

Airneth research memorandum 004-007

Solving the lack of price data availability in (European) aviation economics?

Guillaume Burghouwt¹
Aleid van der Flier
Jaap de Wit

Version 20 June 2007

¹corresponding author
SEO Economic Research/Airneth
Roetersstraat 29
1018WB Amsterdam
The Netherlands
g.burghouwt@seo.nl
www.seo.nl
www.airneth.com

Paper to be presented at the 2007 ATRS World Conference, 21-24 June 2007,
Berkeley, USA

Abstract

This paper introduces a new and innovative method to solve the lack of ticket price data availability in aviation economics, by using internet booking websites. Ticket price databases do exist but cover primarily the US domestic and US international markets (e.g. 10% ticket sample). Outside the US, 'umbrella' databases on airfares are not accessible to researchers or are scarce and incomplete. The data handicap makes a number of analyses, such as the computation of welfare impacts of airport projects/mergers/network changes as well as price monitoring and benchmarking exercises difficult.

This paper presents a new way of filling in the price data gap by introducing a web mining methodology in aviation economics, called *Avifare*. For both researchers and policy makers, *Avifare* may help to solve the lack of price data availability.

Using an information extraction agent, *Avifare* automatically fills in internet booking websites and extracts and stores the search results in an output database. Airport-pairs, departure and return dates, fare class, number of connections etc. are predefined by the user of *Avifare*. *Avifare* is able to extract automatically per day, on average, the price data for 1900 airport pairs.

The conclusion can be drawn that web mining is a promising solution in order to solve the lack of price data in aviation economics, which is in particular relevant to aviation markets outside the US. Although the results of the pilot-study are promising, further research and development of *Avifare* is clearly needed. Given the large body of knowledge with respect to web mining in Computer Sciences and the available literature on ticket price dispersion in aviation, it is expected that these knowledge gaps can be filled in.

1. Introduction

Without exception, the many studies on determinants of pricing and price dispersion of the American aviation industry are based on the DB1A database of the US Department of Transport. This database is also known as the 10% ticket sample. The DB1A database has been a major source for researchers to investigate issues such as the impact of the Airline Deregulation Act on consumer welfare, competition issues, small community air service, airfare development and hub concentration.

Aviation economists in Europe would wish they had a European 10% ticket sample at their disposal. Price data on European aviation are not freely available to researchers or are very scarce and incomplete.

Although umbrella databases on airfare data in European aviation do exist (such as the ATPCO database)¹, these databases are generally only accessible for airlines. Airlines carefully protect their price data because of competitive considerations (Swan 2002). Although some researchers managed to access airline data (see for example the research Cento (2006), supported by a major airline), in general this data handicap makes a number of empirical exercises for the European aviation market extremely difficult, in particular with respect to:

- The validation of ticket price models in European aviation
- The determination of welfare impacts of deregulation, capacity expansion or restrictions, bilateral negotiations, network scenarios and airline market entry or exit in European aviation
- The monitoring and benchmarking of ticket prices at the route, airline or network level in European aviation
- Identification of predatory pricing practices in European aviation

The price data gap has been mentioned by many researchers, in many studies and at many conferences. Is there any way out of this empirical desert? We argue there is.

This paper presents a new way of filling in the price data gap: *Avifare*. The conceptual framework, data mining methodology and price model has been developed by SEO Economic Research. *Avifare* allows researchers but also policy makers, airports and airlines to collect, monitor and use ticket price data on a continuous and flexible basis by using internet booking websites.

This paper is structured as follows. First, we give a brief overview of the existing data sources used in various studies as well as their pros and cons. Section 3 then addresses the use of internet booking websites as an alternative data source. The methodology of

¹ See for example, www.atpco.net

data extraction by using specialized software is described. In section 4 and 5 we describe some pilot results of *Avifare*. In sections 6-8 we wrap up, identify knowledge gaps and indicate avenues for further research.

2. Available airfare databases

Let us first briefly overview the available airfare databases. Which airfare databases do exist, to what extent are these databases available to researchers and what is their geographical market coverage? We conclude that price data for European aviation are not freely available to researchers or are very scarce and incomplete.

2.1 Traditional databases

DB1A The Origin and Destination Databank 1A Ticket Dollar Value O&D (DB1A) database contains all data from the continuous 10% random sample from each ticket that originates in the United States on U.S. carriers. Prices are measured as one-way fares (computed as half of the return fare). All fares other than one-way or roundtrip are excluded. The database is only accessible to U.S. citizens. Virtually all studies on pricing in the US airline industry used the DB1A database (see for example, Borenstein 1989; Evans 1993; Graham et al. 1983; Hurdle et al. 1989; Morrison & Winston 1995; Nadja 2003; Strassman 1990).

BACK Aviation O&D-lux Origin-Destination Fare Data The source data of this database is a 10% sample of airline tickets from reporting carriers collected by the DOT's Office of Airline Information of the Bureau of Transportation Statistics (DB1A) (see above). Data can be accessed through an easy-to-use platform. The database covers US domestic traffic in the US.

ATPCO Airline Tariff Publishing Company (ATPCO) collects and distributes airline fare data for more than 500 airlines. ATPCO distributes these data to the Global Distribution Systems (Amadeus, Sabre, Galileo, Worldspan) and computer reservation systems. ATPCO is owned by a number of airlines. ATPCO contains published fares. As far as our information goes, it does not include the number of passengers paying each fare. Data are only distributed among Global Distribution Systems, airlines and travel agencies. Data are not freely available for research. We do not know of academic studies using ATPCO data.

American Express European Quarterly Travel Index American Express used to publish average price indices of ticket prices in certain European markets on a quarterly basis. Since beginning of the 2006, American Express has cancelled this index.

Intervistas origin & destination data In cooperation with IATA, Intervistas offers a commercial database including passenger O&D routings and ticket prices for Canada and tickets destined to Canada.

Fare Basis Survey The Fare Basis Survey of Statistics Canada represents a regular and comprehensive source of fare type-specific data on passengers, revenues, and average air fares. The Fare Basis Survey estimates the average air fare paid and the proportion of passengers for each fare type (first class, business class, economy class, discount and other) for Canadian scheduled air carriers. The data are available by domestic and international sector, by province, and for selected cities.

The results are used by Transport Canada and the Canadian Transportation Agency for planning functions and evaluating the impact of regulatory reform and establishing policies for the exchange of air services with foreign countries. The information is also used by the System of National Accounts Branch of Statistics Canada to provide estimates of quarterly and annual business versus personal travel by province/territory and to provide interprovincial revenue estimates, by the Prices Division of Statistics Canada and by the Aviation Statistics Centre of Statistics Canada to provide a statistical service for other federal and provincial government departments, carriers, industry, consultants and the public.

In addition, some researchers have created their own databases. In his PhD-thesis, Cento (2006) investigates the relationship between airfare and determinants such as distance, GDP, Hirschman-herfindahl index and the presence of low-cost carriers. Price data were collected for 41 destinations in Germany, The Netherlands, Italy and the UK (intra-European, non-stop services), totaling more than 14.000 observations between April 2001 and July 2003. The unique dataset available to Cento was collected from Galileo by the KLM Revenue Department, where Cento was affiliated to when carrying out the research.

For their research on pricing in the European aviation market, Giaume & Guillou (2004) used a database consisting of 2592 ticket prices on 20 routes from Nice for October 15 2002. Data were extracted from the GDS Amadeus. It is not clear how the researchers were able to access Amadeus.

2.2 Web mining

With respect to ticket purchase, four sales channels are available to consumers:

- Direct off-line ticket sales: direct sale of ticket by the airline through sales office or call center.
- Indirect off-line ticket sales: indirect sale of ticket by a travel agency
- Direct on-line sales: direct ticket sale through airline booking website (e.g. www.klm.com)

- Indirect on-line sales: indirect ticket sale through intermediate internet booking website (e.g. www.travelocity.com, www.cheaptickets.com, www.orbitz.com).

Recently, researchers have discovered the value of on-line direct and indirect sales channels as an alternative airfare database, albeit not at a very large scale. Information on airline ticket prices have become freely available to consumers without interference of travel agents or airline call centers/ sales offices.

Lijesen (2002) was one of the first to use internet as an airfare data source. To answer the question how carriers price connecting flights, Lijesen manually collected ticket price data. Ticket price data collection was restricted to 29 direct and 117 indirect counterparts from seven major hubs in Europe to six primary intercontinental destinations in February 2000.

Coming from a Computer Science background focused on website extraction and search engines, Etzioni and colleagues (2003) went a few steps further than Lijesen. Etzioni et al. describe the procedure for *automatically* extracting airfare data from internet booking websites, rather than manually collecting them. The airfare database was build by AgentBuilder, a data collection agent. Furthermore, the researchers use a self-learning algorithm, called Hamlet, to predict ticket prices based on collected, historical price data. Consequently, Hamlet uses these predictions to advise passengers when to buy and when to postpone a ticket purchase. In a pilot-simulation study, they showed that Hamlet was able to save 607 simulated passengers a total of \$283.904 in a 41-day pilot run. The Hamlet fare predictor has been used to set up a fare prediction website on the World Wide Web, called Farecast². It tells passengers when to buy their ticket for a growing number of North-American cities. Eventually, the ambition of Farecast is to provide a fare predictor for the worldwide aviation network.

The paper of Etzioni and colleagues is in fact the only known research paper today, which uses automated data extraction software to mine airfare data. Most surprisingly, their work has not yet been picked up in or cited by any aviation economic study.

3. Web mining of airfares

Although web mining has become a major research issue in computer and information sciences, our review of literature and databases has shown that the potential value of internet as a source of airfare data has only been used marginally.

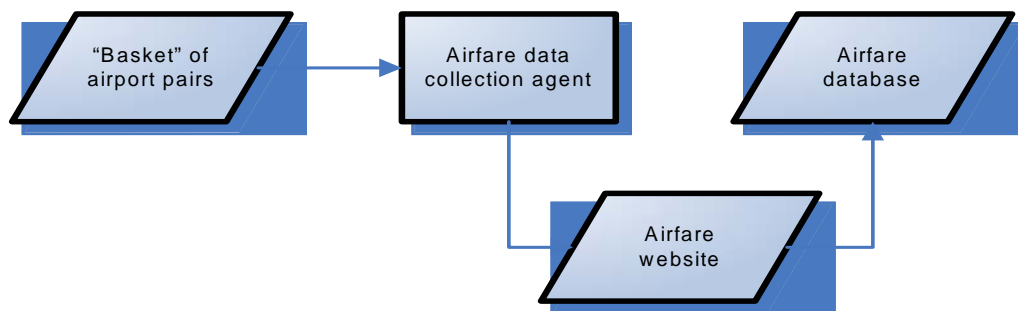
From the perspective of computer sciences, the automated use of the World Wide Web for the collection of airfare data is a form of web mining. Web mining is 'the use of data mining techniques to automatically discover and extract information from Web documents and services' (Kosala & Blockeel 2000, p. 2).

² www.farecast.com

Information Extraction and Information Retrieval are two important branches of Web Mining. Whereas Information Retrieval has the goals to select relevant documents, Information Extraction extract relevant facts from the documents. Consequently, mining airfare data from specific web sites and convert these data in a structured database is a form of Information Extraction.

We have developed an Information Extraction System, *Avifare*, to collect airfare data from a major internet booking website and convert these data into a structured database. Figure 1 summarizes the methodology of *Avifare*.

Figure 1 Structure of the Information Extraction System “ Avifare”

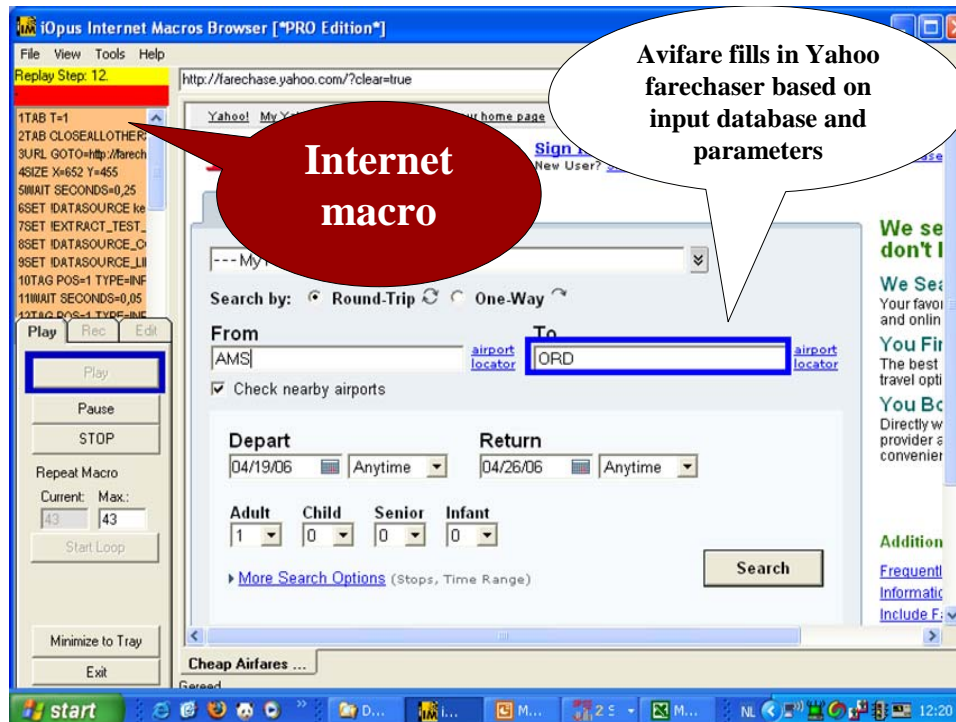


Central to *Avifare* is the data collection agent. In order to extract data on a continuous basis for a large set of airport pairs, *Avifare* uses the software package Imacros as a data collection agent. Imacros is one of the available data collection agents, but many more software packages of this kind are available.

The data collection agent automatically fills in an internet booking website. For our pilot-study, we have used Yahoo Farechase, which has resulted to be a relatively easy to ‘scrape’ site.

Then, the agent gathers the resulting airfares and other related data from internet websites and stores these results of the data mining process into a database. *Avifare* runs on a schedule, which can be defined by the user. For example, the extraction schedule can be twice-daily, daily, weekly or monthly.

Figure 2 Avifare fills in booking website



In addition, a set of parameters has to be defined by the user including:

- Origin and destination airport
- Departure and return date of the flight
- Connection type: non-stop, multi-stop, both
- If multi-stop, maximum connecting time
- Fare class: economy/lowest fare, economy unrestricted, business, first class
- Number of passengers (default=0)
- Number of airfares for each travel option that have to be stored.

With respect to the origin and destination airports, *Avifare* uses an input database. This database includes the 'basket' of airport pairs for which airfares have to be collected (AMS-JFK, AMS-LAX, AMS-SFO etc). *Avifare* automatically fills in the airfare website with these airport pairs.

The input database can be further detailed by departure and return date. This allows researchers to collect fare data for a certain basket of routes for a continuous series of departure and return dates starting today (today, today+1, today+2, today+3 etc.).

Figure 3 Example of input database

Microsoft Excel - keydestinations

Bestand Bewerken Beeld Invoegen Opmaak Extra Data Venster Help

Arial 10 B I U

A1 AMS,"ORD"

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|----|-----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AMS,"ORD" | | | | | | | | | | | | | | | | |
| 2 | AMS,"IAH" | | | | | | | | | | | | | | | | |
| 3 | AMS,"LAX" | | | | | | | | | | | | | | | | |
| 4 | AMS,"JFK" | | | | | | | | | | | | | | | | |
| 5 | AMS,"SFO" | | | | | | | | | | | | | | | | |
| 6 | AMS,"YYZ" | | | | | | | | | | | | | | | | |
| 7 | AMS,"IAD" | | | | | | | | | | | | | | | | |
| 8 | AMS,"DTW" | | | | | | | | | | | | | | | | |
| 9 | AMS,"MEX" | | | | | | | | | | | | | | | | |
| 10 | AMS,"AUA" | | | | | | | | | | | | | | | | |
| 11 | AMS,"BON" | | | | | | | | | | | | | | | | |
| 12 | AMS,"CUR" | | | | | | | | | | | | | | | | |
| 13 | AMS,"PBM" | | | | | | | | | | | | | | | | |
| 14 | AMS,"SXM" | | | | | | | | | | | | | | | | |
| 15 | AMS,"CAI" | | | | | | | | | | | | | | | | |
| 16 | AMS,"CPT" | | | | | | | | | | | | | | | | |
| 17 | AMS,"JNB" | | | | | | | | | | | | | | | | |
| 18 | AMS,"MBO" | | | | | | | | | | | | | | | | |
| 19 | AMS,"ACC" | | | | | | | | | | | | | | | | |
| 20 | AMS,"DAR" | | | | | | | | | | | | | | | | |
| 21 | AMS,"LOS" | | | | | | | | | | | | | | | | |
| 22 | AMS,"DXB" | | | | | | | | | | | | | | | | |
| 23 | AMS,"TLV" | | | | | | | | | | | | | | | | |
| 24 | AMS,"AUH" | | | | | | | | | | | | | | | | |
| 25 | AMS,"ALA" | | | | | | | | | | | | | | | | |
| 26 | AMS,"AMM" | | | | | | | | | | | | | | | | |
| 27 | AMS,"BAH" | | | | | | | | | | | | | | | | |
| 28 | AMS,"BEY" | | | | | | | | | | | | | | | | |
| 29 | AMS,"DAM" | | | | | | | | | | | | | | | | |
| 30 | AMS,"KVI" | | | | | | | | | | | | | | | | |
| 31 | AMS,"THR" | | | | | | | | | | | | | | | | |
| 32 | AMS,"BKK" | | | | | | | | | | | | | | | | |
| 33 | AMS,"PEK" | | | | | | | | | | | | | | | | |

keydestinations/ NUM

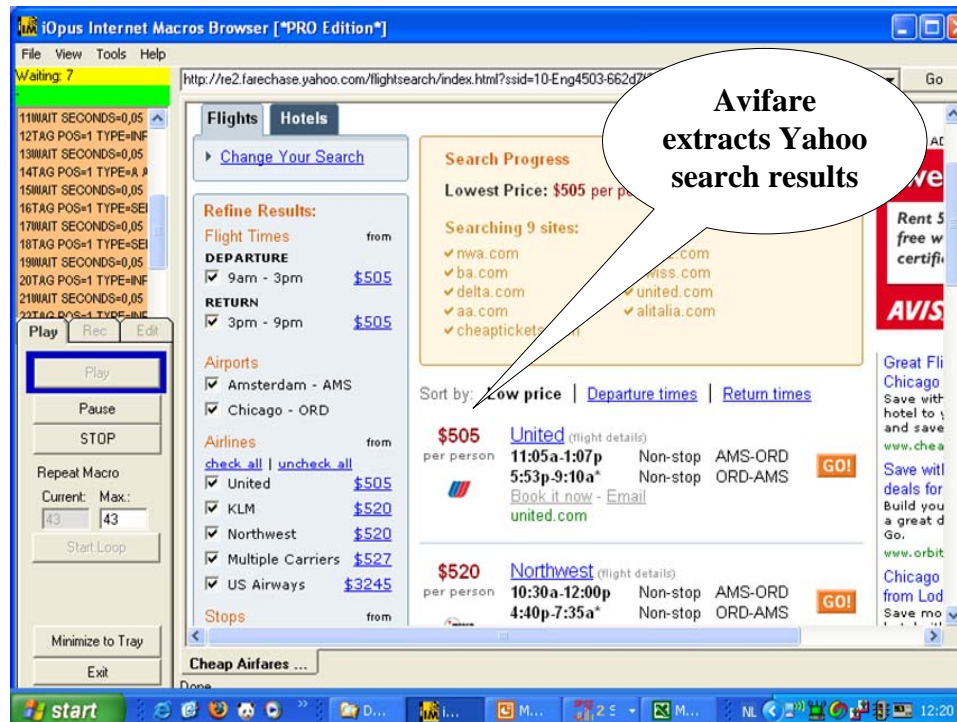
Gereed

start

12:22

Example of Input database in Avifare

Figure 4 Results of Avifare airfare query



Avifare stores the results of the search and extract process in an output database. This database includes:

- Date of website access
- Departure and return date
- Airline on departure and return legs
- Operating airline (if different from published carrier)
- Origin and destination airport
- Departure and arrival times at both the departure and return legs
- Type of flight (non-stop/multi-stop)
- Fare class (economy/lowest, economy unrestricted, business, first class)
- Airfare (\$) including taxes and fees
- Aircraft type (depends on availability on website)
- Hub airport (in case of multi-stop connection) (depends on availability on website)
- Transfer time (depends on availability on website)

4. Example of web mining results

In the pilot analysis, we have scraped prices for a number of routes between Amsterdam and North-America (New York JFK, Newark, Boston, Philadelphia, Montreal and Detroit). In this section, we present some results of the pilot analysis in order to demonstrate the potential and scope of *Avifare*.

Ticket prices can be monitored and displayed along different dimensions, depending among other things on fixation/flexibility of the booking dates and the departure/return dates (see table 1). At this moment, the monitoring system with both flexible booking and flexible departure dates is still under development. The latter type of monitoring is needed to get the full picture of the influence of the departure date and the advance booking period on ticket price levels.

Table 1 Basic dimensions of air fare monitoring and operational status in Avifare

| | Departure date | |
|--------------|----------------|------------------------------|
| Booking date | Fixed | Flexible |
| Fixed | 1) Operational | 2) Operational |
| Flexible | 3) Operational | 4) Not yet fully operational |

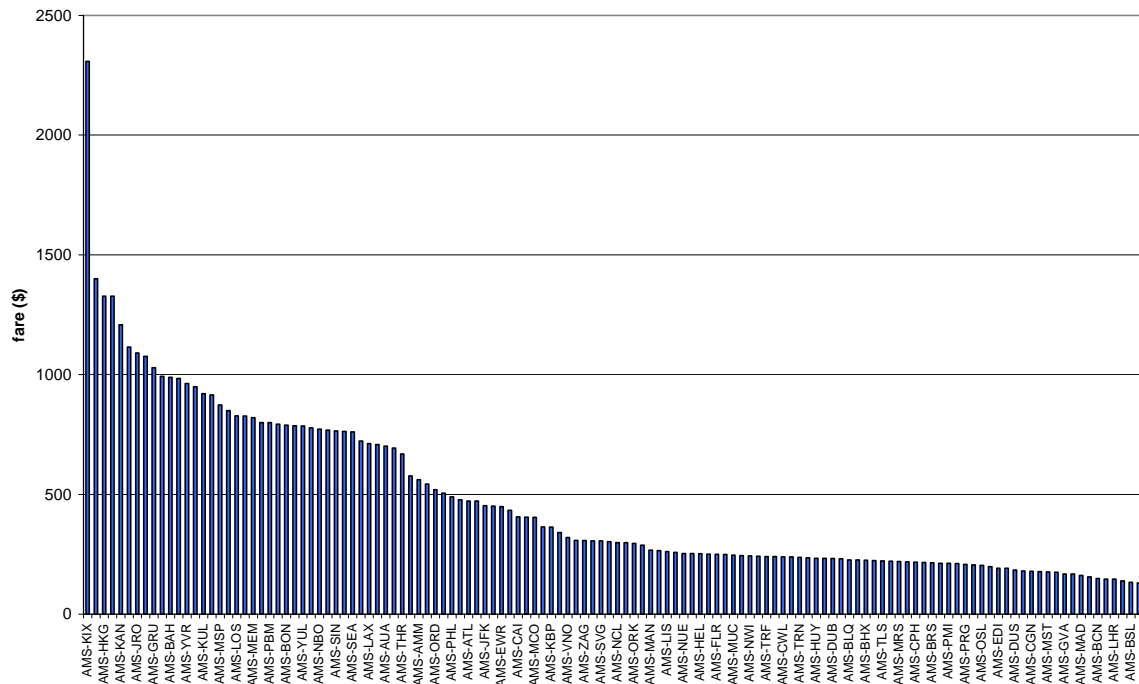
Let us illustrate the various monitoring methods with the results obtained.

1) Fixed booking date and fixed departure date

The most straightforward way of monitoring ticket prices is to scrape the prices for both fixed booking and departure dates. The advantage is that a large number of routes can be scraped in a relatively short period of time. The exercise can be repeated every x days. The figure below show an example for a large number of intercontinental routes from Amsterdam.

The disadvantage of such a monitoring system is that it is more sensitive for outliers and errors (such as the service to KIX in the figure below). Nevertheless, for complete airline networks or airports it is useful and not very time consuming to get an impression of fare levels. This is especially true when the day of departure is somewhat further away and outside the typical fare peaks (holidays and festivities).

Figure 5 Lowest available ticket price (economy) between Amsterdam and selected intercontinental destination. Booking date 30 March 2006, departure date 20 April 2006 (return one week later)

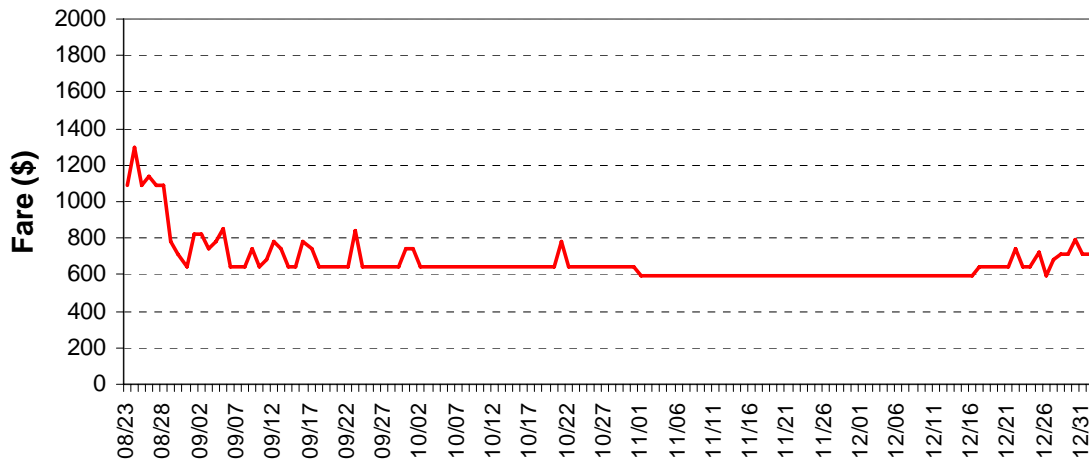


2) Fixed booking date, flexible departure date

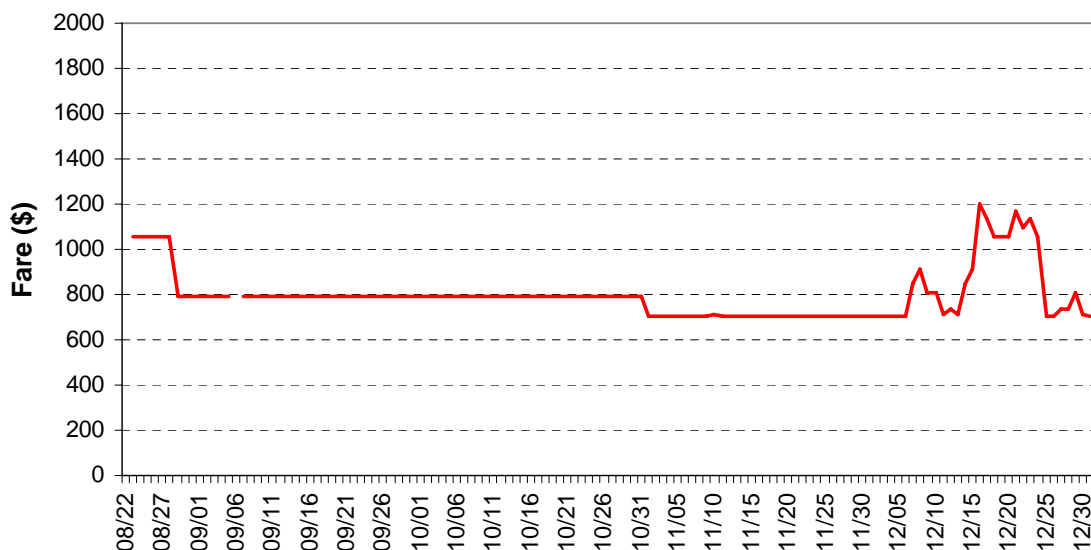
The ticket prices in figures 6, 7, 8 and 9 are fixed on one booking date but have flexible departure dates. At the specific booking date, Avifare looks ahead for a certain period (in this case about three months) and mines the lowest available economy class ticket price for each departure day (return one week later) within this time window of three months. Of course, the monitoring process can be further complicated by choosing various return dates, fare classes (economy, business) and number of connections (non-stop, multi-stop).

Figure 6a&b Lowest available ticket price (economy) between Amsterdam and Washington Dulles (above) and Amsterdam and Detroit (below) per departure date (return one week later)

AMS-IAD (booking date 23 Aug 06)



AMS-DTW (booking date 22 Aug 06)



The airfare development over the various departures dates shows intuitively sound results. Ticket prices are typically high within two weeks before departure³. This is also mirrored in the Avifare results for the various citypairs: at all citypairs spikes can be seen just after the booking date. In particular for New York JFK, fares skyrocketed just after the booking date.

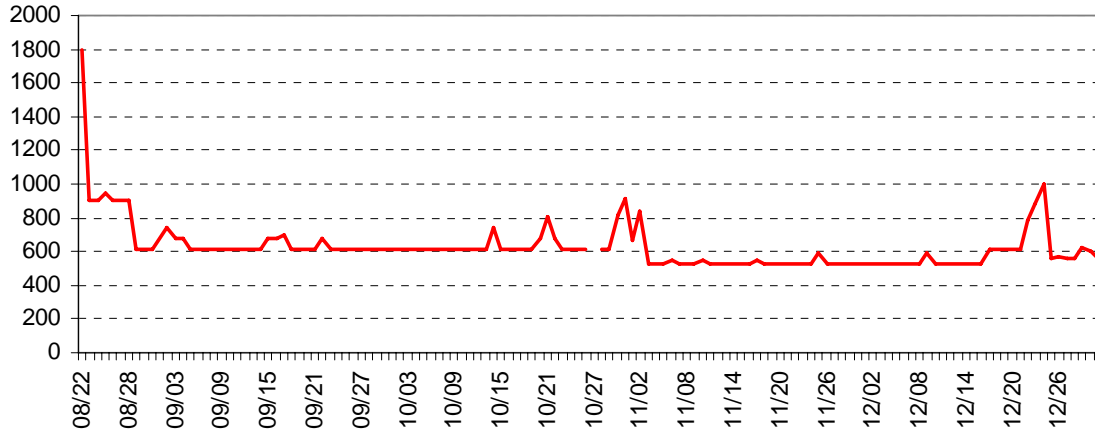
Other typical features of fare development mirrored in our results are the difference between summer and winter fares (dropping around 1 November) and the sharp peaks around Christmas Holidays. Furthermore, the fare graph of Montreal seems to reflect a peaked business travel pattern during the week.

³ See e.g. <http://farecastblog.com/blog/>

A striking feature of the results in figures 6-8 is the difference in volatility in fares. Whereas Amsterdam-Philadelphia shows a quite stable fare development, the Boston and Montreal markets are much more volatile (figures 8b and 9).

Figure 7a&b Lowest available ticket price (economy) between Amsterdam and New York (JFK above, Newark below) per departure date (return one week later)

AMS-JFK (booking date 22 Aug 06)



AMS-EWR (booking date 23 Aug 06)

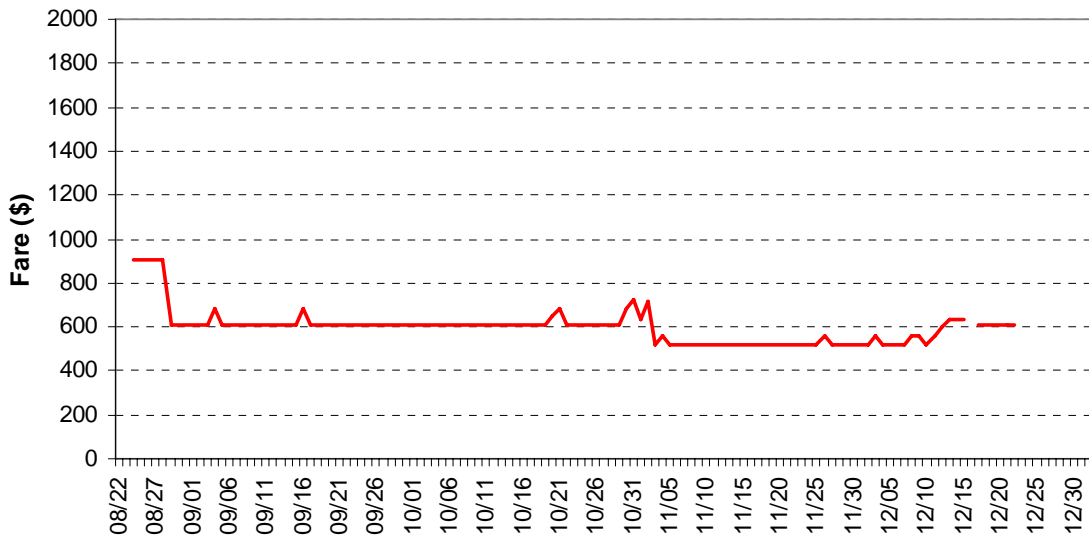


Figure 8a&b Lowest available ticket price (economy) between Amsterdam and Philadelphia (above) and Amsterdam and Boston (below) per departure date (return one week later)

AMS-PHL (booking date 25 Aug 06)

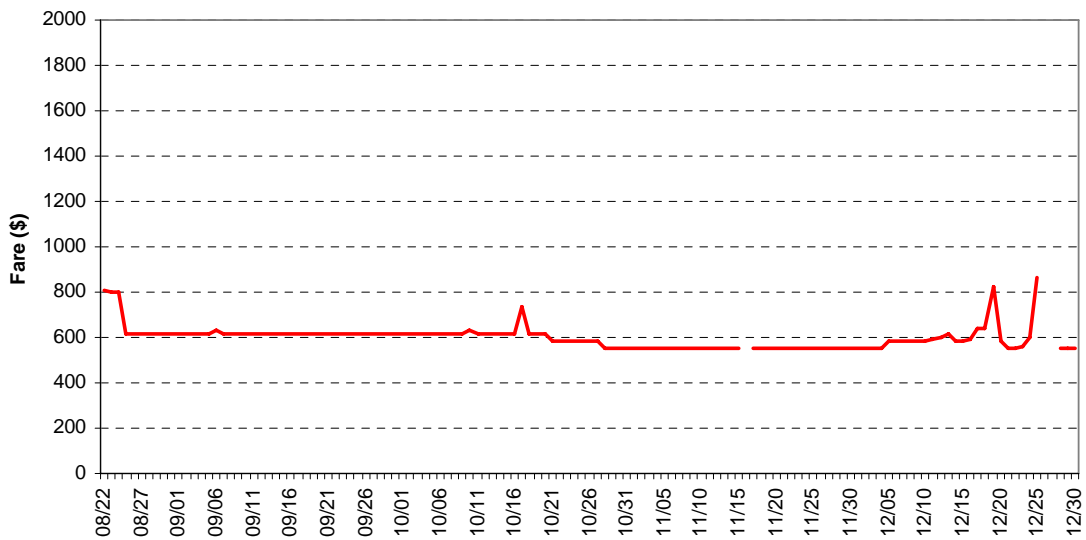
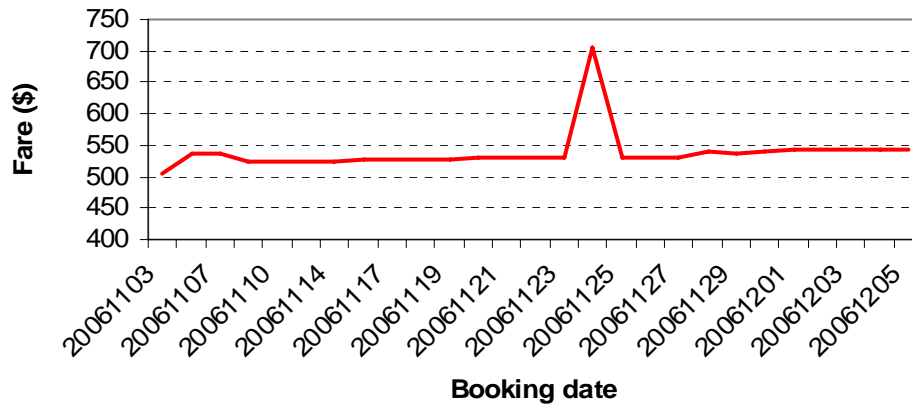


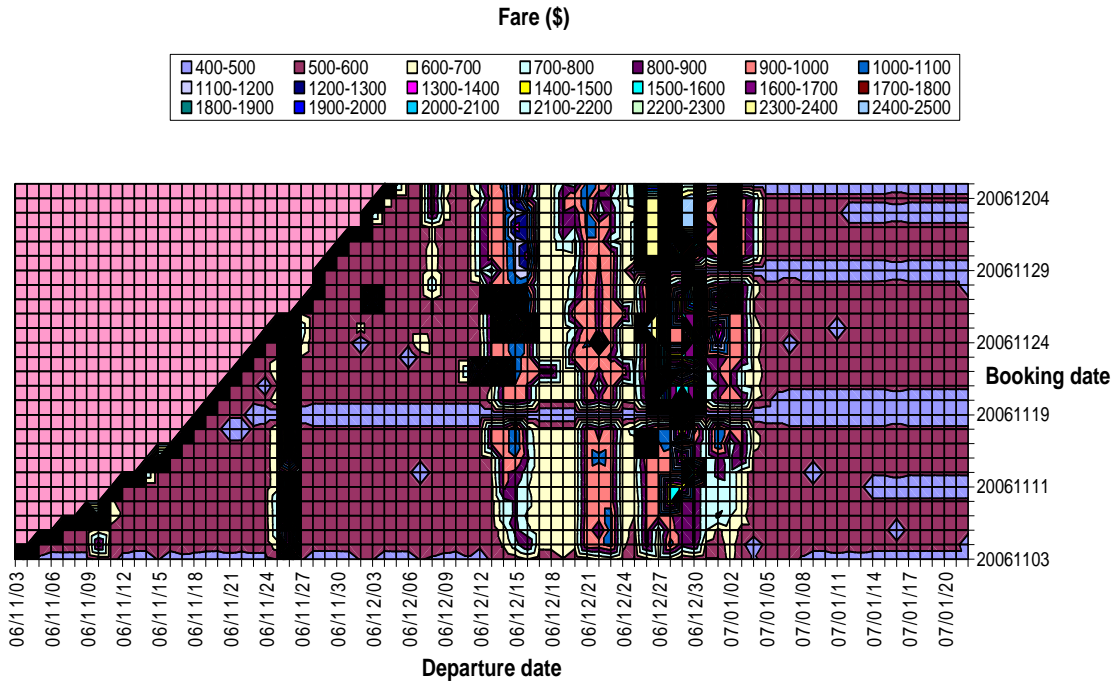
Figure 10 Lowest available fare (economy) between Amsterdam and New York JFK per booking date, departure date 6 December 2006 (return one week later)



4) Flexible booking date, flexible departure date

The above monitoring methodologies can also be combined. This yields a three-dimensional monitoring result in which fares vary both for the departure and booking date as figure 9 depicts. The denser the color, the higher the ticket price per combination of departure and booking date. Such figures can be made for extended booking periods.

Figure 11 Lowest available ticket price (economy) between Amsterdam and New York JFK per departure date and booking date (return one week later)



5. Current research project

As a test of the usefulness of *Avifare* and as a part of a Master's thesis project, *Avifare* results are currently used as a data source for an analysis of pricing at routes originating in Europe.

The aim of the research project is the validation of a ticket price model. The model estimates average fares, based on route characteristics. These characteristics are: non-stop or indirect flight, competition on route-level (measured by Herfindahl index) and distance. To verify the correctness of the current parameters, we need fare data to test the model.

First we selected a route sample, consisting of 350 routes originating from Europe. Since the model estimates average fares, we want a set with average (economy) fares for these routes. Creating this dataset is not an easy task. As previous *Avifare* results have shown, a large variation in fares exists. For each route, we might need to examine every departure date and every booking date to get all the different fares. Even then, we do not know the number of passengers paying a certain fare. Therefore, we have to make some simplifications.

Previous Avifare results show high prices within a few weeks before departure. Furthermore, we see a difference between summer and winter fares, and typical fare peaks around holidays. This is why we choose to monitor fares with *departure dates* in November only. For each route we let Avifare search for the lowest possible fare on 4 consecutive Wednesdays in November (return one week later). This procedure is followed for two consecutive days (*booking dates*), because of the limited time for this particular research project. Finally, the average fare is calculated. This way the average fare is underestimated for a number of reasons:

- We monitor fares with departure date in November only, which is the start of the winter season;
- We monitor fares with departure date on Wednesdays. Fares with departures around weekends tend to be higher;
- For each route we take the lowest possible fare;
- Booking dates are in June (4 months before departure).

Although the average fare will be underestimated, with this dataset the influence of the relevant variables in our model can be made clear.

During the monitoring process we faced some difficulties. We tried to scrape almost 20 internet booking websites: some did not represent the lowest available fares, others seemed to block the computer's IP address after some days of scraping. The best results were found on farechase.yahoo.com. Even this website changed its lay-out sometimes, forcing us to do some re-programming of Avifare.

Nevertheless, average fares could be calculated for almost all our routes. We present some first results here:

Figure 12 Fare distribution (non-stop flights)

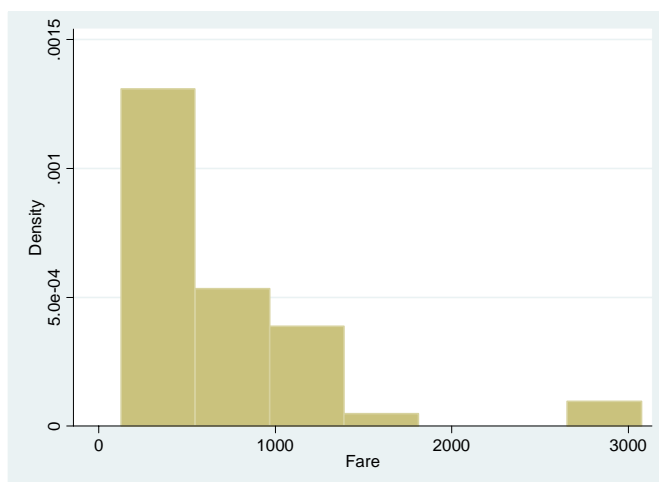


Figure 13 Fare distribution, flight distance 0 – 2000 km (non-stop flights)

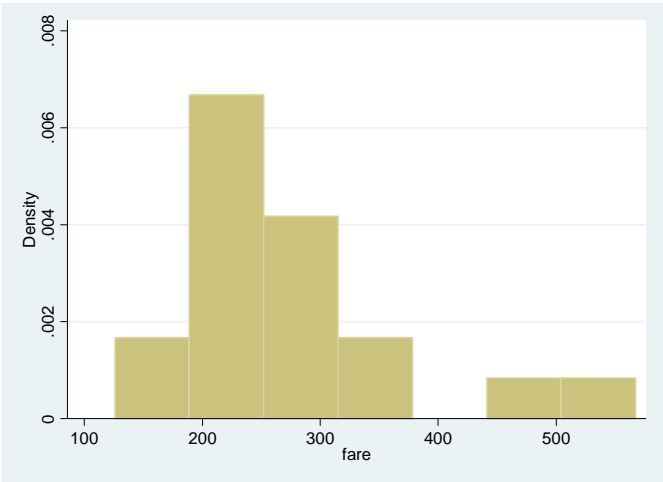
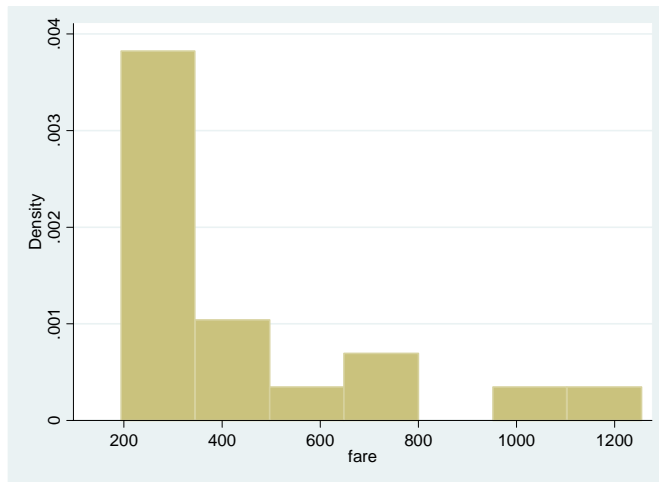


Figure 14 Fare distribution, Herfindahl index ≥ 0.5 (non-stop flights)



These histograms clearly indicate the influence of our variables: Shorter flight distance and higher competition on route-level lead to lower fares. Further research will determine the impact of each of the variables on the ticket price.

6. Knowledge gaps and questions

A number of pilot-analyses have been done with Avifare. At this moment, the following issues have been identified, which have to be addressed in further research. These issues can be categorized as technical and methodological.

6.1 Methodological issues

First of all, in contrast to the DB1A database discussed in section 2, web mining does not give any insight into the number of passengers paying a certain fare. Even if all available fares for a single flight have been extracted by *Avifare* (which is not likely given the rapid price changes on many flights) and are known to the researcher, the average of these fares is not likely to mirror the real average fare paid by the passengers. This would only be true if each fare was paid by the same amount of passengers. However, it is more likely that certain fares are paid more frequently than others.

Therefore, an important question is if it is possible to assume a certain distribution of airfares and number of passengers paying these airfares on a single flight or route. The literature on price dispersion and discrimination in the airline industry might give some useful insights to this respect (Borenstein & Rose 1994; Dana 1998; Giaume & Guillou 2004; Hayes 1998; Swan 2002) as well as interviews with airline revenue managers. Another option would be to compare the distribution of Avifare airfares with known city-pair averages of domestic US routes provided by the US Department of Transport⁴.

⁴ <http://ostpxweb.dot.gov/aviation/index.html>

Since we know the real average fare paid on certain routes, it is likely that 'rules of the thumb' can be developed in order to derivate the average from a set of airfares collected by Avifare from the WWW for a single flight or route.

Second, and related to the first issue, is the question how many fares for each query should be extracted and stored in a database. For example, a price query on an internet booking website for the market Amsterdam-New York JFK, yields a large number of ticket prices and travel options, even when constrained to a certain travel class. Since we do not know the actual fare paid by each passenger, the question is how many and which fares should be mined and stored by Avifare. Is the extraction of the lowest available fare for each fare class and airline flight a representative choice? Or should the x-th number of lowest fares be extracted? In this respect, interviews with revenue managers of large airlines may shed some light on this issue.

Third, an important issue is the monitoring methodology. As concluded earlier, different monitoring methodologies are possible:

- Airfares can be monitored with a fixed departure and return data and with a progressing booking date. For example, *Avifare* may monitor each day the available business fares for a flight departing on 8 November 2006.
- Airfares can be monitored with a fixed booking period and a progressing booking date. For example, *Avifare* may monitor daily the business fares for a flight departing three months from now.
- Airfares can be monitored with a fixed booking date for a continuous series of departure and return dates. For example, *Avifare* may monitor the fares on 23 May 2007 for same-day departure (return one week later), 24 May, 25 May, 26 May etc.
- Also combinations of these methodologies are possible.

6.2 Technical

The first and most important problem we encountered using the *Avifare* system is the fact that data collection agent was banned from certain websites after a few days of scraping, based on recognition of the computer's IP address. Hence, we are at this moment investigating what the possibilities are to use dynamic IP addresses.

Another major technical drawback of *Avifare* in its current pilot-state is certainly the possible bias in the internet booking websites, which have been used. At the moment, it is not clear to what extent the major internet booking website, which has been used in the pilot study reported in section 5, covers the real pricing behavior of airlines. It has been noticed that the specific booking website does not cover all low cost carrier services. Moreover, the booking website seems to be more accurate in the United States

and Europe than the Middle East and Africa. The use of different internet booking websites as a data source for web mining simultaneously may be one of the solutions to cope with the potential bias.

Related to this issue is the fact that *Avifare* is currently working with a single internet booking website, which is an indirect on-line sales channel. In other words, this channel collects airfare data from various airline booking websites, GDS and other indirect on-line sales channels. Would the use of direct on-line sales channels only (airline websites), instead of indirect on-line sales channels, deliver more accurate price data? Do direct and indirect sales channels deliver the same results?

Furthermore, at the moment, the data collection agent of *Avifare* requires the user to label manually the data that have to be extracted and stored. Although this system seems to be quite robust of small changes in the structure of the booking side, larger changes may result in errors in the output database. Furthermore, for large amounts of data, the labeling task becomes time intensive. Solutions may be found in Computer Sciences, where self-learning algorithms (see for example, Lerman et al. 2001) can completely automatically extract data from lists and tables without labeling them first.

7. Future research

Although the results of the pilot-study are promising, further research and development of *Avifare* is clearly needed. Given the large body of knowledge with respect to web mining in Computer Sciences and the available literature on ticket price dispersion in aviation, it is expected that these questions can be answered.

In short, research efforts should be directed at:

- Computation of an average airfare without knowing pax numbers. Research efforts should focus on:
 - reviewing insights from the price dispersion literature in aviation economics
 - interviews with airline revenue managers
 - comparing *Avifare* price averages for selected US routes and known US DoT averages
- Assessment of possible bias of indirect on-line sales channel and comparison with direct on-line sales channels
- Further development of the data collection agent of *Avifare*, possibly including self-learning algorithms.
- Mapping of the pro's and con's of different on-line sales channels as an airfare data source.

8. Conclusions

The conclusion can be drawn that web mining is a promising solution in order to solve the lack of price data availability in aviation economics, in particular with respect to

aviation markets outside the US. In computer sciences, a field of research is dedicated to automatically extract, select, store and exploit the information available on the WWW using specialized software and self-learning algorithms (web mining). Insights from web mining science are well suited to be used for airfare data collection in aviation economic research. Currently, a research project is undertaken to actually use the data which have been collected with Avifare for a analysis of ticket price determinants in Europe.

The pilot studies show promising results on the data which are delivered by Avifare. Yet, further technical streamlining is needed and methodological steps will have to be taken in order use web mined airfare responsibly on a large-scale and continuous basis. Then, we are a few steps closer to solving the lack of ticket price data in (European) aviation economics.

References

Web mining

- Etzioni, O., R. Tuchinda, C.A. Knoblock and A. Yates (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. SIGKDD '03.
- Knoblock, C.A. (2004). Building software agents for planning, monitoring, and optimizing travel.
- Kosala, R. and H. Blockeel (2000). Web mining research: a survey. ACM SIGKDD Explorations Newsletter, 2000.
- Lerman, K., C.A. Knoblock and S. Minton (2001). Automatic data extraction from lists and tables in web sources. Proceedings of the IJCAI 2001 Workshop on Adaptive Text Extraction and Mining, Seattle, WA, 2001.

Pricing

- Abramovitz, A.D. & S. Brown (1993). Market share and price determination in the contemporary airline industry. In: Review of Industrial Organization 8, pp. 419-433.
- Borenstein, S. (1989). Hubs and high fares: dominance and market power in the US airline industry. In: Rand Journal of Economics, vol 20, no. 3, pp. 344-364.
- Borenstein, S. (1992). The evolution of U.S. airline competition. In: Journal of Economic Perspectives, vol. 6, no. 2, pp. 45-73.
- Borenstein, S. & N. Rose (1994). Competition and price dispersion in the U.S. airline industry. In: Journal of Political Economy, vol. 102, no. 4, pp. 653-683.
- Carlsson, F. (2004). Prices and departures in European domestic aviation markets. In: Review of Industrial Economics, vol 24, pp. 37-49.
- Giaume, S. & S. Guillou (2004). Price discrimination and concentration in European airline markets. In: Journal of Air Transport Management, vol. 10, pp. 305-310.
- Hayes, K.J. & L.B. Ross (1998). Is airline price dispersion the result of careful planning or competitive forces. In: Review of Industrial Organization, vol. 13, pp. 523-541.
- Lijesen, M. G. (2004). Home carrier advantages in the airline industry. PhD Thesis, Free University of Amsterdam.
- Liu, Q. & K. Serfes (2005). The effect of market structure on price dispersion: an analysis of the U.S. airline industry. Paper. Department of Economics, University of Oklahoma.
- Marin, P. (1995). Competition in European Aviation: pricing policy and market structure. In: Journal of Industrial Economics, vol. 43, no. 2, pp. 141-159.
- Morrison, S.A. & C. Winston (1990). The dynamics of airline pricing and competition. In: The American Economic Review, vol. 80, no. 2, pp. 389-393.
- Morrison, S.A. & C. Winston (1990). The evolution of the airline industry. Washington: The Brookings Institution.
- Najda, C. (2003). Low-cost carriers and low fares: competition and concentration in the U.S. Airline Industry. Department of Economics, Stanford University.

Nichols (undated). Concentration and airline ticket prices: how low cost carriers changed things.

Stavins, J. (2001). Price discrimination in the airline market: the effect of market concentration. In: *The Review of Economics and Statistics*, pp. 200-203.

Swan, W.M. (2002). Prices, fares and yields. Paper presented at ATRS 2002 Conference, Toulouse.