two products for sale, i.e. economy class seats and business class seats. For each product differentiating the conditions gives rise to subclasses. Each subclass has its own set of conditions and also its own price range. The mechanisms that determine the sales price will not be discussed in this thesis, because they are not relevant for the project. However it is useful to have a general idea of the problems arising from unknown demand. This will be done using some simple examples.

Suppose flight KL0000 has 100 seats available for sale. There are only two subclasses, A and B. Bookings for these seats can be made right up until the departure time of the flight. However KLM has the possibility of accepting or rejecting bookings. Three different cases will be examined. In each case the price for a ticket for subclass A has been set at € 50 and for subclass B at € 100. Also demand for subclass A is known to be infinite. However demand for subclass B is different in these three cases. In the first case the demand

Subclass	Case 1	Case 2	Case 3
A	∞	∞	high
B	150	80	low

Table 2.1: Demand per subclass

for subclass B is known to be 150 and maximizing revenue is simple. Only bookings for subclass B should be accepted and the revenue will be 100∗€ 100,– = € 10000,–. In the second case the demand is known to be 80 and the optimal strategy is still more or less the same as in the first case. Accept all bookings for subclass B and only the first 20 bookings for subclass A. The revenue will now be 80∗€ 100,– + 20∗€ 50,– = € 90000,–. In both these cases the optimal strategy is straightforward due to the fact that demand is known in advance. Unfortunately this is not the case in real life. The third case is an example of this. Now all that is known about the demand is that demand for subclass A is high, but that it is not unlimited. For subclass B even less is known about the demand, only that demand is low. The problem that arises now is which bookings to accept and which to reject. Bookings for subclass B will always be accepted if possible, but how many seats should be reserved for these bookings? Rejecting too many bookings for subclass A might result in unoccupied seats. This phenomenon is called *spoilage*. These seats could have been sold, but the rejection policy was too strict. On the other hand rejecting too few bookings for subclass A could result in filling the aircraft too soon. This could mean that bookings for subclass B will be rejected, because simply no seats are available any more. This is called *spillage*. Customers that were willing to pay a high amount have to be rejected, because lower paying customers were accepted at an earlier stage. Both spillage and spoilage will result in lower revenue than the maximal revenue possible. A complicating factor is that in general customers who are willing to pay a high amount book later than those who are not.

In general, flights at KLM have a lot more subclasses available than two. Also, within a subclass prices may vary for different type of customers. For instance, different countries may charge different prices for the same product. This is simply due to the fact that in a certain country people might be willing to pay more than in another country. The process of accepting or rejecting bookings, called *inventory steering*, is therefore very complex and accurate demand forecasting is an absolute necessity to minimize spillage and spoilage.

An important technique in generating maximum revenue is the possibility of *overbooking*. Experience shows that every flight encounters a number of cancelations and no-shows. These no-shows can be passengers who arrive late for their flight or simply never show up at all. If one assumes that these cancelations or no-shows will occur for a certain flight it can be profitable to sell more tickets than there are seats available for that flight. This is the so-called technique of overbooking. Of course, passengers that have a reservation and can not board the airplane due to overbooking must be compensated by KLM. Accepting too many bookings is therefore not advisable.

2.1.2 Flight information

The revenue management department at KLM is divided into three separate revenue groups, namely Europe (from now on abbreviated as RU), North-America (RW) and the rest-of-the-world (RV). This means that each revenue group consists of all destinations in that region. For example, Rome, Aberdeen and Stockholm are all in revenue group RU, since they all lie in Europe. Also Lima, Tokyo and Cape Town are in the same group, namely RV. These destinations are referred to as sublines. A subline consists of all flights between Amsterdam and a certain destination, e.g. all flights from Amsterdam to Rome and vice versa. The way in which destinations are grouped together to form a higher level is called a *flight hierarchy*. Figure 2.1 illustrates a small part of the KLM flight hierarchy. In total there are 116 sublines and 23 complexes.

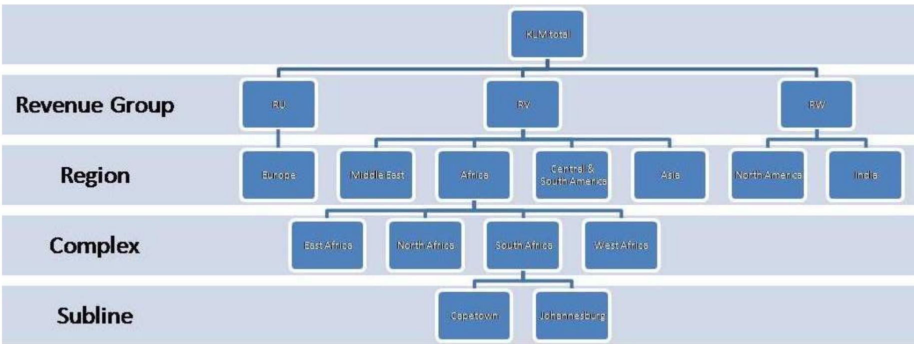


Figure 2.1: KLM flight hierarchy

A single flight is referred to as a flight leg. As it is not always possible to get from one origin to another destination multi-leg flights will sometimes be necessary. These flights also have the same flight number. For instance if a passenger wants to fly from Amsterdam to Taipei this is only possible by making a stop at Bangkok. The route Amsterdam-Taipei is called a segment, which consists of the flight legs Amsterdam-Bangkok and Bangkok-Taipei, both of them having flight number KL0877. For this reason the number of flight leg passengers and segment passengers in a revenue group may differ. A passenger is counted for each individual flight leg, but only once for the higher levels. In the above example the passenger would be counted twice for each flight leg, but only once for the segment Amsterdam-Taipei. It is important to note that this is different from the case of a passenger who wants to fly from Rome to Atlanta. Because there is no direct flight available this passenger will have to fly from Rome to Amsterdam first and then from Amsterdam to Atlanta. These are also two flight legs, but these two flights have different flight numbers and are a part of different sublines. The passenger is therefore counted once for segment Rome-Amsterdam and once for Amsterdam-Atlanta. In general, the price this passenger pays for a ticket from Rome to Atlanta is less than if he had bought two separate tickets, one for Rome-Amsterdam and one for Amsterdam-Atlanta. This is why his so-called *traffic type* is stored. The main distinction made in traffic type is local traffic or connecting traffic. Also a passenger's *True Origin & Destination* is stored.

Besides a passenger's flight information a lot more information is known. Two key elements of information that will frequently be used in this thesis is a passenger's point of sale (PoS) and his subclass.
The PoS is simply the country where he bought his ticket. This is an important factor because different countries have different prices for tickets. Similar to the flight hierarchy there is a *PoS-hierarchy* that determines how different PoS's can be grouped together. The highest level in the PoS-hierarchy is called PoS-all. Countries may have a different currency than the euro, the currency that KLM works with. These different currencies will be subject to fluctuations in the rate of exchange (RoX).
A passenger's *subclass* is denoted by a single capital letter and refers to the conditions under which he bought his ticket. When purchasing a ticket at KLM there are two types of tickets available at KLM, Business Class tickets and Economy Class tickets. These types determine a passenger's *cabin* in the aircraft. Cabin C relates to business class tickets and cabin M to economy class tickets. Within these cabins the different subclasses are defined. Examples of these subclasses are 'K', 'Z', 'E' and 'V'. Subclass 'C' also exists and is a subclass within the C cabin.

2.1.3 Decision Support Tools

Several decision support tools have been developed at KLM in the last few years. Two tools that play a major role in the process of revenue forecasting are: Monet and DeLorean.

Monet Monet contains all realized revenue data of flown months. Around the third Saturday of each month the revenue data of the previous month becomes available. This data holds among others the following details

- Flight number
- Flight date
- Point of Sale
- Subclass
- True Origin and Destination.

For each of these combinations the following information is available

- Revenue
- Passengers
- Available seats.

From this information multiple other important facts can de derived. Also Year-over-Year (YoY) indices, current year numbers divided by previous year numbers, are easily viewable in Monet.
All information is based on flight details not passenger details. As a result, if a passenger flies from Rome to Atlanta his details (PoS, subclass, true O&D) will be found on a flight from Rome to Amsterdam and a flight from Amsterdam to Atlanta. It is however not possible to tell if those details belong to the same person.

One important extra element of information available in Monet is a passenger's ticketing month. This is the month in which he actually paid for his ticket, which is not necessarily the month in which he flies. There are two possibilities for making a RoX correction in Monet. The first method is to use the previous year's RoX for passengers that flew this year. This means that if a passenger's ticketing month is November 2007, the revenue earned from this passenger is not calculated using the RoX of November 2007, but of November 2006. The second method is to use this year's RoX for last year's passengers. The goal of both these methods is to compare year over year revenues without the influence of a changed RoX.

DeLorean DeLorean is a forward booking application. In general, bookings for flights at KLM can be made up to 340 days in advance. Every day of the year a snapshot is made of all KLM flights. The snapshot contains two important facts about every flight. These facts are the number of booked passengers and the scheduled number of seats, the *capacity*. The current data holds among others the following details

- Flight number
- Flight date

- Point of Sale
- Subclass
- True Origin and Destination
- Passenger type (Individual/Group/Duty)
- Award

For each of these combinations the number of bookings and capacity is available. Flights that have already flown in the past are also included in the snapshot. For these flights the so-called *day-before-departure* (DBD) snapshot is stored. This is the number of bookings and the scheduled capacity for a flight as it is one day before the scheduled departure date. It is very important to note that this might differ considerably from the actual flown passengers and capacity. In general, actual flown passengers numbers will be lower than the DBD number of bookings. Several reasons can cause this *drop-off*. Some examples are passengers that do not show up at the flight, passengers that cancel their flight at the last minute or passengers that miss their connection. Also passengers might be requested to change flights due to overbooking. Of course, new bookings can be made on the day of departure as well. This increase in passengers due to new bookings is usually significantly lower than the drop-off.

From this detailed information several other points of information can be derived. Namely,

- Booked Load Factor
- Bookings index(bookings cy/bookings py)
- Scheduled Capacity index (scheduled seatkm cy/scheduled seatkm py)
- Days before departure.

Actual flown passenger and capacity totals are also available in DeLorean, co-called *SLS-information*. However, for these totals the only details available are flight number, flight date and cabin. Therefore the actual passenger and capacity are mostly studied in Monet where more detailed information is available.

Both Monet and DeLorean use the same flight hierarchy and PoS-hierarchy. The flight hierarchy consists of 8 levels. The most important levels are given in table 2.2. Levels F1, F3 and F5 are called Flight Pair, Flight Pair Group and Complex Group respectively. These levels are generally very similar in structure to levels F0, F2 and F6, and therefore not often studied separately.

The PoS-hierarchy consists of 6 levels in Monet and DeLorean. Here the structure is not as straightforward. In general, level P0 is the actual country where the ticket has been sold. In all cases the level P5 corresponds to PoS-all. The levels in between can be very different for each country. For instance for

Level	KLM Flight hierarchy
F0	Flight
F2	Subline
F4	Complex
F6	Region
F7	KLM total

Table 2.2: KLM flight hierarchy in Monet and DeLorean

France levels P0, P1, P2, P3 and P4 are simply "France" and level P5 is PoS-all. For Honduras however the sequence is totally different, namely Honduras, Panama Region, Central America, Latin America, Americas and PoS-all.

Finally, for each flight the scheduled capacity is available in DeLorean and the actual flown capacity is available in Monet. Seats are only divided into business class seats or business class seats. Therefore if one wants to know, for instance, what the capacity for subclass 'V' seats on flight KL0000 is, the number of economy class seats will be given. This will be the same number for subclass 'N', 'K', or 'X' as these are all subclasses within the economy class cabin. Cabin capacity can be subject to great last minute changes. This is due to the fact that for some airplanes the only difference between business class and economy class seats is the service and not the actual seat. In these aircrafts the two cabins are simply separated by a curtain. This curtain can be moved at the last minute in the occurrence of many (or few) business class passengers. As a result the actual flown number of business class seats can be very different from the scheduled number of business class seats. This phenomenon is mostly seen in European flights.

2.2 Current forecasting methods within KLM

Each of the three revenue groups at KLM has an analyst that provides the short-term revenue forecasts for his group. These forecasts are made at the linegroup level. This flight level is non-existent in Monet and DeLorean, but it is comparable to the level of Region (F6) in Monet and DeLorean. The forecasts are calculated by forecasting two factors: volume and yield. The multiplication of these two gives the revenue forecast.

At KLM a distinction is made between revenue that is earned purely through ticket sales, so-called *Net1* revenue, and revenue that is earned through ticket sales plus fuel surcharges paid by the passengers, *TFSR* revenue. Every passenger of KLM is required to pay a fuel surcharge on top of their ticket price. This fuel surcharge is however not exactly the same for each passenger. A passenger flying from Amsterdam to Paris for instance pays a smaller fuel surcharge than a passenger flying from Amsterdam to Tokyo.

The analysts are only required to forecast Net1 revenue totals. Revenue earned from fuel surcharges is forecasted separately by the controllers at KLM.

2.2.1 Volume

The volume is given in paxkm. This is the total number of kilometers flown by the passengers. If, for instance, 100 passengers are on flight KL0000 and the length of this flight is 1000 km then the number of paxkm $100 * 1000 = 100000$ paxkm. Adding the paxkm of all flights within a certain revenue group leads to the total number of paxkm for that revenue group.

Actually the analysts do not forecast a number for the total paxkm, but they forecast a so-called Load Factor (LF). In every aircraft there is a certain number of seats available for passengers. Every seat travels the same distance as the passengers, whether it is occupied or not. The total number of seats on an airplane multiplied with the distance each seat travels is called seatkm. The Load Factor is now defined as the total number of paxkm divided by the total number of seatkm.

$$LoadFactor = \frac{\#paxkm}{\#seatkm}$$

Note that this is the same as the number of passengers divided by the number of seats at the level of flight, but not necessarily for higher levels. This is due to the fact that at higher levels seats may be from different flights and therefore may have a different distance to travel. In general, at the level of subline this difference is still non-existent, but even at this level differences may occur as a result of multi-leg flights. Table 2.3 illustrates the difference between the two methods for calculating load factors.

Flight (distance in km)	Pax	Seats	LF	Paxkm	Seatkm	LF
London (100)	50	50	100%	5000	5000	100%
Tokyo (5000)	30	50	60%	150000	250000	60%
combined (-)	80	100	80%	155000	255000	61%

Table 2.3: Two possible ways for calculating LF

The reason for forecasting a Load Factor is the fact that capacity is subject to change. Flights may initially be scheduled to fly with a certain type of aircraft, but unexpected circumstances may lead to a change in aircraft or perhaps even a cancelation of the flight. Circumstances might be: unexpected high/low demand, personnel strikes or bad weather. Forecasting the Load Factor ensures that the number of paxkm is adjusted when the number of seatkm changes.

2.2.2 Yield

Yield is defined as the revenue obtained from a single paxkm and is given in euro cents per paxkm. If in the above example the 100 passengers generate a total revenue of € 4000, the yield for this flight is $400000/100000 = 4$ eurocents/paxkm.

As in the case of volume the analysts do not forecast an absolute yield number but a yield factor. In this case this is the factor compared to the yield number of the previous year. The reason for doing this is that a split is made between manageable yield fluctuations an unmanageable fluctuations yield. Unmanageable yield fluctuations are caused by different rates of exchange. If a person buys a ticket for \$100 for flight KL0000 and pays the same amount for the same flight one year later, his yield in euro's can be significantly different merely due to the different rate of exchange. Manageable yield fluctuations can be caused by several factors such as shift in Economy/Business Class traffic, difference in contributing sublines or difference in pricing. All these different influences on the final yield number have led to the concept of yieldmixes.

Yieldmixes Knowing the underlying reasons for a change in yield can be important information for the revenue analysts. If a change is caused by a big shift from business class traffic to economy class traffic or as a result of a sudden devaluation of the dollar, the resulting conclusions can be completely different. In the former case this might mean that perhaps the prices for business class are too high and a price change needs to be made. In the latter case it might mean that the sales focus should be less on countries that use the dollar as currency. In order to study the change in yield a few new definitions have to be introduced.

- **Cabin mix**: That part of the percentile change between current year- and previous year yield, that can be accounted for by the traffic shift between C-class and M-class traffic.
- **Linegroup mix**: That part of the percentile change between current year- and previous year yield, that can be accounted for by the traffic shift between linegroups.
- **Point of Sale mix**: That part of the percentile change between current year- and previous year yield, that can be accounted for by the traffic shift between contributing points of sale.
- **Price mix**: That part of the percentile change between current year- and previous year yield, that is caused by difference in price level.
- **Rate of Exchange mix**: That part of the percentile change between current year- and previous year yield, that is caused by difference in rates of exchange between the euro and currencies of the contributing Points of Sale.
- **Subline mix**: That part of the percentile change between current year- and previous year yield, that is caused by difference in contributing sublines.

Determining the mix factors is probably best explained by a short example. Table 2.4 shows revenue data for flight KL0000.

All tickets for this flight have been sold in the USA. Therefore the earned

Flight	paxkm	py rev($)	py rev(€)	paxkm	cy rev($)	cy rev(€)
M cabin	10000	800	1200	20000	1600	1600
C cabin	5000	800	1200	2500	400	400
Total	15000	1600	2400	22500	2000	2000

Table 2.4: Revenue data flight KL0000

revenue in dollars is given as well as the earned revenue in euro's after conversion. The yield for this year's flight is $200000/22500 = 8.89$ $eurocents/paxkm$. Compared to last year's yield of $240000/15000 = 16$ $eurocents/paxkm$ this is a drop of is a drop of 44.4%. Because all tickets were sold in the USA there has been no shift in Points of Sale and this has been no factor in the change in yield. Also the subline mix and linegroup mix cannot be a factor since a flight is lowest in the hierarchy. This means the change in yield could have been caused by different ticket prices this year compared to last year, a different rate of exchange for dollars to euros or perhaps a shift in business/economy traffic. Determining these mix factors is done stepwise. It is important to note that the increase of $22500/15000 = 1.5$ in paxkm does have an influence on the revenue totals, but this factor does **not** influence the yield change! As the unit of yield is cents per paxkm the total number of paxkm is irrelevant.

Step 1 (RoX mix): The difference between this year's RoX and last year's RoX is $\frac{2000}{2000} - \frac{2400}{1600} = -0.50$ €/\$. As a result this year's revenue is €1000 less than it would have been last year (last year the same \$2000 would have converted to € 3000). Dividing these €1000 by this year's number of paxkm gives the absolute yield change caused by the RoX change. This is $\frac{-100000}{22500} = -4.44$ $eurocents/paxkm$, which implies a change in terms of percentage of $\frac{-4.44}{16} = -27.8\%$.

Step 2 (Price mix): The price mix is determined by calculating the yield changes in local currency (in this example \$) at the lowest detail level. For both the C-cabin and the M-cabin the yield change is 0 (M-cabin: $\frac{80000}{10000} - \frac{160000}{20000} = 0$ and C-cabin:$\frac{80000}{5000} - \frac{40000}{2500} = 0$). This means there have been no changes in the prices and consequently no change in yield is caused by the price mix.

Step 3(Cabin mix): The cabin mix is determined by calculating the yield change that is not caused by the RoX mix and the price mix. Because the absolute yield change $= 8.89 - 16 = -7.11$ $eurocents/paxkm$ the expected revenue change would be $-7.11 * 22500 =$ € -1600. Of this € 1600 we already know that € 1000 was caused by the RoX change. Therefore the remaining € 600 must have been caused by the different ratio between C-traffic and M-traffic, corresponding with an absolute yield change of $\frac{-60000}{22500} = -2.67$ $eurocents/paxkm$. The cabin mix is then $\frac{-2.67}{16} = 16.7\%$.

As the example shows, the idea behind determining the yield mix factors is to adjust all variables (e.g. RoX, ticket price or C/M ratio) to last year's value except one. The difference in expected revenue and actual revenue must then be caused by the variable that has not been adjusted. If we were interested in the yield mix factors at a higher level, for instance revenue group level, and assume sales can be made in all countries of the world the method would be exactly the same. In that case the PoS mix, subline mix and linegroup mix could also be calculated. In fact, other factors such as subclass mix, flight mix or complex mix could also be calculated. It is crucial though to start from the lowest level of detail and work up from there. A more extensive example as well as the algorithm for calculating the different mix factors is given in Appendix A.

2.2.3 General ideas of forecasting methods at KLM

Although the three analysts have their own way of making the monthly forecasts, there are some general ideas and methods that apply to each revenue group.

Forecasting Load Factor especially is more or less done in the same fashion. The main graph for studying LF is the so-called *booking curve* in DeLorean. This curve shows the development of all bookings. This can be done at any level, e.g. flight, subline, linegroup etcetera. For flights that have already occurred in the past the booking curve will start at 340 days before the departure date right up to 1 day before departure. As explained earlier DeLorean contains detailed booking information. For flights that are scheduled to depart at some time in the future the start of the booking curve will again be at 340 days before departure. However, in this case the curve will stop at the last available snapshot date. The simplest way to forecast a load factor is to compare the current booking curve to booking curves of previous years. Figure 2.2 shows an example of such a graph.

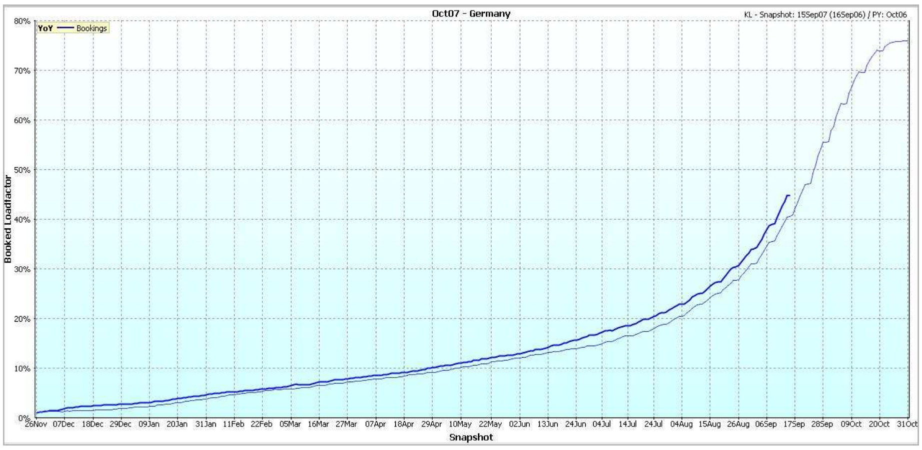


Figure 2.2: Booking curve

The booking curves of years in the past give the analyst a basic idea of how a booking pattern might develop. Combining these booking curves with his knowledge of inventory steering mechanisms, possible trends observed in previous months or perhaps plain intuition lead the analyst to make a forecast of where the booking curve will finish.

Forecasting yield is found by the analysts to be more difficult. Also the methods for forecasting yield differ considerably between the three analysts. However also in this case there are some ideas they have in common. First of all the historical yield development is studied in Monet. An example of a yield graph is given in figure 2.3.

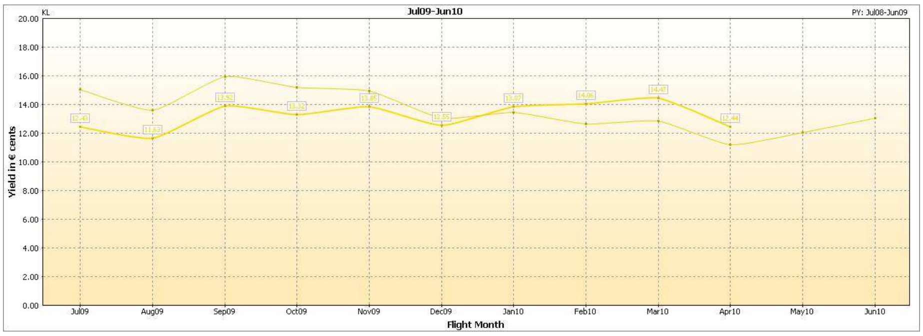


Figure 2.3: Yield graph

As in the case of the booking curve the goal is to forecast how the graph will continue. Unfortunately there are a lot more factors that can influence the yield number. Examples of these factors have already been given in the section on yield mix factors. Forecasting each mix factor separately and then combining all these factors to find the total yield change would be a possible method for forecasting yield. In practice however this is found to be too time consuming and more importantly too difficult. Therefore generally the analysts only study the factors that based on their experience influence yield the most. For instance during the financial crisis it was observed that the number of business class passengers declined rapidly. Consequently it was observed that the cabin mix had a major negative effect on yield indices. In order to forecast the separate mix factors historical revenue data from Monet is combined with booking data from DeLorean. Current bookings for the upcoming months are compared to months from the past. The impact a certain factor had on yield in the past is then used to predict the influence that factor will have on future yield indices. Figure 2.4 shows how this could work for the cabin mix.

In January we see that there has been a negative change in the C/M cabin ratio compared to previous year's January. The share of "cheap" economy class bookings became larger and as a result the share of "expensive" business class bookings became smaller. This change will cause the combined yield number to decrease. Monet tells us that the yield index for January was 0,95. The shift in C/M-traffic was less negative in February. Consequently, in February the yield index was observed to be 0,97. Since the bookings for April show that the change in C/M-traffic will be even less negative (perhaps a positive change will be observed) than February, we might predict that the yield index will be somewhere in the region of 1,00. Because there is no revenue information available for March, it is not possible to directly use this month for predicting yield

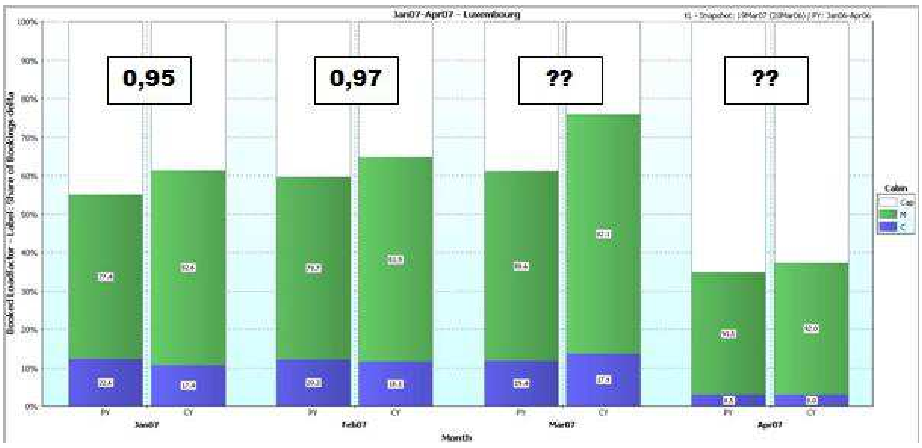


Figure 2.4: Cabin share of bookings

indices. The current booking information for March can be used to examine the changing cabin ratio's.

2.2.4 Controlling

Although it is impossible to forecast with complete accuracy it is vital that the forecasts are fairly reliable. If the difference between actual outcomes and forecasts are frequently very large then decisions can never be made on the basis of these forecasts. Therefore accuracy targets are set for the analysts. Once all facts of a certain month are available controllers compare these outcomes to the forecasts made by the analysts for this month. Before the comparisons can be done however a capacity correction and a RoX correction are made for the forecasts. These corrections are necessary, because these two variables may have changed considerably since the moment the forecasts were made.

Capacity correction As explained in section 2.2.1 capacity is subject to change. These changes can directly influence the analyst's forecasts. Because a load factor percentage is given to forecast volume, a decrease/increase in capacity will result in a decrease/increase in revenue. This could mean an analyst might forecast the exact load factor percentage and the exact yield number but still be totally inaccurate for the total revenue forecast. It is clear that this is unfair. The revenue total calculated at the moment of forecast is therefore corrected using the actual capacity.

$$Rev_{fc} = (LF_{fc} * Cap_{act}) * Yield_{fc}$$

RoX correction Another complicating factor in judging revenue forecasts is the fact that rates of exchange fluctuate. The analysts forecast a manageable yield factor. The controllers then add a RoX factor to this number in order to calculate a revenue forecast.

$$Rev_{fc} = Volume_{fc} * (1 + Manageable\ yield_{fc} + RoX_{fc}) * Yield_{fc}^{py}$$

A simple example will clarify this function. Suppose an analyst forecasts a volume of 10000 paxkm and that the manageable yield factors will have add up to -10%. If last year's yield total was 10 *eurocents/paxkm* and the controllers calculate a RoX factor of -5% then the forecasted revenue total will be:

$$Rev_{fc} = 10000 * (1 - 0.1 - 0.05) * 10 = 85000\ eurocents$$

If the RoX factor used at the moment of forecasting is very different from the actual RoX factor this again directly influences the forecasted revenue totals. Therefore the controllers correct the forecasted revenue totals by using the actual RoX factors instead of the forecasted RoX factors.

One final correction is made in the occurrence of extraordinary circumstances. Sometimes extraordinary events, which could have impossibly been foreseen by the analysts, affect the eventual revenue totals. Specific examples of these type of events are the Mexican Flu, personnel strikes and most recently the eruption of the volcano Eyjafjallajkull in Iceland. In these special cases the analysts are allowed to estimate how much effect the event had on the eventual outcome. The controller then corrects the forecasts accordingly to score the revenue forecasts.

Targets After the corrections have been made the forecasts can be compared to the actual outcomes. Table 2.5 shows the error margins allowed for the analysts. The absolute percent error (APE) is calculated and every time this

Months ahead	APE bound
1 month	1.5%
2 months	4%
3 months	6%

Table 2.5: Allowed margins of error

error is smaller than the allowed margin the analysts score a point. Besides the revenue group level scores are also kept at the linegroup level. These scores can give good insights into the quality of the forecasts.

Chapter 3

Research

Three different types of forecasting methods have been studied before it was decided to create a new method. The three types were

1. Bayesian Belief Nets (BBN's)
2. Time Series
3. Logistic Regression

Each method will be shortly explained before discussing how it could be used for our project.

3.1 Bayesian Belief Nets

3.1.1 Introduction

Bayesian Belief Nets (BBN's) are directed acyclic graphs (DAG's). The nodes of the graph represent univariate random variables, which can be discrete or continuous, and the arcs represent direct influences[2].

One of the basic rules of probability theory is that the probability of event A given event B is

$$P(A|B) = \frac{P(A,B)}{P(B)},$$

where P(A,B) is the probability of event A and event B occurring. Rearranging this formula leads to the so-called *product rule*

$$P(A,B) = P(B) \cdot P(A|B).$$

For three variables this can be written as

$$P(A,B,C) = P(A|B,C) \cdot P(B|C) = P(A|B,C) \cdot P(B|C) \cdot P(C) \ ,$$

which can be extended to the *joint probability function* for the n-dimensional case

$$P(A_1, A_2, \ldots, A_n) = P(A_n) \prod_{i=1}^{n-1} P(A_i | A_{i+1}, \ldots, A_n).$$

Bayesian Belief Nets make it possible to simplify the above formula by making explicit the dependencies between variables. This simplification is done through *conditional independency*.
Two variables are said to be *independent* if

$$P(A|B) = P(A),$$

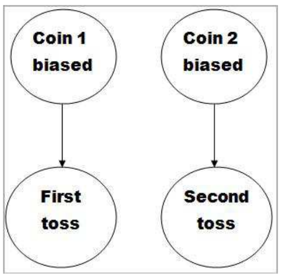
from which follows

$$P(A, B) = P(A) \cdot P(B).$$

A simple example of this idea is two persons tossing a different coin. If event A represents the outcome of the first person's toss and event B the outcome of the second person's toss, then it is clear that events A and B are independent. This also means that if we know the outcome of event B this will not affect our belief about the outcome of event A.
The previous example changes however when we let both persons throw the same coin. Suppose also that there is a possibility of the coin being biased towards Heads, although this is not certain. In this case events A and B are no longer independent. If, for instance, the outcome of event B is Heads, then this will increase our belief in event A being Heads as well. In this case events A and B are dependent on a separate variable C, namely "the coin is biased towards Heads". This variable can take on the values True or False. Events A and B are dependent, but once the outcome of C is known, they become independent. In other words, if we know the value of C the outcome of B will not affect our belief about A. Events A and B are said to be *conditionally independent given* C. In formula this can be written as

$$P(A|B, C) = P(A|C).$$

As mentioned earlier BBN's are directed acyclic graphs. The nodes represent variables and arcs are added between nodes if there exists a direct influence between them. The variables can be discrete or continuous. A BBN for the coin tossing example is shown in figure 3.1.



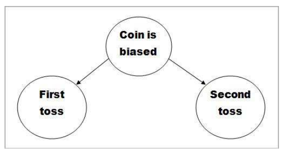


Figure 3.1: Left: two different coins, right: one coin

The left figure illustrates the case where two coins are tossed. Events A and B are independent and therefore no arcs are added. In the second case there still is no direct influence between events A and B, but now an extra node is added for the event "coin is biased towards Heads". Because the outcome of this new event directly influences the outcome of events A and B, arcs are added from event C to events A and B. The main reason for using BBN's is that they make it possible to model and reason about uncertainty. If the coin is biased towards Heads this does not mean that the outcome of events A and B will be Heads. It only means that the probability of Heads will increase.

Coin	Biased	Unbiased
Heads	0.8	0.5
Tails	0.2	0.5

Table 3.1: Probability table

The probabilities given in table 3.1 are so-called *conditional probabilities*. If it is known whether the coin is biased or not the corresponding probabilities for events A and B are known. If this is not known, but the probabilities of the coin being biased or not are known, it is possible to calculate the *unconditional probabilities* of events A and B.
Suppose empirical evidence has shown the probability of a coin being biased towards Heads is 0.1. The unconditional probability of a coin toss being Heads then becomes:

$$\begin{aligned} P(H) &= P(H|biased) + P(H|not\ biased) \\ &= P(H\ and\ biased) \cdot P(biased) + P(H\ and\ biased) \cdot P(not\ biased) \\ &= 0.8 \cdot 0.1 + 0.5 \cdot 0.9 = 0.53. \end{aligned}$$

The unconditional probability distribution is also called the *marginal distribution*.
Conversely, if we observe that the first coin toss is Heads, the probability that the coin is biased towards Heads also increases. This can be seen using *Bayes' theorem*:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \ ,$$

which can easily be derived from the fact that

$$P(A|B) \cdot P(B) = P(A, B) = P(B|A) \cdot P(A).$$

Using Bayes' theorem we find for our example that

$$\begin{aligned} P(biased|H) &= \frac{P(H|biased) \cdot P(biased)}{P(H)} \\ &= \frac{0.8 \cdot 0.1}{0.53} \approx 0.15. \end{aligned}$$

Finally, even though the two coin tosses seem to be independent, the result of the first coin toss influences the probabilities of the outcome of the second coin toss. For instance, if the first coin toss is observed to be Heads the probability of the coin being biased has already been shown to change. This change directly influences the probability of the second coin toss being Heads to:

$$\begin{aligned} P(H) &= P(H|biased) + P(H|not\ biased) \\ &= P(H\ and\ biased) \cdot P(biased) + P(H\ and\ biased) \cdot P(not\ biased) \\ &= 0.8 \cdot 0.15 + 0.5 \cdot 0.85 \approx 0.55. \end{aligned}$$

It is important to realize that, if given that the coin is biased, the outcome of event A no longer influences event B.

All direct predecessors of a node i in the graph are called *parents* of that node and form the set *Pa(i)*. Nodes without parents are called *root nodes*. The directed graph of a BBN that each variable is conditionally independent

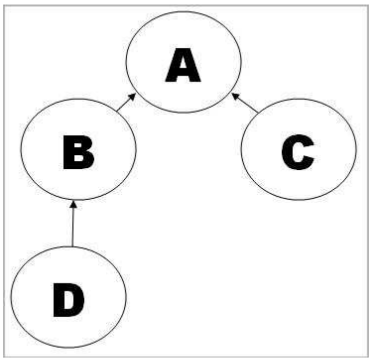


Figure 3.2: Nodes **D** and **C** are *root nodes*, nodes **B** and **C** are *parents* of node **A**

of all predecessors in the ordering given its direct predecessors (parents)[2]. As a result the joint probability function can be simplified accordingly. The joint distribution for the n-dimensional case now becomes.

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i|Pa(A_i))$$

Suppose we have a network of four nodes *A, B, C* and D then without knowing anything about the dependencies between the variables the joint probability function is $P(A, B, C, D) = P(A|B, C, D) \cdot P(B|C, D) \cdot P(C|D) \cdot P(D)$. However if the dependencies are explicitly modeled as the BBN in figure 3.2 the joint probability distribution becomes $P(A, B, C, D) = P(A|B, C) \cdot P(B|D) \cdot P(C) \cdot P(D)$

3.1.2 Forecasting using a BBN

BBN's are generally used for two reasons *diagnosis* and *forecasting*. Diagnosis is is sometimes called *bottom-up reasoning*. An event is observed and the effect this observation has on its predecessors is studied. As a result the main reason(s) for this observed event might be "diagnosed". In the same manner forecasting can be done. An event is observed and the effect this has on its descendants is studied.

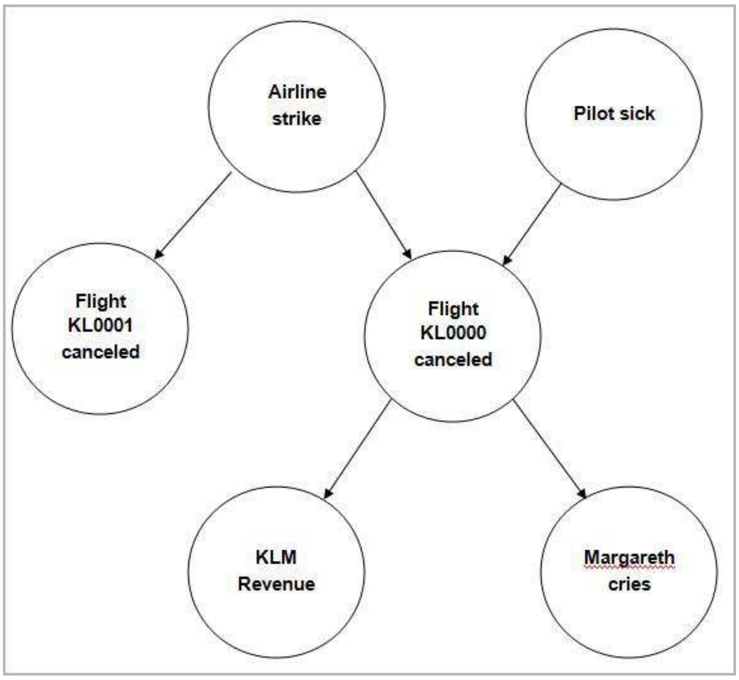


Figure 3.3: Diagnosis and forecasting using a BBN

Figure 3.3 is an example of the manner in which BBN's can be used for diagnosis and forecasting. The BBN contains 5 boolean random variables, meaning the outcome is either True or False. The node called KLM revenue is a continuous random variable. Suppose we observe that flight KL0000 has been canceled. Given the BBN there can be two reasons for this cancelation, either there was an airline strike or the pilot is unable to fly due to sickness. The observed cancelation increases the probability of an airline strike as well as the probability that the pilot is sick. The probability that the pilot is sick might be significantly larger than the probability of an airline strike. However, if we observe that flight KL0001 has been canceled the probability of an airline strike will increase even more and perhaps become larger than the probability the pilot being sick. In this case we say that the cancelation of flight KL0000 has been 'explained away'. This explaining away technique can be used to diagnose the cause of certain observed events.

The same observation of a canceled flight can however also be used for forecasting. It will for instance influence the probability that Margareth will cry (perhaps because her father was supposed to come back home on that flight).

It will also influence the revenue totals of KLM. The cancelation will mean that for this flight will no revenue will be earned. The expected value of the KLM revenue will therefore be lower given the event of the cancelation.

For all examples discussed in this chapter so far the dependencies are known and the corresponding structure of the DAG is given. The main difficulty in forecasting using BBN's is that these dependencies are not known. Therefore the structure of the DAG is also unknown. Anca Hanea of the Technical University of Delft has suggested a method for determining this structure.

Hanea method The main idea behind the Hanea method is to build a joint probability distribution for ordinal data using the joint normal copula[2].

Definition 3.1.1. *The* copula *of two continuous random variables X and Y is the joint distribution of $F_X(X)$ and $F_Y(Y)$, where F_X, F_Y are the cumulative distribution functions of X, Y respectively. The copula of (X, Y) is a distribution on $[0,1]^2 = I^2$ with uniform marginal distributions.*

This definition uses the fact that the random variable X'defined as $X' = F_X(X)$ has a uniform distribution. The joint normal copula is then constructed from the bivariate normal distribution. This idea can be extended to the general multivariate case.

All the underlying mathematical ideas of the Hanea method are well described in her PhD thesis "Algorithms for non-parametric Bayesian Belief Nets" and will therefore not be discussed in this thesis. Very simply put the Hanea method for determining the structure of a BBN is as follows:

1. Represent the element that needs to be forecasted and variables that may influence this element as unconnected nodes of a graph

2. While the determinant of the rank correlation matrix of a BBN using the normal copula is not within a certain confidence band, add arcs to the graph

This "simple" definition of the Hanea method already illustrates that some mathematical computations will have to be made. Calculating the rank correlation matrix and its determinant as well as determining the confidence band may require mathematical knowledge that is not known to the forecast analyst. Even if the analyst is familiar with the necessary calculations, the number of variables may be so large that this becomes a very time consuming ordeal. For this reason a software package called *Uninet* has been developed at the Technical University of Delft. The main focus of Uninet is dependence modeling for high dimensional distributions. It can handle discrete distributions, continuous distributions and distributions imported from a sample file. This last option has been used to experiment with forecasting revenue and will be explained a little bit further.

In order to examine data in Uninet this data has to be imported using a *.sae* file. This is a comma separated file containing the names of the variables on the first row and the multidimensional samples on the following rows. Based on the historical data Uninet then can calculate three types of correlation matrices and their determinants. These three are:

- DER = determinant of empirical rank correlation matrix
- DNR = determinant of the rank correlation matrix obtained by transforming the univariate distributions to standard normals, and then transforming the product moment correlations to rank correlations using the Pearson's transformation. This matrix is also called the normal rank correlation matrix.
- DBBN = the determinant of the rank correlation matrix of a BBN using the normal copula.

The DNR is in fact the same is the DBBN if the constructed BBN is the saturated graph containing all variables. The goal is to keep the BBN as sparse as possible, because this makes it easy to interpret the influences that are most important. This is why the variables are added to the graph first without connecting them by any arcs. Arcs are then added one by one until the DNR is within the $\alpha\%$ confidence band of the DBBN. The value of α can be chosen to be 90, 95 or 99. Crucial in this process is to determine how arcs are added to the graph. This can be done by simply choosing arcs that are thought to be of important influence. A more scientific approach is to connect variables that correspond to the highest correlation coefficients in the normal rank correlation matrix.

Once the structure of the BBN has been fixed the forecasts can be made in the manner explained earlier. Input variables are observed and entered in the graph. As a result the marginal distributions of all other variables, including the variable to be forecasted, are updated. Once all observations have been entered the forecasted value is the expected value of the marginal distribution.

Forecasting revenue at KLM An important challenge when forecasting revenue using a BBN is to determine a structure and candidate variables for the graph. Revenue needs to be forecasted, which would imply creating a BBN where revenue is the variable to be forecasted. However revenue at the level of revenue group is the aggregation of revenue at lower levels. Therefore it would also be possible to create BBN's for each lower level and finally aggregate these low level forecasts to obtain the high level forecasts. Also it is possible to split revenue into yield and volume and create two separate BBN's for these two variables. Obviously, as the level of forecast is lowered the number of necessary BBN's increases.

Possible input variables can consist of variables that are observed and variables that need to be forecasted. For instance, a possible input variable could be to use the share of bookings percentages for cabins or subclasses. These percentages can be taken at the moment of a certain fixed snapshot, e.g. 10 days before the next month, or the actual percentages can be used. When forecasting one month ahead the snapshot percentages will be available as forecasting is generally done at the end of the month. The observed percentages will then be entered in the corresponding node of the graph, which will alter the distribution of the forecast variable. When actual share of bookings percentages are chosen as an input variable these percentages will have to be forecasted themselves.

Many different input variables for revenue forecasting can be chosen. All types of booking information can be used, such as the aforementioned share of bookings percentages as well as differences in year over year booking patterns. Besides booking information historical revenue information can be used for input variables. For instance, the difference between the actual totals of a certain month and the average totals the last five years of that month can be used as an input variable. This type of variable could be used as an indication of "good" or "bad" times. Again, many different input variables can be chosen. Examining the correlation coefficients will then reveal which factors are important and which are not.

3.1.3 Advantages and disadvantages

There are some great advantages to using a BBN for forecasting purposes. The most important advantage being the fact that interactions and dependencies between variables can be made clearly visible through the graph structure of the BBN. Not only are the interactions visible in the graph, it is also possible to experiment a little with the variables. The use of Uninet makes it possible to choose different values for a certain variable and see what the effect is on the other variables. Also the use of Uninet makes it possible to work with large numbers of variables and dependencies. Uninet was created to assist in the examination of a large project called the *CATS* project. For this project a graph consisting of 1359(!) nodes was created. This means that many potential input variables can be tested by the analyst.

Unfortunately there are some major obstacles that need to be taken before this method can be implemented at KLM. The first obstacle is the availability of the software package Uninet. This package was made available for free for this project. However it is very probable that this is not available for free for KLM. A commercial company like KLM would most likely have to pay to use this package. Even if this obstacle could be taken, a very serious problem is the compatibility of Uninet with the KLM decision support tools. At this point it is only possible to import *.sae* files into Uninet. This means that data from DeLorean or Monet has to be exported to an Excel file, which in turn has to be converted to a *.sae* file before it can be imported into Uninet. This is possible, of course, but not very practical for the analyst to work with. Exporting data from Uninet into Monet or DeLorean is impossible at this moment. Finally, even without these practical problems the method of forecasting using a BBN has some drawbacks. An important drawback is the necessity to update the graph. Dependencies may vary in time. As a result factors that were seen to be of great influence one year might not be so important the next year. This means the structure of the graph and its corresponding dependencies should be updated at some point. This updating would require repeating the whole process of building the BBN, which is very time consuming and also requires the understanding of the analyst.

3.2 Time Series

3.2.1 Introduction

A time series is a sequence of *data points*, measured typically at successive times spaced at uniform intervals [3]. The main difference between time series analysis and other common data analysis problems is that there is a natural ordering of the data. If there are T observations, they may be denoted by y_t, with $t = 1, \ldots, T$. Forecasting a future observation $\hat{y}_{T+1|T}$ using time series is essentially an extrapolation technique.

Several methods of time series forecasting are available. An important family of mathematical models has been introduced by Box and Jenkins. These models are often called *autoregressive moving average* models (ARMA). The basic notion of these models is stationarity. Very roughly speaking this means that a time series has no timelike trend. If a time series is not stationary, transformations such as differencing can be applied in order to derive a new time series that is. In these cases we speak of *autoregressive integrated moving average* models (ARIMA).

The dependence of the current value y_t on previous values is called an autoregressive (AR) process. The relation of the current value y_t to previous forecasting errors is called an moving average (MA) process. An ARMA(p,q) process combines autoregressive and moving average components. The general notation for an ARMA(p,q) model is:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_q \epsilon_{t-q} + \ldots + \theta_1 \epsilon_{t-1} + \epsilon_t$$

, where ϕ_i are parameters of the model and ϵ_i are *white noise* error terms. The constant c is often omitted for simplicity. An ARMA(p,0) is sometimes called an AR(p); and ARMA(0,q) is sometimes called MA(q). The parameters p and q are the number of autoregressive parameters and moving average parameters respectively. For non-stationary time series ARIMA(p,d,q) models can be used. The parameter d implies that the d-th order difference is taken from the original time series to remove the trend effect.

Before the actual forecasting can be done a suitable model has to be selected. This is generally done in three steps.

1. Analyze the *autocorrelation function* (ACF) and *partial correlation function* (PACF)
2. Select candidate models
3. Determine best model using *Akaike information criterion* (AIC) or *Bayes information criterion* (BIC)

Instead of analyzing the ACF and PACF to select candidate models it is possible to just consider a lot of different ARIMA models. Plotting the ACF and PACF can however give an indication of the data containing a trend or seasonal pattern. As a result the total number of models to be examined can be reduced. The AIC and BIC are often chosen as a measure for the best model, because a compromise must be found between the *goodness of fit* and the number of parameters of the model. Simply taking the MAPE scores (or perhaps the Mean Squared Errors) of all models would result in a very complicated model and possibly *overfitting*. This is due to the fact that the MAPE or MSE can always be made smaller by simply adding an AR or MA term. The AIC and BIC are defined as follows:

$$AIC = -2 \log L + 2m$$

$$BIC = -2 \log L + m \log n$$

, where L is the likelihood of the data given a certain model, n is the number of observations and $m = p + q$ is the number of parameters of the model. As the model gets bigger and more complicated it will fit better. The first term $-2 \log L$ will get smaller, but the second will get bigger, because m gets bigger. The model that has the minimal value of the AIC (or BIC) should be chosen as the best compromise.

3.2.2 Revenue Forecasting using Time Series

A study on revenue forecasting using time series models has already been done in the past at KLM. This study[6] was done by Roger Hendriksz and focused on *smoothing techniques*, such as *exponential smoothing* and the *Holt-Winters method*. Just as the ARMA and ARIMA models these techniques can reveal more clearly underlying effects such as trends or seasonality. In fact, ARIMA(0,1,1) and ARIMA(0,2,2) are the same as the exponential and double exponential smoothing techniques.

The smoothing techniques were applied to the total KLM revenue totals. The same can be done for other ARMA and ARIMA models. The method for selecting the best model would be the same as suggested in the previous section. Statistical software packages such as SPSS and R can be used to make the necessary calculations.

Not only can the high level totals be forecasted using these models, but also lower level forecasts can be made as well as yield and/or volume forecasts. Combining low level forecasts by aggregation or multiplication can then lead to the desired high level revenue forecasts.

3.2.3 Advantages and disadvantages

There are two main reasons why ARMA and ARIMA models have not been chosen as the eventual forecast method. Firstly, time series forecasting only uses historical data as input for the forecast. KLM Decision Support Tools such as DeLorean however already provide useful information about the future. For example, if bookings for a future month show a great increase of business class tickets compared to previous months, this is an indication that revenue will be higher. It makes sense to select a forecast method that incorporates this up to date information.

The second reason for not choosing time series models is practicality. The main goal of the project is to find an accurate forecast method. Of course, if this can be achieved without using booking information, this will be done. However, the forecast method will also have to be used by the analysts. Because of this reason the goal was to implement the model in Monet instead of learning the analysts how to work with R or SPSS. This implementation is not so straightforward. Simple moving average models can be programmed in the C++ code of Monet. More complicated models are not that easy. During this project a link has been created between Monet and R. This link makes it possible to open R from Monet after which a script can be run automatically.

3.3 Regression analysis

3.3.1 Introduction

Regression analysis is a statistical tool for the investigation of relationships between variables. Variables are divided into inputs, often called *independent variables* or *predictors*, and outputs, called *dependent variables* or *responses*. In general, a regression model is of the form,

$$Y \approx f(X, \beta)$$

, where Y is a vector of dependent variables, X is a matrix of independent variables and β is a vector of parameters. More formally this is sometimes written as $E(Y|X) = f(X, \beta)$. The function f needs to be specified in order to be able to perform the regression analysis.

An important family of regression models are the so-called *linear regression models*. In linear regression, the model specification is that the dependent variable is a linear combinations of the parameters. It is important to note that the dependent variable does not need to be a linear combination of the independent variables. The general form of a linear regression model is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{ip} x_p,\ i = 1, \ldots, n$$

, which can also be written in the vector notation

$$Y = X^T \beta$$

In many models the first column of X is set to 1 and the corresponding element of β, usually denoted as β_0, is then called the *intercept*.

A popular method for estimating the unknown parameter vector $\beta^T = (\beta_0, \beta_1, \ldots, \beta p)$ is *least squares*. This method selects the values of β that minimize the residual sum of squares,

$$
\begin{aligned}
RSS(\beta) &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \\
&= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2
\end{aligned}
$$

Differentiating this equation with respect to β and setting the derivative equal to 0 leads to the solution

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

It can be shown that this $\hat{\beta}$ is an unbiased estimator and that the variance is proportional to the number of inputs. In other words, least squares models using many predictors will show large variance and as a result might have low prediction accuracy. This prediction accuracy can sometimes be improved by using only a subset of the input variables and eliminating the rest from the model. The resulting estimate of β will perhaps show some bias, but because the variance has decreased an improvement in prediction accuracy may be witnessed. Examples of these subset selection methods are *Best-Subset Selection*, *Forward Stepwise Selection* and *Backward Stepwise Selection*.

Instead of totally discarding certain variables from the model it is also possible to use *shrinkage methods*. These methods impose penalties on the size of the coefficients. Examples of these methods are *Ridge Regression* and *The Lasso*[7].

All models discussed in this section are linear regression models. Many other relationships may exist between the dependent variable and its predictors. However often transformations such as log, square-root and square make it possible to study the data using linear models. For instance the non-linear regression model $Y = ae^{X^T\beta}$ can be transformed to the linear model $\log Y = \log a + X^T\beta$.

3.3.2 Revenue Forecasting using Regression analysis

The most complex part of forecasting using regression analysis is the selection of the best model and the calculation of its parameters. Once a model has been determined new observed input variables can be entered into the model and the forecast value is then outcome of the model. Obviously this requires all input variables to be available or to be estimated by the analyst.

Before the best regression model and its parameters can be determined a decision has to be made concerning the level of the forecast model. Forecasts can be made at the level of revenue group or at a lower level after which aggregation leads to a higher level forecast. Also it is possible to use different models for yield and volume. Multiplication of the outcomes of these separate models gives the desired revenue forecast.

For regression models the same types of input variables can be used as with the BBN models. Historical data from Monet and booking information from DeLorean should probably be combined for an optimal result. This was confirmed by some preliminary tests done in the software package R. Low level yield totals were used as input variable in a simple linear regression model for a higher level yield total. The idea behind this test was that, in contrast to revenue or paxkm totals, yield totals can not be added up to obtain higher level yield totals. Using the low level totals as input variables in a regression model could make it possible to determine the higher level totals. The problem with this model was that when passenger traffic is distributed differently than "normal" months the model shows poor forecast results. For instance, suppose passenger traffic to Spain is usually distributed equally over high-yielding Barcelona and low-yielding Madrid. If for some reason in a certain month 80% over the passengers traveling to Spain go to Barcelona and the remaining 20% goes to Madrid, the yield totals for Spain will be higher than normal. This is not recognized by this model, but could have been recognized if the bookings for that month had been taken into account.

3.3.3 Advantages and disadvantages

Regression methods for forecasting were not chosen as the main focus of this project for several reasons. One important reason was practicality. Because the model might need to be updated at some point, either the model has to be created automatically from Monet or the analysts will need some knowledge of regression analysis. Also knowledge of a software package like SPSS or R will be required. It is not certain that the analysts will possess this knowledge. As mentioned in section 3.1.3 a link has been created between R and Monet, making it possible to run a script in R from Monet. This could make it possible to use a regression model, but more work will have to be done.

Again, even besides the practical problems there are other objections to regression models. The amount of available data is the main problem. The decision support tools Monet and DeLorean contain very detailed information. However, this detailed information only goes back to January 2005. As this project started in September 2009, only 57 months of historical data were available. If factors such as seasonality effects are also taken into account in the model this total becomes even less, because different models might be needed for different seasons. This number of observations is not very much for determining a trustworthy model.

Chapter 4

The model

The crucial question in this thesis is:

At what level should forecasts be made to obtain the most accurate results at the highest level?

As KLM is mostly interested in the revenue forecasts at the level of the revenue groups it is essential that these forecasts are as accurate as possible. This does not mean however that the forecasting methods need to be applied to the totals at this level. It is also possible to use a *Bottom-up* approach where forecasts are made at a lower level and then aggregated to determine the forecast at the higher levels. The great advantage of this bottom-up approach is that the different dynamics at the lower levels can be taken into account. For instance, the seasons of Cape Town and Tokyo are complete opposites. Both destinations are within the RV revenue group though. It makes sense to try to make an accurate forecast at the level of subline in order to take this difference in seasonality into account. Another advantage of bottom-up forecasting is that low levels errors may cancel each other out at a higher level.

Unfortunately, bottom-up forecasting often has very poor accuracy at higher level forecast levels. This may be a result of forecast error at intermediate (middle) levels accumulating as data moves up to higher levels[4]. The alternative is to take a *Top-down* approach. Forecasts are made at the highest level. The high level totals can then be distributed over the lower levels to obtain the lower level forecasts. In this case low level forecasts tend to be very inaccurate though.

All methods described in this chapter were tested at several different levels. Always the results at the highest level determined the quality of the method.

The main idea for determining an accurate forecast model was based on three steps. First try to find a method to forecast the number of paxkm (volume) accurately. Then do the same for yield. And finally combine these two methods and hope that this combination achieves good results for the total revenue. It should be noted that since the revenue analysts are required to forecast Net1 totals, this is always the type of revenue that is forecasted in this thesis.

The analysis of the results has all been done in the Decision Support tool Monet. The main advantage of working in Monet is that the large amounts of data are easily accessible and do not need to be exported to other programs as Excel or R. Also making comparisons between actually realized totals and forecasted totals is very simple in the graphical environment of Monet. As forecasts are generally made at the end of a month all revenue and passenger data of the previous month is presumed to be available. The last month for which all actual data is available is called the *last data month*.

All forecast methods were tested for the months January 2008 to November 2009. These months were chosen partly for practical reasons, but also because this period contains "normal" and "abnormal" months. The biggest part of 2008 was a relatively steady period where no unusual circumstances took place. This changed at the end of 2008 when the airline industry was hit hard by the global financial crisis. This is best visible for the months December 2008 to March 2009 when revenue totals plummeted. From April 2009 onwards a gradual recovery was visible, at least in the sense that the market stabilized again. One exception being May 2009 when especially RU and RW were hit hard by the Mexican Flu. More months were available in Monet and DeLorean for testing. This was not done, because many different methods had to be tested and these tests are very time consuming.

For each month the absolute percent error (APE) of the forecast is determined. The forecast method achieving the lowest Mean Absolute Percent Error (MAPE) is considered to be the best method when forecast paxkm or yield. When forecasting the total revenue, the number of times a model is within the margins allowed to the analysts is decisive. The eventual suggested model will finally be compared to the results of the analysts using a *validation set*. This validation set consists of the entire year 2007 plus the months December 2009 to March 2010. Again these months were chosen, because they form a good mix of "normal" months with "abnormal" months.

In contrast to the scores of the analysts no corrections are made in the case of extraordinary events. As a result sometimes MAPE scores may be very negatively influenced by a single month. When this occurs this will not always be mentioned. Only when a method is considered to be a suitable forecast method will this be taken into account.

Finally, a small remark concerning examples in this thesis. All examples have been chosen to best clarify a certain method or problem. Totals used in these examples are totally fictitious. They can therefore seem very unrealistic.

Notation Before the actual models are discussed it is necessary to introduce some notation. Superscripts are used to indicate if for some variable this year's total is meant or last year's. In the first case the superscript cy is used. For the latter case the superscript py is used. When more information is needed to distinguish variables a subscript is used. Examples of subscripts that are frequently used in this thesis are € or lc when discussing revenue. The subscript in these cases clarifies if a given total is given in euro's or perhaps in a passenger's

local currency.
Throughout this thesis a distinction is also made between *pkm* and *rpkm*. Paxkm totals derived from the booking information are indicated with pkm. These totals can be retrieved in DeLorean. Actual flown paxkm totals are indicated with rpkm. These totals are the revenue paxkm totals which can be found in Monet. The same distinction is made when discussing load factors. Here ALF represents actual load factors and BLF booked load factors.

4.1 Volume

Forecasting volume at the highest level eliminates the possibility of taking into account the processes going on at the lower levels. A very important process being inventory steering. Inventory steering is done at the flight level and will normally be done in the same manner year over year. If however a flight is filling up a lot slower (or quicker) than last year the inventory analysts will accept more (or fewer) new passengers compared to last year. A bottom-up approach would possibly make it possible to make adjustments for these cases. Three different methods were examined to forecast volume:

1. Paxkm without capacity
2. Paxkm with capacity
3. Load Factor

For all these methods an important decision that had to be made was whether to work with *absolute* differences or *relative* differences. Working with absolute differences induces the risk of ending up with negative totals. As we are forecasting paxkm (or yield) this should not be possible. Relative differences on the other hand run the risk of dividing by zero. This is something that will generally not occur that often in the case of forecasting paxkm. In the case of forecasting yield however this problem arises much more frequently.

4.1.1 Paxkm without capacity

The main assumption behind this approach is that capacity does not influence the number of sold tickets. Capacity is assumed to be determined by the volume. If demand is great and there are a lot of bookings then capacity will be increased to accommodate this demand. Conversely, capacity is decreased when there are few bookings. In reality this is generally not the case, because capacity is scheduled in advance. The reason for excluding the influence of capacity is to keep the model as simple as possible. Forecasting volume can be done by solely studying paxkm curves. Several ideas have been tested using only paxkm curves. The two most important ideas were:

1. The absolute/relative difference in paxkm between current year and previous year at the moment of departure is the same as at the moment of snapshot

2. The absolute/relative difference in paxkm between current year and previous year at the moment of departure is the same as at the moment of snapshot multiplied with an adjustment factor based on the last data month

In formula this can be written as

1. (a)
$$\overline{rpkm} = pkm_{ss}^{cy} + (rpkm^{py} - pkm_{ss}^{py})$$
or

   (b)
$$\overline{rpkm} = pkm_{ss}^{cy} * \frac{rpkm^{py}}{pkm_{ss}^{py}}$$

2. (a)
$$\overline{rpkm} = pkm_{ss}^{cy} + (rpkm^{py} - pkm_{ss}^{py}) * factor$$
or

   (b)
$$\overline{rpkm} = pkm_{ss}^{cy} * \frac{rpkm^{py}}{pkm_{ss}^{py}} * factor$$

The factor is the difference between the forecasted total using formula 1a or 1b and the actual total of the last data month, $factor = rpkm_{ldm} - \overline{rpkm_{ldm}}$ or $factor = rpkm_{ldm}/\overline{rpkm_{ldm}}$. For example, in February we are forecasting for the month of March. The last month for which all data is available is January. Using formula 1b a forecasted paxkm total of 200 km is calculated. The correction factor is then found by simulating that we are forecasting the paxkm total of January. Say we find that our forecast would be 120 km. Looking back we now read in the data that the actual paxkm is 100 km. Then the correction factor is $\frac{120}{100} = 1.2$. In other words we overestimated with a factor 1.2. The forecasted paxkm total for March becomes $200 * 1.2 = 240$ km.

The results of these four different methods are given in table 4.1

Method	RU	RV	RW
1a	2,23	4,55	2,75
1b	2,86	4,79	3,31
2a	2,11	4,68	2,55
2b	2,77	4,83	3,36

Table 4.1: MAPE scores for paxkm without capacity

Figure 4.1: Correction factor

4.1.2 Paxkm with capacity

In actuality capacity does have an influence on ticket sales. If capacity for a destination is significantly larger this year than for the same destination last year it makes sense to assume that more bookings will be accepted. As a result the number of paxkm that is added from the moment of snapshot to the moment of departure will be greater. Figure 4.2 shows an example of this idea. On the $22^{\text{nd}}$of April the two paxkm totals were roughly equal. In the end this year's total finished higher than last year's.

Figure 4.2: paxkm with capacity

As in the previous section several methods have been tested that take capacity into account. The four most important methods are identical to the methods of the previous section. However, this time an extra capacity factor is introduced. Again in formula this can be written as

3. (a)
$$\overline{rpkm} = rpkm^{py} + (paxkm_{ss}^{cy} - paxkm_{ss}^{py}) * capfactor$$
or

(b)
$$\overline{rpkm} = rpkm^{py} * \frac{paxkm_{ss}^{cy}}{paxkm_{ss}^{py}} * capfactor$$

4. (a)
$$\overline{rpkm} = rpkm^{py} + (paxkm_{ss}^{cy} - paxkm_{ss}^{py}) * factor * capfactor$$
or

(b)
$$\overline{rpkm} = rpkm^{py} * \frac{paxkm_{ss}^{cy}}{paxkm_{ss}^{py}} * factor * capfactor$$

, where $capfactor = \frac{seatkm_{ss}^{cy}}{seatkm_{ss}^{py}}$. The results of these four different methods are given in table 4.2

Method	RU	RV	RW
3a	2,17	4,63	2,56
3b	2,58	4,93	3,12
4a	2,08	4,89	2,47
4b	2,36	5,01	3,07

Table 4.2: MAPE scores for paxkm with capacity

4.1.3 Load Factor

As forecasting volume is done by predicting a load factor percentage a natural method is to use load factor curves. This is also how it is mostly done by the analysts. The main ideas are exactly the same as in section 3.1.1 Thus

5. (a)
$$\overline{ALF} = ALF^{py} + (BLF_{ss}^{cy} - BLF_{ss}^{py})$$
or

(b)
$$\overline{ALF} = ALF^{py} * \frac{BLF_{ss}^{cy}}{BLF_{ss}^{py}}$$

6. (a)
$$\overline{ALF} = ALF^{py} + (BLF_{ss}^{cy} - BLF_{ss}^{py}) * factor$$
or

(b)

$$\overline{ALF} = ALF^{py} * \frac{BLF_{ss}^{cy}}{BLF_{ss}^{py}} * factor$$

A big advantage of working with load factor curves is that this automatically takes the capacity into account. A change in load factor of 1% results in a different change in paxkm when capacity is different. The results for these methods are given in the following table.

Method	RU	RV	RW
5a	1,72	1,25	1,58
5b	2,45	1,45	1,76
6a	1,78	1,32	1,43
6b	2,67	1,51	1,72

Table 4.3: MAPE scores for Load Factor

All methods discussed so far have the serious drawback that inventory steering is not taken into account. Inventory steering regulates which bookings are accepted or rejected and therefore has a major impact on booking curves. In general, inventory steering is done in the same way year over year. If bookings are made in the same manner year over year then the booking curve will obviously also more or less be the same year over year. However, if the booking behavior differs year over year inventory steering will cause the booking curve to also be different. For instance, once a flight starts to fill up the inventory analyst will reject more and more bookings. As a result the booking curve will start to slow down, i.e. increase slower. This slowing down effect can result in the booking curve of one year catching up with that of another year.

Figure 4.3: Catching up Booking Curve

Another drawback of the aforementioned methods is that it is possible to forecast paxkm totals which are higher than the available capacity. In other words the load factor percentage is larger than 100%. Although overbooking is common practice at KLM, and therefore load factors larger than 100% can be found in the bookings, an actual flight can never have a total paxkm which is larger than the available seatkm. Therefore the load factor will always be at most 100%.

The final methods for forecasting volume that have been tested were essentially the same as methods 5 and 6. The difference being that now adjustments were made to compensate for the two drawbacks. The first drawback is presumed to occur when the absolute difference between this year's load factor percentage and last year's load factor percentage is larger than a certain threshold value. When this occurs the difference is multiplied with a *catch-up-factor*. Several values have been tested for this catch-up-factor. Also models have been tested with two threshold values. If the absolute difference is larger than both threshold values a certain catch-up-factor is used, if the difference is larger than one value but smaller than the other, another catch-up-factor is used.
An adjustment for the second drawback is made after the initial forecast has been made. If the load factor percentage is predicted to be larger than 100% this percentage is set to a certain value α below 100%.

Table 4.4 shows the results for several combinations of threshold values, catch-up-factors and α's.

T1	T2	CU-factor 1	CU-factor 2	α	RU	RV	RW
2	-	0,5	-	98	1,41	1,6	1,12
2	4	0,5	0,25	98	1,51	1,19	1,33
4	-	0,5	-	98	1,46	1,66	1,21
4	8	0,5	0,25	98	1,56	1,27	1,32

Table 4.4: MAPE scores paxkm using CU-factors

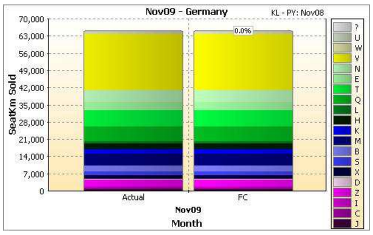
Best method The final method chosen for forecasting paxkm was to forecast at the level of subline using method 5a. Although table 4.4 might suggest another method, the method using two CU-factors was chosen. The corresponding threshold values are 4% and 8%. The reason for this choice was that the MAPE scores were very negatively influenced by two months, namely May 2009 and November 2009. Without these two months the MAPE scores were at least as good as those for the other methods. In fact, when forecasting 2 or 3 months ahead, results were better than for the other methods.

4.2 Yield

The concept of yieldmixes make yield forecasting much more complex than volume forecasting. If, for instance, 100 people would travel to Barcelona in the business class one year and in economy the next their paxkm contribution will be exactly the same. Their yield contribution however will be totally different. Even if they travel in economy class both years their yield contribution can still be totally different. As explained in the section on yieldmixes this can be caused by a different rate of exchange, different ticket prices or a shift in subclass traffic.
These yieldmix factors also complicate aggregation of low level outcomes. Simply adding to low level yield totals does not result in the higher level yield total. Suppose the yield number for revenue group RU, RV and RW are all 10 eurocents/paxkm, then the yield number for KLM total will not be 30 eurocents/paxkm but also 10 eurocents/paxkm. If yield is 10 eurocents/paxkm for RU, 15 for RV and 20 for RW it is not as straightforward. The paxkm ratio's need to be known in order to know the yield number for KLM total.

One mix factor that can be taken out of the equation is the RoX. By using the same RoX's for the current year totals and previous' year totals this mix factor no longer is of any influence. As forecasting is always done in the current year it seems logical convert last year's totals using this year's RoX. It is important to note that since testing is done for months in the past all RoX corrections can be made year over year. For instance, if we want to forecast totals for our test month December 2008, we can convert all December 2007 totals using the RoX's of December 2008. When forecasting months in the future however the RoX details of these future months are not yet available. If a year over year correction can not be made then the RoX of the last available month is used. If today is May 22$^{\text{nd}}$and we want to forecast totals for June 2010, it is not possible to convert totals from June 2009 using the RoX's of June 2010. In this case the RoX's of April 2010 will be used if available, otherwise March 2010 will be used. Afterwards the controller can use the actual RoX from June 2010 to make the correct year over year RoX correction when scoring the forecasts made by the analysts.

Testing yield forecast methods has been done a little bit differently than in the case of forecasting volume. The fact that low level yield totals can not simply be added to obtain higher level yield totals makes it necessary to include paxkm totals. Therefore instead of comparing a forecasted yield number to the actual yield number forecasted revenue totals are compared. For all months in the test set the actual passengers are known and these can be used instead of having to forecast them. A yield forecast can then be made at a chosen level using the actual passengers. These yield forecasts are multiplied with the actual paxkm for that level, and this multiplication gives a forecasted revenue total. Because actual paxkm totals are taken, differences between forecasted revenue totals and actuals revenue totals can only be caused by a difference in yield. The two figures in figure 4.4 illustrate this idea. In the right figure the forecasted revenue totals for Germany are compared to the actual revenue totals. A difference of -1.4% is observed. In the left figure it can be seen that the paxkm totals are all exactly the same. Therefore the difference in revenue must have been caused by a difference in yield.



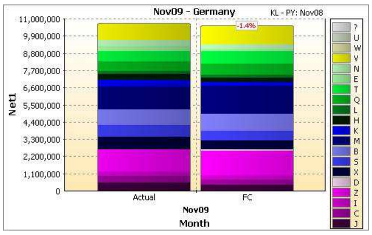


Figure 4.4: Left: actual paxkm vs. fc paxkm, right: actual revenue vs. fc revenue

There are a lot of possible ways of varying the level at which we forecast. Choosing a level within the flight hierarchy is one of them, but there are actually many more possibilities. Other possibilities to differentiate between passengers are

- PoS
- Subclass/cabin
- True O&D
- Traffic type

Every time an extra piece of information is taken into account a new lowest forecast level is defined. Forecasting is always done by first making a yield forecast at the lowest level. Higher level forecasts are subsequently obtained by aggregation of the lowest level revenue forecasts.

Because there are so many possible levels at which to forecast, making the actual forecast was initially done in the simplest way possible

> The absolute/relative difference in yield between current year and previous year of the last data month is added to the previous year of the forecast month

Again this can be written in the following formula,

1. (a)
$$\overline{yield_{fcm}} = yield_{fcm}^{py} + (yield_{ldm}^{cy} - yield_{ldm}^{py})$$
   (b)
$$\overline{yield_{fcm}} = yield_{fcm}^{py} * \frac{yield_{ldm}^{cy}}{yield_{ldm}^{py}}$$

Even this simple way of forecasting has one major problem connected to it. This problem is:

What if one of the input variables is not available ?

If for some reason the historic yield data is not known then the forecasted yield can take on strange, unrealistic values. This problem will seldomly occur when making very high level forecasts. However as more detailed information is taken into account it becomes more likely that certain types of passengers will not occur in some months. For example, if only a forecast is made at the level of revenue group and no other details are taken into account there will always be yield numbers available at this level. If however his subclass, PoS and true O&D are also taken into account then forecasting a yield number can become difficult when his subclass is 'H', PoS is Guadeloupe and True O&D is Buenos Aires to Tokyo. The chance of zero passengers of the same type in all three input months is pretty large. The problem of missing data will be called the *empty bucket problem*.

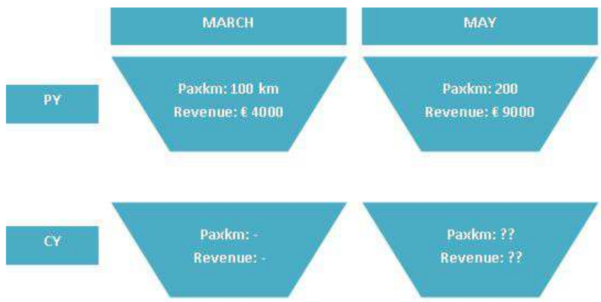


Figure 4.5: Empty bucket

The empty bucket problem is the main reason why True O&D is never taken into account when testing yield forecast methods. So many combinations of origin and destination are possible that the event of an empty bucket becomes a recurring problem. Also for some reason a passenger's origin or destination sometimes is unknown. This makes it impossible to match this passenger with others.

Empty bucket problem Empty buckets occur when yield forecasts are made at a low level. Low level forecasting on the other hand has the great advantage that less factors can influence the yield total. Yield totals at the level of subline, for instance, are not influenced by the linegroup mix or the subline mix. Every extra element of detail that is taken into account eliminates a yieldmix factor. Two different solutions have been tested for the empty bucket problem. In the occurrence of an empty bucket there is the possibility to a) change the forecast method for this passenger type or b) apply the forecast method to a higher level that has the same characteristics and no empty buckets. Figure 4.6 shows an example where both solutions are possible.

Type A has the problem of an empty bucket. Therefore forecast method 1a) is not possible for this type of passenger. The most straightforward way to

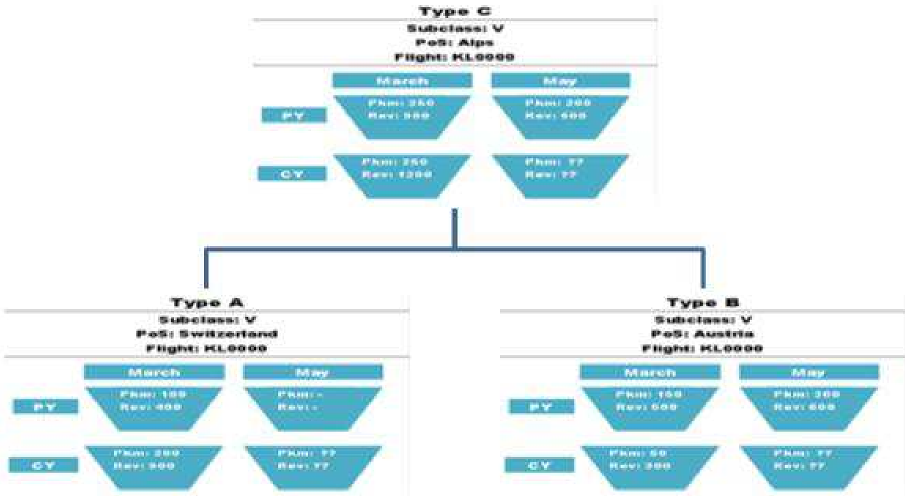


Figure 4.6: Solving empty bucket problem

deal with this problem is to simply use an available bucket to forecast the yield total. Also the average of the two available buckets could be chosen. Finally, perhaps the bucket of two years ago is available and the corresponding yield total could be used to make a forecast. All these options involve changing the forecast method.
The other possibility is to apply forecast method 1a) to Type C. For this passenger type all buckets are available. The yield total that is forecasted for this passenger type is then used as the forecasted yield total for passenger type A. It is true that type A and type C do not exactly share the same passenger details. However the only difference is the PoS and therefore it is reasonable to assume that their yield totals will not be that different. It is important to note that for type B the forecast can be made at this level. As the forecast is made at the lowest possible level, for type B the forecast is made at his own level. This means that aggregating the forecasted revenue totals of type A and type B will not add up to the total used for forecasting type A.
Forecasting at a higher level in the case of empty buckets requires merging of different types of passengers by omitting some point of information. This merging of different passengers can often be done in more than one way. In the example above the choice was made to change the point of sale level. It would also have been possible to change the flight level or perhaps ignore subclass. Every choice of selecting a higher level will result in a different yield total for type A. Sometimes many steps will need to be taken in order to find a passenger type without empty buckets. Every step involves a selection of information that is omitted. The example shows how one choice can be visualized as climbing one step in a *hierarchy tree*. If more than one step is needed the tree will need to be

sufficiently large. The hierarchy is built after all details of the passengers have been determined. Because we are using the actual passengers for forecasting yield, the lowest level of the hierarchy tree can be chosen at the level of flight, subclass and point of sale. The decision how to omit detail determines how low levels are connected to higher levels.

Model selection Many different ways of dealing with empty buckets have been tested. The first main question that needed to be answered was:

How to deal with empty buckets?

The decision was made to focus mainly on changing the forecast level instead of the forecast method. The main reason for this was that it can always happen that there is no (or very little) historic data at all for a certain type of passenger. Changing the forecast method will not help in these situations. Changing forecast level and comparing this type of passenger to a similar passenger will provide a forecast. Another reason for changing forecast level instead of forecast method is that it takes trends and seasonality into consideration. In general, each month and year has its own characteristics and ticket prices fluctuate. Simply using the yield total of a different month would mean ignoring these fluctuations. Finally, a practical reason played a part. The decision which months to include in the forecasting process needs to be done in advance. All data must be stored in Monet before the calculations can be made. At the moment an empty bucket is perceived it is therefore not possible to retrieve information from months that have not already been stored. A simple solution would be to store all months in Monet before starting the calculations. This is unfortunately not a desirable option as testing will become very slow due to the enormous amount of data that is stored.

The first important result found was that dropping subclass first resulted in bad forecasts. Table 4.5 shows some results.

Flight level	MAPE RU	RV	RW
L2	3,22	2,89	2,89
L4	2,63	4,3	2,71
L6	2,07	2,64	2,27

Table 4.5: MAPE scores when subclass is dropped first

This seems to make sense. Price ranges of subclasses can vary greatly. Using a high yielding subclass like 'S' to forecast yield for a passenger in the low yielding subclass 'V' will probably produce unrealistic outcomes. Also the difference between economy class yield and business class yield is large. Dropping cabin too early is therefore not sensible.

Raising the forecast level will therefore have to be done by raising either the PoS level or the flight level. Table 4.6 shows the test results for two different methods.

Method	RU	RV	RW
PoS first	1,29	1,39	2,14
Flight first	1,27	1,39	2,08

Table 4.6: MAPE scores when changing forecast level

The first method is to raise the PoS-level until level PoS5 is reached. If there still is an empty bucket then the flight level is raised. This process continues until there are no empty buckets. The second method is the opposite. First the flight level is raised until it reaches level L5 and then the PoS-level.
It is also possible to use a combination of both methods. For instance, first the PoS-level is raised. If there are still empty buckets at level P5, the PoS-level is set back to P0 and the flight level is raised one level. In case of an empty bucket the PoS-level is once again raised until it reaches P5. The PoS-level is reset to P0 and the flight level is once again raised, etcetera. Test results for this method are given in table 4.7.

Method	RU	RV	RW
PoS first, Flight second	1,27	1,4	2,03
Flight first, PoS second	1,22	1,16	2,00

Table 4.7: MAPE scores for combined methods

4.3 Revenue

4.3.1 Combining Volume model with Yield model

In sections 4.1 and 4.2 the best methods for forecasting volume and yield were determined. The main problem when combining these two models is that the forecasts are made on different levels. Paxkm forecasts are made on the level of subline, but without the detail of PoS and Subclass. These details are used when forecasting yield. Thus, a decision has to be taken whether to forecast yield at the level of subline or to distribute the forecasted paxkm at subline level over lower level buckets.
The first method is relatively straightforward. The best method for forecasting paxkm, as suggested in section 4.1, is used. For yield forecasts both methods suggested in section 4.2 have been tested. The MAPE results for the combination of these two methods are given in table 4.8

Method	RU	RV	RW
1a	4,58	2,83	4,50
1b	4,63	3,31	4,74

Table 4.8: MAPE scores when forecasting yield at subline level

Table 4.8 also shows the MAPE results if the paxkm forecasts are made at a higher level. It is clear that these results are not satisfactory. Underlying mix factors are not taken into account anymore and as a result the revenue forecasts are inaccurate.

In order to achieve accurate forecasts it is therefore necessary to distribute the forecasted paxkm over the lower level buckets. A possible method for this distribution is to forecast the shares of the bookings (SoB) that are made in a certain subclass, cabin or PoS.

Actual subclass and cabin ratio's Before deciding whether to try to forecast subclass SoB's or Cabin SoB's tests were done with actual subclass and cabin SoB's. This testing worked as follows:

1. Forecast the number of paxkm on the level of subline, $\overline{rpkm}$.
2. For all subclasses prevalent in the current bookings determine the share of the bookings (SoB) percentages as they actually ended up.
3. Forecast paxkm totals based on these percentages, $\overline{subclrpkm} = \overline{rpkm} * \frac{SoB}{100}$
4. Calculate subclass factor, $subclassfactor = \frac{\overline{subclrpkm}}{subclpkm_{ss}}$
5. Calculate correction factor, $corr = \frac{\overline{rpkm}}{\sum \overline{subclrpkm}}$
6. For all types of bookings the forecasted number of paxkm is, $\overline{rpkm} = pkm_{ss} * corr * subclassfactor$

Step 4 is necessary due to the fact that in step 2 only those subclasses that are already observed in the current bookings are forecasted. If for a certain subclass no bookings have been made yet then the forecasted number of bookings will also be set to 0. As a result the share of bookings percentage is also 0. However if in fact there were some passengers in this subclass this percentage was larger than 0. Only using the subclassfactor for the other subclasses would now result in a lower paxkm total at the subline level, because the percentages don't add up to 100%. This difference is adjusted using the Paxkm factor. It should be noted that due to this adjustment the percentages may not exactly be the same as the actual percentages. The small example given in table 4.9 illustrates the testing method as well as the problem of changing percentages.

subcl	pkm_{ss}	$rpkm$	SoB	$\overline{rpkm}$	subclfactor	corr	$\overline{rpkm}$
E	100	300	37,5%	270	2,7	1,1429	308,6
N	200	400	50%	360	1,8	1,1429	411,4
V	0	100	12,5%	0	0	0	0

Table 4.9: Forecasted rpkm = 720 km

The actual number of paxkm for this subline was 800 km distributed over the three subclasses. The forecasted number was 720 km. At the moment of forecasting no bookings had been made for subclass V. Therefore no paxkm are forecasted for this subclass and the 720 paxkm will have to be distributed amongst the 2 remaining subclasses. However using the actual subclass ratio's results in a paxkm total of $270 + 360 = 630km$ which is less than the desired total of 720 km. The Paxkm factor,$\frac{720}{630} = 1,1429$ is calculated to correct this difference.
The reason for using the this method is that this makes it possible to use the PoS information of the bookings. For instance, if the 100 km of subclass E are divided 70:30 over PoS's Germany and USA then the forecasted number of paxkm will be respectively $70 * 2,7 * 1,1429 = 216$ km and $92,6$ km. This makes it possible to forecast yield at a more detailed level, which has been found to be more accurate. Table 4.10 shows some results of these testing methods. As always these results are MAPE scores for revenue forecasts at the level of the revenue groups.

Testing Method	RU	RV	RW
Actual Subclass ratio	2,74	1,95	2,77
Actual Cabin ratio	4,29	3,46	4,62

Table 4.10: MAPE scores when actual cabin ratio or subclass ratio are used

Clearly, results are better when the actual subclass ratio is used instead of the actual cabin ratio. Unfortunately forecasting the subclass ratio is much more complex than forecasting the cabin ratio.

Forecasting subclass or cabin SoB's The easiest method for forecasting subclass or cabin shares of passengers is to use the shares of the bookings. In other words the actual share of passengers for a subclass is the same as the SoB-percentage at the moment of snapshot. The results for this method proved unsatisfactory. An important reason for this is that late high yielding bookings are not taken into account.
A method for forecasting these ratio's that does take this late booking behavior into account is the day-before-period (DBP) method. This method is similar to the method for forecasting volume. For multiple months the booking information, specifically subclass ratio or cabin ratio, is compared. The idea is to forecast the actual subclass or cabin ratio by analyzing what happened in previous months from the same number of days before those months. In Appendix B a more lengthy explanation is given about the details of this DBP method. Also the function that has been created in Monet to facilitate this analysis is explained.
The two main methods that have been tested using the DBP function only use the booking information at the moment of snapshot (SS) and the Actual data (Act) as variables. In formula these can be written as :

- **Method 1:** $Act^{cy} = SS^{cy} + Act^{py} - SS^{py}$

- **Method 2:** $Act^{cy} = SS^{cy} * \frac{Act^{py}}{SS^{py}}$

The variables Act and SS can be a subclass SoB-percentage or a cabin SoB-percentage. Immediately some problems arise with these two methods. Method 1 has the possibility of forecasting negative percentages and method 2 runs the risk of dividing by zero if at the moment of snapshot there are no bookings for this subclass/cabin. In these cases another method should be chosen to determine SoB-percentages. In general, the alternative method chosen was to use the actual percentage of a month in the past as the forecasted percentage. Once the subclass SoB-percentages have been forecasted the same steps are applied as in the testing method.

1. Forecast the number of paxkm on the level of subline, $rpkm$.
2. Forecast SoB-percentages using DBP function.
3. Forecast paxkm totals based on these percentages, $\overline{subclrpkm} = \overline{rpkm} * \frac{SoB}{100}$
4. Calculate subclass factor, $subclfactor = \frac{subclrpkm}{subclpkm_{ss}}$
5. Calculate paxkm factor, $corr = \frac{\overline{rpkm}}{\sum \overline{subclrpkm}}$
6. For all types of bookings the forecasted number of paxkm is, $\overline{rpkm} = pkm_{ss} * corr * subclfactor$

After the paxkm have been forecasted the next step is to forecast yield. This is done at the lowest available level of detail. Some test results are given in table 4.11.

Forecasted ratio	RU	RV	RW
Subclass	3,83	2,67	3,43
Cabin	4,72	4,13	4,92

Table 4.11: MAPE scores when using method 2

Closer examination of these results showed that a major problem occurred. Cabin ratio's can be forecasted with greater accuracy than subclass ratio's. This is partly due to the fact that there may be several subclasses that have no bookings yet and are therefore forecasted to be 0. Cabins without bookings are very uncommon. Yield forecasting at the level of cabin however is less accurate than at the level of subclass. This has lead to the idea of using *clusters* of subclasses.

Cluster methods Combinations of subclasses joined together are called clusters. These clusters can be chosen in every possible way, although it makes sense to only cluster subclasses when they are related in some way. For instance, the C and M cabins can technically be seen as clusters. Here subclasses are joined together on the basis of business class or economy class conditions. The possibility of clustering was already available in DeLorean. In Monet it unfortunately was not yet possible to cluster subclasses. Therefore clusters have only been used when analyzing volume. In the new version of Monet it now is possible to analyze revenue information, like yield, using clusters. This makes it possible to apply the same testing and forecasting methods to cluster ratio's as discussed for subclass and cabin ratio's. Several different sets of clusters have been examined. A natural way of creating clusters is to define yield categories. Economy class is split into low yield, medium yield and high yield. Business class is split into low yield and high yield. In practice this results in the following clusters:

- High Yield Business = Subclasses J and C
- Low Yield Business = Subclasses I,Z and D
- High Yield Economy = Subclasses X,S,B and M
- Medium Yield Economy = Subclasses K,H,L,Q and T
- Low Yield Economy = Subclasses E,N,V,U and others

The forecast results using these clusters are shown in table 4.12.

Months ahead	RU	RV	RW
1 month	3,2	2,0	3,6
2 months	4,4	3,4	4,8
3 months	4,8	4,3	5,2

Table 4.12: MAPE scores when using 5 clusters

These results of forecast method 2 are very encouraging. A slight improvement was found when the C cabin was not split into categories. Ergo, subclasses J,C,Z,I and D are grouped together in one cluster. The results of using forecast method 2 with these clusters are given in table 4.13.

Months ahead	RU	RV	RW
1 month	3,0	2,0	3,0
2 months	4,2	2,9	4,6
3 months	4,4	4,0	5,1

Table 4.13: Best MAPE scores

4.3.2 Other models

Many different types of models have been studied and tested during the research of this project. The main ideas have been discussed in the previous section. Some ideas are not worthy of being discussed in this thesis. However two ideas should be mentioned. First of all, all methods discussed so far only use a minimal amount of input variables. One snapshot moment or one revenue month, but never more than that. It seems logical to use more information if this is available. Some tests have been done using more information than this minimal amount. These were not encouraging and therefore this line of investigation was stopped. One example of forecasting using more information is a test done for forecasting volume. In this case not one but two snapshots were used. The booking curves were studied at these two moments. If the difference between this year's curve and last year's curve had become smaller from the first snapshot moment to the second this would imply the current curve was catching up. The forecasted total for this year would then be closer to last year's total than it was at the last snapshot moment. Conversely if the difference had become larger, the final difference would also be forecasted to be larger. In formula,

- $Trend = (BLF^{cy}_{ss2} - BLF^{py}_{ss2}) - (BLF^{cy}_{ss1} - BLF^{py}_{ss1})$.
- $LF^{cy} = LF^{py} + factor * (BLF^{cy}_{ss2} - BLF^{py}_{ss2})$, where $factor > 1$ if $Trend > 0$ and else $0 < factor \leq 1$.

The scaling factor was determined in several different ways. However none of them produced better results than the methods using the minimal amount of information. Similar experiences were encountered when forecasting yield using more information. More information did not directly lead to better forecast results. An example of this is given in table 4.14 where paxkm were predicted using this scaling factor.

Months ahead	RU	RV	RW
1 month ahead	1,82	1,45	1,68
2 months ahead	2,41	2,77	2,38
3 months ahead	2,81	23,13	2,83

Table 4.14: Paxkm using scaling factor

Another line of investigation was to work with small buckets instead of empty buckets. The idea behind this *small bucket problem* was that it is not sensible to make forecasts based on a very small number of people, but a minimum number $\alpha$of passengers is required to make a reliable forecast. This means that if for some input variable the data is based on less than α passengers the level at which the forecasts are made has to be raised.
Again a couple of values of α have been tested, but forecast results were not found to improve.

α	RU	RV	RW
10	3,3	2,0	3,9
20	3,5	2,1	4,0
50	3,9	2,4	4,3

Table 4.15: Cluster method using different values for α, 1 month ahead

Chapter 5

Conclusions and recommendations

When trying to determine the best forecast method it can be comforting, as well as frustrating, to realize that it is impossible to exactly predict the future. No single method is perfect. However, the results of the research done during this project do offer some conclusions as well as recommendations for further research and possible improvements in the forecasting process.

5.1 Forecast methods

After all methods have been tested the best forecast results were produced with the cluster method consisting of 4 clusters. These four clusters are business, economy low yield, economy medium yield and economy high yield. The only exception is found when forecasting 1 month forward for revenue group RV. In this case splitting the business cluster into high yield and low yield produces slightly more accurate results. Besides the MAPE scores the number of times the forecast is within the allowed error margin has been counted. Table 5.1 shows these scores as well as the scores of the analysts for the same months.

Months ahead	RU	RV	RW
1 month	8	14	6
2 months	15	18	13
3 months	16	18	17

Months ahead	RU	RV	RW
1 month	9	8	10
2 months	12	11	14
3 months	15	11	17

Table 5.1: Left: Best scores after research, right: Analysts's scores

As can be seen in the table, in 6 out of the 9 cases the suggested forecast methods performed at least as well as the analysts did for these months. Lower level comparisons are not as easy to make as linegroups are not defined in Monet.

It is possible to study the forecast results at the level of region. The MAPE scores are given in table 5.2.

Months ahead	Europe	AP	Africa	MESA	CSA	NA	India
1 month	3,0	2,2	3,8	4,2	5,1	3,8	5,9
2 months	4,2	4,7	4,3	6,6	5,2	4,5	6,7
3 months	4,4	7,2	5,2	7,7	5,7	5,4	9,4

Table 5.2: MAPE scores for regions

As might be expected the MAPE scores are worse than at the levels of revenue group. This is due to the greater variability at lower levels.

The real test however is to calculate these scores for the validation set. This validation set consists of the months January 2007 to December 2007 and December 2009 to March 2010. The results for the validation set as well the scores of the analysts are given in table 5.3.

Months ahead	RU	RV	RW
1 month	5	7	8
2 months	9	11	11
3 months	9	11	11

Months ahead	RU	RV	RW
1 month	7	4	5
2 months	10	13	10
3 months	12	14	11

Table 5.3: Left: Best scores after research, right: Analysts's scores

The results for the validation set are already pretty encouraging. The 1,5% margin for forecasting one month ahead proves very difficult to achieve. The relative errors are smaller than when forecasting two or three months ahead, but not yet small enough to come within the bound. The results for forecasting two or three months ahead are even more encouraging. Taking a closer look at the results shows another very interesting fact.

Table 5.4 shows the relative errors when forecasting two months ahead. Of the twelve months in 2007 the suggested method is within the allowed margin of 4% in 11 out of 12 cases for RV and RW. For RU this is a little bit less, but still a score of 9 out 12 is reached. Unfortunately, the last four months of the validation set show very poor results. The forecasts are much too low. In some cases more than 15% too low! The exact same pattern can be seen when forecasting one and three months ahead. All points have been scored for the months of 2007. For the remaining four months very large deviations were found.

There are two reasons for the poor forecast results from December 2009 to March 2010. One reason is the fact that the peak of the financial crisis was reached in the months December 2008 to March 2009. Yield numbers deteriorated rapidly during these months instead of showing the "normal" patterns. As these months are important input variables when forecasting December 2009 to March 2010 and the resulting forecasts are too low. The second reason is that ticket prices have been raised again as the economy started to recover. It is not possible for the forecast method to foresee these price changes.

Months ahead	RU	RV	RW
Jan 2007	-5,0	-1,6	-2,3
Feb 2007	0,1	1,9	2,5
March 2007	-0,7	-4,5	-3,0
April 2007	-3,2	-1,2	-3,0
May 2007	-1,0	2,4	5,9
June 2007	-3,2	1,5	2,3
July 2007	-0,2	-1,3	-0,3
Aug 2007	-1,4	-2,9	-0,5
Sep 2007	-3,5	-0,2	3,2
Oct 2007	-6,8	-1,2	1,5
Nov 2007	-6,2	2,0	2,1
Dec 2007	-3,9	3,6	1,0
Dec 2009	-10,5	-9,5	-16,8
Jan 2010	-9,1	-8,1	-18,1
Feb 2010	-7,7	-11,4	-15,8
March 2010	-9,6	-14,4	-19,1

Table 5.4: Percent Errors for validation set

The main goal of this project was to examine if, given the very detailed data available nowadays at KLM, it is possible to develop a forecast method that improves the accuracy of short-term forecasting within KLM. The research done during this project is a first step in showing that this is possible. The results at the level of linegroup are already very encouraging. It is now also possible to view in more detail the build-up of the forecasts.
The main problem with the current method is the lack of flexibility. The suggested forecast methods have been programmed in Monet and therefore have a fixed structure. However, sometimes it might be useful to make small adjustments to the forecast methods in order to get better results. A good example of this can be seen when studying the results for the months December 2009 to March 2010. The forecast results were far below the actual outcomes. For these months it might be better to use the data of two years ago as input variables. Also a single extraordinary month may negatively influence forecast results. For instance, April 2010 is not a sensible input variable due to the very uncommon revenue results caused by the ash cloud from Iceland. In this case it may be better to use March 2010 as an input variable. Currently it is not possible to make these types of adjustments in the forecast function in Monet.
It could also be useful if it was possible to use expert knowledge a little bit. Load factor forecasting, for instance, is something that is done quite well by the analysts. It would be nice if an analyst could enter his own LF forecast and see what happens to the final result. The same holds for price changes. If the analyst knows a price change is being made by KLM, he should be able to adjust the forecast method accordingly.
For the current forecast function in Monet only one adjustment can be made

by the user. This is the grouping of the subclasses, the clusters. Editing the clusters can be done in the filter function in Monet. All in all, the next step should be to expand the current forecast function and make it more interactive. At this moment advice is to let the analysts use the suggested forecast methods as a different perspective. When forecast totals are very different from their own forecast totals, this could be a signal to look at their forecasts a little bit closer. The analysts should always be ware that the forecasted totals in Monet already include the RoX factor. Therefore they should use their own forecasts combined with the controller's RoX factor. This advice is in fact similar to a study by prof. Franses from the Erasmus University that was done in 2009 at KLM. His advice was "to let the experts use a simple model and add their managerial intuition, but add in a symmetric way"[5]. The final part of this advice refers to the fact that analysts tend only to adjust their forecasts when they are lower that that of the model. As a result forecasts tend to have a bias for overestimating.

5.2 Other recommendations

Scoring At this moment it is a little bit difficult to score the forecasts made by Monet. One reason for this is that in the programming code the number of months that is looked ahead is fixed for months in the past. This means that, if one wants to check the forecast made for March 2010, the program immediately assumes that the forecast date is February 19th 2010. If the forecast for March 2010 of January 22ndis wanted this can only be found by changing the programming code. It is not very difficult to make this more simple and adjust while using Monet. For months in the future there is no problem.

The RoX correction is automatically done in Monet for months in the past as for these months the actual RoX's is now available. This is in contrast to future months for which the RoX's are not yet known. If at some point it is possible for the analysts to make their final forecasts in Monet, this would automatically imply that the RoX factor would no longer need to be forecasted by the controllers.

The capacity correction is also easily done in Monet. A function has been created that can make the switch between the capacity at the moment of forecasting and the final actual capacity. In contrast to the ASK correction made by the controllers the ASK correction in Monet is done at the level of subline. This is done because the forecasts are also made at this level. If in the future the analysts would also make these lower level forecasts it would only be fair to make the ASK corrections at the corresponding level.

Fuel surcharges All revenue forecasts have been made for Net1 revenue totals. Some research into forecasting revenue totals earned from fuel surcharges has been done during this project. As this was not within the scope of this project this research was not followed up on. Improvements can be made in the accuracy of these fuel surcharge forecasts. The main idea worth investigating would be to combine forecasted volume totals, using the methods suggested in this thesis, with the Decision Support tool that contains all possible fuel surcharge rates. This would not be very difficult to implement, and might well improve accuracy greatly.

Unused data sources Finally, the main reason for this project was to study if the large amounts of data available at KLM could make it possible to produce more accurate and more structured forecasts. The research done during this project is an indication that this is possible. It is therefore very unfortunate that not all data sources at KLM were made available for the research. More up to date information like *daily revenue* and *forward looking revenue* will undoubtedly even further improve the accuracy of forecasting. Follow up research should therefore focus on gaining access to these data sources.

Appendix A

Yieldmix

A.1 Method for calculating yieldmixes

The reason for calculating yieldmixes is to quantify the different factors that cause a change in yield. This change is usually compared year over year, but also different comparisons could be made. Factors that cause these yield changes can be a change in cabin ratio, PoS ratio, RoX, subline ratio, etcetera. Step by step a factor is singled out and the change in revenue caused by this single factor is determined. The difference between the total change in revenue and the change caused by this single factor must then be caused by the remaining factors. This process is then repeated until all factors have been examined.

Unfortunately, the mix factors are not uncorrelated. This means that the order in which the different factors are chosen influences the outcome of the corresponding mixes. The easiest way to understand this is to consider the example of Houston. There are two sublines that fly to Houston, the D40 and D42. The D42 subline is a business class subline only, i.e. only business class tickets can be bought for this subline. Suppose the KLM decides to stop flying the D42 subline this will most likely have a negative effect on the yield totals for Houston. But is this negative effect caused by the change in cabin ratio or by the change in subline ratio? If the cabin mix is calculated first, the majority of the yield change is attributed to it and a small part to the subline mix. Conversely, if the subline mix is calculated first it will be the other way around. The majority of the yield change will now be attributed to the subline ratio change and a smaller part to the change in cabin ratio. However, the sum of both mix factors must be the same for both cases.

The first step in calculating yield mixes is to determine last year's yield, $yield_e^{py}$, and this year's yield, $yield_e^{cy}$. The subscript is added to distinguish if yield is given in € or in its local currency (lc). Using these two numbers the year over year absolute and relative yield change can be calculated. If all mixes are calculated these should at up to the relative yield change.
There are three different cases for calculating the actual yield mixes. The RoX-

mix and Price mix have their own specific method. For all other mixes there is a general method. Each method will be clarified in a mathematical formula as well as in words.

- **RoX-mix**:
  1. $RoX\_net1 = (RoX^{cy} - RoX^{py}) * rev_{lc}^{cy}$
  2. $Rox\_net = \frac{\sum RoX\_net1}{paxkm^{cy}}$
  3. $RoXmix = \frac{RoX\_net}{yield^{py}} * 100$
- **Price mix**
  1. $pr\_net1 = (yield_{lc}^{cy} - yield_{lc}^{py}) * paxkm^{cy} * RoX^{py}$
  2. $pr\_net = \frac{\sum pr\_net1}{paxkm^{cy}}$
  3. $Pricemix = \frac{pr\_net}{yield^{py}} * 100$
- **General method**
  1. $gen\_net1 = (yield_{e}^{cy} - yield_{e}^{py}) * paxkm^{cy} - RoX\_net1$
  2. $gen\_net = \frac{\sum gen\_net1}{paxkm^{cy}}$
  3. $Generalmix = \frac{gen\_net}{yield^{py}} * 100$

The main difference between the three methods is the level at which the first step is performed. The RoX-mix and price mix are both calculated at the lowest level of detail. All other mixes are calculated one level above the lowest level of detail. When a certain yield mix is calculated the corresponding factor is not taken into account anymore. The lowest level of detail is thereby raised one level and the next yieldmix can be calculated. Although the order in which yield mixes are calculated is more or less arbitrary, the raising of the lowest level of detail necessitates that the RoX mix and price mix are calculated first. Also the subline mix must be calculated before the linegroup mix.
In all three methods the first step is to calculate the difference in revenue caused by a single factor. For the RoX-mix this is done by multiplying the difference in RoX with the actual earned local revenue. For the other two factors the difference in yield is multiplied with the current year's paxkm. For the price mix the yield difference in local currency is used, for the others the yield in € is used.
Steps 2 and 3 are the same in all cases. All lower level revenue changes are added up to get the revenue change at the highest level. This sum is then divided by total number of paxkm at the highest level, to obtain the the absolute yield change at the highest level. Finally, the yield mix factor is found by dividing this absolute change in yield by last year's total yield.
It is important to always check if the correct units are used. Revenue totals are generally given in €. On the other hand the KLM uses $eurocents/paxkm$ for calculating yield. This means either the revenue totals should initially be multiplied with 100 or for these calculations yield is given in $€/paxkm$.

					rev_e^{cy}	$paxkm^{cy}$	rev_e^{cy}	$paxkm^{py}$	rev_{lc}^{cy}	rev_{lc}^{py}	pr_net1	RoX_net	PoS_net1	Cabin_net1	Subline_net1	Linegroup_net1
Subline	A05	C	1	USA	6701	6705	7381	9387	3953	4713	919	510				
			431	Switzerland	32029	40827	19975	25479	28546	17830	-27	49				
			530	GB	218137	268431	156165	199926	80482	63115	-10539	19001				
			900	Netherlands	250475	249204	234958	255139	250475	234957	20984	-1				
					507342	565167	418479	489931	363456	320615	11336	19559	5041			
		M	1	USA	1966	5811	2014	6473	1158	1281	13	145				
			431	Switzerland	25853	47382	11620	24910	23042	10373	3709	41				
			530	GB	349928	972831	225394	567319	146225	90735	-23266	31692				
			900	Netherlands	188440	384201	143835	321840	188440	143834	16736	-1				
					611187	1410225	382863	920542	358865	246223	-2808	31877	-7217			
Sublinetotal					1118529	1975392	801342	1410473	722321	566838	8527	51436	-2177	-55200		
	A06	C	1	USA	28566	40030	16713	24864	16808	10661	-558	2216				
			431	Switzerland	109007	145299	75183	106494	97154	67132	6227	202				
			530	GB	352305	534944	245118	411127	130082	98893	3485	29881				
			900	Netherlands	287095	383801	337304	497165	287095	337302	26705	-2				
					776973	1104074	674318	1039650	531139	513988	35859	32298	28572			
		M	1	USA	19709	65454	9359	33643	11599	5957	15	1486				
			431	Switzerland	39174	83638	42132	94280	34915	37614	1733	65				
			530	GB	394630	1781171	319197	1308931	146860	128256	-68860	29132				
			900	Netherlands	156547	399214	131716	373564	156547	131715	15788	-1				
					610060	2329477	502404	1810418	349921	303542	-51325	30682	-67069			
Subline total					1387033	3433551	1176722	2850068	881060	817530	-15466	62980	-38497	-93574		
Linegroup total					2505562	5408943	1978064	4260541	1603381	1384368	-6938	114416	-40674	-148775	-120092	
Subline	F00	C	1	USA	75915	284267	64528	301300	44656	41160	9129	5906				
			431	Switzerland	232781	864762	150237	558886	207472	134140	-92	412				
			530	GB	204384	664207	305478	1339459	75505	122705	36493	16412				
			900	Netherlands	865488	2611327	772172	2666267	865488	772167	109233	-6				
					1378568	4424563	1292415	4865912	1193121	1070172	154762	22725	180653			
		M	1	USA	8730	84909	18142	175295	5135	11607	-761	704				
			431	Switzerland	254876	3641245	446554	7157877	227218	398677	27340	371				
			530	GB	292441	5190219	221251	4057016	109974	88592	-8400	17790				
			900	Netherlands	410922	5272687	360465	4042111	410922	260463	-59280	-2				
					966969	14189060	1046412	15432299	753249	859339	-41101	18863	-14006			
Sublinetotal					2345537	18613623	2338827	20298211	1946370	1929511	113661	41588	166647	159226		
Subline	F01	C	1	USA	27385	117356	17598	59304	16109	11167	-9438	1999				
			431	Switzerland	205116	734197	150287	531016	182820	134185	-3033	358				
			530	GB	203709	608816	171445	579027	75332	69183	6418	17026				
			900	Netherlands	928482	2713317	560604	1753193	928482	560600	60874	-7				
					1364692	4173686	899934	2922540	1202743	775135	54820	19376	60118			
		M	1	USA	9205	48865	7537	78624	5404	4739	3910	610				
			431	Switzerland	316314	5624233	423295	6418740	281972	377903	-55059	473				
			530	GB	262054	3847380	160841	2378759	97983	64553	-16007	17918				
			900	Netherlands	351587	3779933	246990	2533763	351587	246989	-16878	-1				
					939160	13300411	838663	11409886	736946	694184	-84033	19000	-57463			
Subline total					2303852	17474097	1738597	14332426	1939689	1469319	-29213	38376	2655	145778		
Linegroup total					4649389	36087720	4077424	34630637	3886059	3398830	84448	79964	169302	305004	320443	
KLM total					7154951	41496663	6055488	38891178	5489440	4783198	77510	194380	128629	156229	200351	499400

Table A.1: Example of yieldmix calculation

Example In table A.1 an example is shown of a yieldmix calculation. The revenue totals are given as well as the outcomes of the first steps of the mix factor calculations. Some examples of how these first steps are calculated are given below:

- The first RoX_net1 total of 510:

$$(\frac{6701}{3953} - \frac{7381}{4713}) * 3953 = 510$$

- The first Price_net1 total of 510:

$$(\frac{3953}{6705} - \frac{4713}{9387}) * 6705 * \frac{7381}{4713} = 919$$

- The first Cabin_net1 total of -55200:

$$((\frac{1118529}{1975392} - \frac{801342}{1410473}) * 1975392) - 51436 = -55200$$

In table A.2 the actual yieldmixes are calculated.

Mix factor	Calculations	Outcome
RoX-mix	Step 2 : $\frac{194380}{41496663} * 100$	0.468423
	Step 3 : $\frac{0.468423}{15.57034} * 100$	3.008431%
Price mix	Step 2 : $\frac{77510}{41496663} * 100$	0.186785
	Step 3 : $\frac{0.186785}{15.57034} * 100$	1.199622%
PoS mix	Step 2 : $\frac{128629-77510}{41496663} * 100$	0.123188
	Step 3 : $\frac{0.123188}{15.57034} * 100$	0.791173%
Cabin mix	Step 2 : $\frac{156229-128629}{41496663} * 100$	0.066512
	Step 3 : $\frac{0.066512}{15.57034} * 100$	0.427174%
Subline mix	Step 2 : $\frac{200351-156229}{41496663} * 100$	0.106326
	Step 3 : $\frac{0.106326}{15.57034} * 100$	0.682877%
Linegroup mix	Step 2 : $\frac{499400-200351}{41496663} * 100$	0.720659
	Step 3 : $\frac{0.720659}{15.57034} * 100$	4.628408%

Table A.2: Yieldmixes

Because all revenue totals in table A.1 are given in € an extra multiplication with 100 is performed in step 2. This ensures that the absolute change in yield caused by the corresponding mix factor is given in the standard unit of $eurocents/paxkm$.

A.2 Function in Monet

In Monet a function has been created that visualizes the impact of the different yieldmixes. This function is more or less similar to an already existing website maintained by the revenue accounting department at KLM. The mix factors which are calculated are:

1. RoX-mix
2. Price mix
3. PoS mix
4. Cabin mix
5. Subline mix
6. Linegroup mix

This order is also the order in which the calculations are carried out. As explained in the previous section different mix factors can be chosen as well as a different order. If it would be considered better to change the order or the mix factors, this would not require a lot of adjusting of the programming code. For now, the revenue accounting website has been used as a guide.

Most of the options available in Monet are also available within the yieldmix function. An example is given in figure A.1.

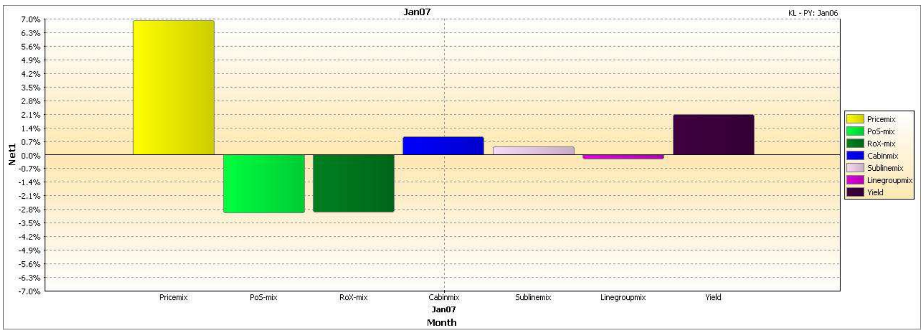


Figure A.1: Yieldmix function in Monet

The biggest advantage compared to the existing website is the availability of these Monet options. Not only can all flight levels be chosen, but a yieldmix analysis can also be made for all combinations of subclass/PoS/period available in Monet.

It is important to note that there are some small differences in the results as shown on the website and in Monet. One reason for this is the fact that different data sources are used. The revenue totals of both data sources are slightly different and as a result the yield factors are also slightly different.

These differences are usually no more than a couple of tenths of a percent. Larger differences have been observed in the Price mix and the PoS mix. The cause of these differences is not known. However the method implemented in Monet has been carefully check and its correctness verified.

Appendix B

Day before period

A frequently used method for forecasting revenue at KLM is to compare booking patterns of different months. This is done by comparing booking curves, but also by analyzing shift in subclass percentages or cabin percentages. The ratio's observed at the moment of forecasting will be different from the eventual actual ratio's. A common phenomenon is that business class passengers have a tendency make their bookings later than economy class passengers. This means that the actual percentage of business class passengers is usually higher than at the moment of forecasting. To analyze these shifts in percentages a function has been created in Monet called the *Day-before-period function* (DBP-function). The DBP-function shows the actual flown passengers for a certain period or periods as stored in Monet as well as the booking information at a certain number of days before these same period(s).

Figure B.1 illustrates a 10 day before period comparison.

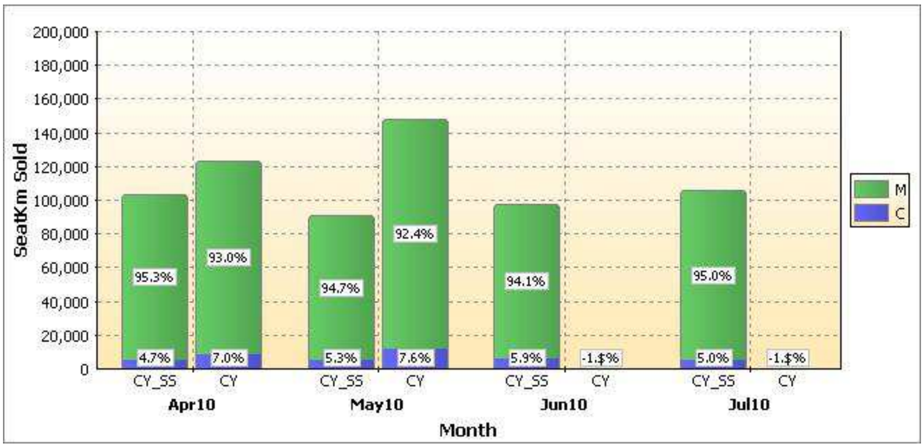


Figure B.1: 10 Day Before Period example

For each month the booking information is shown as it was 10 days before that month. For example, for April the booking information of the 21st of March is shown. For April and May the actual flown passenger information is also shown. The fact that this information is not shown for June and July implies that these totals were not yet available. This is due to the fact that this graph was created June 22nd.

When working in Monet the number of days before the period can be varied by varying the snapshot date. The difference between the first day of the first period and the snapshot date determines the number of days before period. To view figure B.1 in Monet the snapshot date must therefore be set to the 21stof March. The difference between the 21st of March and the 1st of April is 10 days. In similar fashion for these four months a 40 day before period can be obtained by setting the snapshot date to the 19th of February. If the snapshot date is set to the 22nd of March, the snapshot bar of July will disappear, because the snapshot of June 22nd is not available yet. This snapshot will become available on the 23rd of June.

Bibliography

[1] Koole G., *Optimization of Business Processes: An Introduction to Applied Stochastic Modeling*, 2009

[2] Hanea A., *Algorithms for Non-parametric Bayesian Belief Nets*, 2008

[3] Wikipedia, `http://en.wikipedia.org/wiki/Time_series`

[4] Kahn K.B., *Revisiting Top-Down Versus Bottom-Up Forecasting*, The Journal of Business Forecasting, 1998

[5] Franses P.H., *Decomposing bias in expert forecasts*, 2009

[6] Hendriksz R, *Airline Short-Term Revenue Forecasting*, 2005

[7] Hastie T., Tobshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, second edition, 2009

Index