

# Logical analysis of data for estimating passenger show rates in the airline industry

Christine Dupuis<sup>1</sup>, Michel Gamache<sup>1</sup>, Jean-François Pagé<sup>2</sup>

<sup>1</sup> Département de mathématiques et de génie industriel, École Polytechnique de Montréal, Québec, Canada, H3C 3A7

<sup>2</sup> Operations Research, Air Canada inc, Montreal, Québec, Canada, H4Y 1H4

## Abstract

A frequent practice in the airline industry is to overbook flights to make up for losses caused by absent passengers (also called no-shows). This paper proposes an approach for improving the accuracy of show rate predictions, in order to consequently adjust the overbooking levels. The chosen method is known as the «Logical Analysis of Data» (LAD). It differs from other conventional data mining methods by its ability to detect logical combinatory information about the observations (passengers). The LAD method, which is based on optimization techniques, detects sets of conditions (called patterns) for which all the satisfying passengers have a significantly higher or lower show rate than the studied population. The passengers are classified according to the patterns as show, no-show or unknown. When compared to Air Canada's current tool for overbooking forecasts, which is based on historical statistics, LAD model appears to be very competitive.

## Introduction

A frequent practice in the airline industry is to overbook flights to make up for losses caused by absent passengers or to collect double fares if a prepaid passenger doesn't show up. This practice, is not only used by airline companies, but also in other industries, such as hotels or train transportation, where potential additional revenues can be generated when spaces are freed by absent customers.

During the peak years at Air Canada, over 400 millions of dollars were generated by this practice, which can clearly make the difference between a profitable and a non-profitable year. The optimal level for overbooking is when no seats are empty on the flight at departure time, and when no passengers are denied boarding. If the show rate is correctly predicted, it becomes easier to overbook the flight accurately.

The goal of this study is to correctly classify booked passengers as shows or no-shows in order to improve the forecast overbooking level. Many methods exist for classification in conventional data mining, mostly based on statistics: clustering, decision trees, association rules. The Logical Analysis of Data, also referred to as LAD, suggests a new way of analyzing data through combinatorial logic, Boolean functions, and operations research techniques.

A dataset is a set of observations. Each observation represents a passenger and is characterized by a vector of attributes. The LAD searches through the dataset to identify recurrent sets of conditions on attributes, also called patterns, so that the passengers satisfying a pattern have a significantly higher (or lower) show rate than the rest of the population. Then, it is possible to classify the new passengers between two types, positive (show) or negative (no-show), accordingly to the patterns they mostly satisfy. Each type has its own series of patterns. Four cases can be observed. In the first two cases, the new observation only fulfils some of the positive (negative) patterns and is classified positive (negative). In the third case, the new observation satisfies patterns of both types and is classified either as positive or negative accordingly to the weight given to specific patterns of both types covering the observation. Finally, in the fourth case the observation fulfils no patterns at all, and therefore remains unclassified.

In the first section of this paper, we discuss how the airline industry conducts overbooking. We also present the reasons that led the choice of the LAD method for the classification of passengers. In the second section we describe the basic principles of LAD and how it classifies data, while in the third section we present how we adapted the

method to the show rate problem. In the fourth section, tests on real life data confirm that this method is very accurate for predicting show rates. Finally, in conclusion, we propose new avenues of research that could be explored.

### **1. Literature review**

Overbooking practices in the airline industry vary largely from one airline to another. When dealing with multi booking class fares and seat allocation in nesting, like it is done at Air Canada, the authorized capacity can be calculated as follows:

$$\text{Authorized Capacity} = \frac{\text{Actual Capacity}}{1 - \text{Show Rate}}$$

In this formula, the capacity of the plane is taken into account, and it is considered that the cost of a denied boarding is the same as an empty seat, which is rarely the case. In addition, the main input is the show rate for the flight, which is usually a simple average on historical flights (Hillier et al., 1998). No matter which calculation method is chosen, the show rate remains the most important input for the inventory management.

It is also important to consider the risk the company is willing to encounter. For example, some airline companies have a zero overbooking policy in order to guarantee that a booked passenger will never be denied boarding. This is a strategic marketing decision that allows those airlines to publicize around it, and to gain favourable image in the customer's eye. Usually these companies require payment at the time of the reservation.

However, the risk that a passenger will be bumped from a flight is relatively small. According to the U.S Department of Transportation's Office of Aviation Enforcement and Proceedings, in 2008, 63,612 passengers were kept from overbooked flights. This number represents a rate of 1.10 bumps per 10,000 passengers.

Suzuki (2002) has studied the optimal overbooking policies for US major airlines by considering how denied boarding passengers would behave after they are bumped. The results imply that overbooking improves an airline's current revenue, but it also reduces

the airline's future revenues. The results also imply that, although there is a significant negative overbooking effect, no airline should decrease overbooking levels because the positive side of overbooking is so strong that it more than offsets its negative side.

Considering that overbooking is profitable, companies like Air Canada develop research efforts to improve show rates forecasts, in order to adjust the overbooking levels optimally. Currently, most of the models for sales forecast in the airline industry are based on the historical rates of attendance on each flight (sometimes down to seat-class level). These approaches ignored the characteristics of the individuals who bought the seats. In this paper, we investigate these characteristics in order to better predict the show rate.

According to Joe Brancatelli (see Bertoni 2009), editor of the business travel site Joesentme.com, most of the no-shows are business customers while leisure fliers tend to book trips early and show up when they say they will. We are not so categorical in our claims, because several other factors may contribute to the presence or absence of a passenger at the gate. More recently, companies give more importance to studying the various elements that characterize the passengers and flights: the type of ticket, seat type (class), the day of the week, the departure time, the origin, the destination, to name a few. The majority of this information comes from a database called the PNR (Passenger Name Record).

In 2003, Lawrence, R. D., Hong, S. J. & Cherrier, J., respectively from IBM and Air Canada, first developed a passenger-based modelling for estimating no-show rates. Their approach, consisting in adjusting the prediction based on linear regressions of attributes in the PNR, was qualified as efficient by the authors. They pointed out that the most relevant attributes were the presence of a ticket number in the PNR, a frequent flyer program member, whether the flight was the first leg for the passenger or no, the origin of the PNR and the booking class, which we also included in our significant attributes list. The authors used a set of passenger features and the cabin level no-show rate predicted

by the historical model as inputs to predict the no-show probability for the whole cabin. The average features were calculated for the studied group, and used as influences to the historical no-show rate of this group.

Gorin et al. (2006) studied the benefits of adjusting the costs according to three characteristics extracted from the PNR: ticket type (electronic or no), passenger on return trip or outbound, and the number of people travelling together. The adjustment of their prediction model was based on the proportion of passengers on the flight that are on their return segment, the proportion of passengers with electronic tickets and the allocation of passengers between those that are local, compared to those arriving via another flight (connecting).

Thus the idea of using the available information about the passenger to adjust the overbooking is not new. Given the number of observations (passengers) and a binary outcome, i.e. the passenger is present or not. This seemed to be a good fit for the logical analysis of data (LAD) as a clustering method. Up to this day, LAD shows excellent results in diversified fields for similar problems. The following table, from Boros et al. (2000), shows the results for a set of solved problems by different data mining methods. These datasets were drawn from the *Repository of Machine Learning Data-bases and Domain Theories*, maintained by University of California, Irvine.

Table 1: LAD comparison with other methods

| Dataset                       | LAD     |          |         |          | Best found |     |       |
|-------------------------------|---------|----------|---------|----------|------------|-----|-------|
|                               | Average | Std Dev. | Average | Std Dev. |            |     |       |
| <i>Australian credit card</i> | 85,4    | 1,2      | 85,5    | 2,6      | 85,5       | N/D | 71%   |
| <i>Boston housing</i>         | 84,0    | 1,6      | 85,2    | 3,0      | 83,2       | 3,1 | 80%   |
| <i>Breast cancer</i>          | 96,9    | 0,9      | 97,2    | 1,3      | 96,2       | 0,3 | 80%   |
| <i>Congressional voting</i>   | 96,2    | 1,1      | 96,6    | 1,8      | 95,6       | N/D | 66,6% |
| <i>Diabetes</i>               | 71,9    | 1,9      | 72,3    | 2,4      | 76         | N/D | 75%   |
| <i>Heart disease</i>          | 82,3    | 1,7      | 83,8    | 5,2      | 80,6       | 3,1 | N/D   |

Two different series of 20 tests were performed, the first using half of the dataset for training, and the other half for testing, while the second uses 80% for training, and only 20% for testing. The average accuracy for the classification is higher when using a larger part of the dataset for building the patterns, however, more variability is observed from one test to another (higher standard deviations). The right part of the table shows the results obtained by the best known approach for the same problem.

LAD appears to be very competitive with each of the six conventional data mining methods appearing in the right part of the table. It is a robust method, very flexible and adaptable, and it offers great stability, which makes it an excellent tool for most classification problems. For more details on each of these LAD models and their results, the reader can refer to Boros et al. (2000).

Shortly after the new method was proven efficient, in the 90s, the first LAD three pilot studies were launched. They are reported in depth in Boros et al. (2000) and in Hammer (1999): labour productivity in Chinese provinces, choice of grounds for oil exploitation, and psychometric tests for depression diagnosis. All these three studies show unquestionably successful results not only for the classification accuracy, but also for the use of the rich information contained in the patterns. The patterns can indeed be very useful for other purposes than the classification alone. They allow to explain the presence or absence of the studied phenomena by looking at the conditions, and to show which attributes are the most significant.

The LAD method has subsequently been used for diverse applications. To date, the method has been proved effective in different areas: identification of the risks of coronary artery blockage (Alexe et al., 2003), growth of organic polymeric material (Abramson et al. 2005), diagnostic of lung problems (Boros et al., 2000), identification of lymphoma diffuse large B cell (Alexe et al., 2005), control of defective parts on aircraft flights during maintenance (Bennani, A. & Yacout, S., 2009).

This method has proven to be robust and adaptable to essentially any type of classification problem. Given its success in various fields, and the fact that it has not yet been employed to solve problem types where the random factor is major, such as human behaviour phenomena, the LAD method is selected to study the passenger show rates issue in the airline industry.

## **2. Logical analysis of data basic principles**

The LAD general idea is to predict the happening of a given event, in our case the presence or absence of a passenger for his flight, using sets of conditions that describe this event as very likely to happen (positive) or very unlikely to happen (negative), for the subset of individuals that meet these conditions among the studied population. In order to do so, each individual, also called observation, is characterized by a vector of attributes, for which the values are specific to each observation. The sets of conditions on these attributes, also referred to as patterns, are built by putting together conditions (less than, greater than) on cutpoint values. A pattern consists of a set of terms, each composed of an attribute, a cutpoint value and a sign for the restriction.

The LAD can be broken down into four phases: 1) data extraction and preparation, 2) discretization, 3) pattern generation, 4) model construction and validation. Each phase has its parameters and number of possible ways it can be accomplished. To build and evaluate the quality of a model, the dataset is initially separated in two subsets: training and testing. While the first subset is used for the pattern detection, the second is used for the validation of the model. The observations are compared to each pattern, and their score is calculated as described in the fourth phase.

The first phase is to collect the needed data. The data consists of a vector of attributes for each passenger. Examples of attributes are booking class, origin, destination, time of departure, etc. The higher the correlation between the attributes and the outcome of the event, the better they are for use in LAD. It is the combination of two or more of these attributes that can help predict the outcome (positive or negative) for each observation.

LAD requires the data to be numerical. Therefore, when dealing with descriptive or nominative fields, for example city names, each different term must be given a numerical value. In order for the indexation of categorical attributes to be coherent, it must be made consequently with the natural order of the different values. For example, the days of the week can be numbered from 1 to 7, starting on Monday. The LAD method requires doing so, because of the discretization in the second phase. The more this numbering is coherent, the better the results will be. As for a second example, the city names are numbered geographically from West to East. At the end of the first phase, all the pertinent attributes should be extracted and the categorical attributes should be numerically ordered.

The second phase of LAD consists in creating a separation grid for the observations based on a cutpoint system. This is needed for the indexation of the values inside each attribute. This practice is mainly useful to reduce the number of different values that can be checked against for each attribute. Implicitly, the computational time for the pattern generation grows exponentially with the total number of different values through all the attributes, so it is convenient to keep this number as low as possible. This is because the more possible values there are for each attribute, the more combinations are to be examined when searching for patterns. However, the objective remains to the observations in types (presents and absents) so it is necessary to insert a cutpoint as frequently as a significant change is observed in the passenger behaviour. For example, since no change is observed in the show rates from Monday to Wednesday, there is no need for a separation, but on Thursdays, some important changes occur, therefore this will be the first cutpoint, and then the values are consequently readjusted accordingly to the indexation:

Table 2: Days of the week discretization

|                            |   |
|----------------------------|---|
| Monday, Tuesday, Wednesday | 1 |
| Thursday                   | 2 |
| Friday                     | 3 |

|          |   |
|----------|---|
| Saturday | 4 |
| Sunday   | 5 |

This allows us to compute only five different values for the days of the week attribute, instead of the original seven. When choosing the number and the location of the cutpoints for each attribute, different methods can be used such as clustering or statistical analysis tools. This is done for each attribute, and once all the attributes are indexed according to the grid, it is possible to move on to the next step: the pattern generation, also called pandect generation.

Three parameters need to be introduced in order to guide the pattern generation: homogeneity, prevalence and degree. The homogeneity is the proportion of positive observations covered by one pattern, out of the total number of observations covered by this same pattern. For positive patterns, homogeneity should be as high as possible, while for negative patterns it should be low. The prevalence of a pattern is the indicator of its importance: it should be as high as possible for both types of patterns. It is measured by the number of observations covered by the pattern, over the total number of observations in the dataset, for this same type. Finally, the degree is the maximal number of bounds on attributes permitted in one pattern. If the degree is equal to 3, a pattern can either contain bounds on three different attributes, or one condition on an attribute and two conditions on another one (interval). Because the pattern generation algorithm that we used visits all possibilities, each cutpoint is tested as a potential term.

Many algorithms exist to conduct the third phase with success: bottom-up, top-down, consensus (Boros et al., 2000). However, the LAD V2.0 software from Alexe Sorin was used for experimental purposes. The programmed algorithm consists in visiting every single possibility. It is obvious that this is not optimal in terms of computational delays, but it is surely the simplest way to program it.

Once all the patterns satisfying the fixed parameters are listed, we can move on to the fourth and last phase: to select which patterns will constitute the optimal model. To diminish the number of patterns, only some of them are selected from the pandect to be part of the model. In order to do so, this can be addressed as a set covering problem, for example, with a greedy algorithm. Once the patterns are selected, the model is almost complete. The final step is to attach weights to each pattern to indicate its relative importance. The purpose of the weighting is to classify observations that may satisfy only positive patterns, only negative, but possibly both types. Those satisfying none will remain unclassified. A new observation's score is calculated by adding (subtracting) all the weights of the positive (negative) patterns that are satisfied. For each type, a threshold value must be determined. When the score of the new observation is higher (lower) than the positive (negative) threshold, it is classified as positive (negative). If it is in between the two thresholds, or if it satisfies no patterns at all, the observation remains unclassified.

### **3. Show rates with LAD**

In this section, it is shown how the LAD method can be adapted to the show rates problem in the airline industry, from the extraction of the information about passengers through each phase of LAD, until the classification of new passengers, and the prediction of show rates for future flights.

Navigating through airline databases is an uneasy task; their volume and their structure make it difficult to understand for an outsider. Nevertheless, it was possible to extract over 30 attributes for exploration. In order to perform the LAD method on these attributes, we first had to transform all the non numerical data into numerical values. Although we did try some tests on this first set of attributes, it was rapidly noticed that not all attributes appeared in the patterns. Also, the computational delays were considerable, especially when using higher degrees. This is why it was decided to filter the attributes before applying LAD. It was also resolved, for simplicity sake, to combine

the discretization phase with the filtering. The final list of attributes that were made discrete is presented in the first appendix.

Each of these attributes is collected for every PNR of a person travelling from Vancouver to Calgary, on flights commercialized by Air Canada, during March 2009. This is equivalent to a population of 38 501 passengers for the study. Half of them are put into the training set, for the pattern generation; the rest is kept aside for the evaluation of the patterns, i.e. the testing.

Now that extraction of data and discretization are completed, it is time to move on to the pandect generation. As early as the first exploration tests, the booking class appears in almost every pattern, positive as well as negative. Based on this observation, we separated the data respectively to their product, thus creating five populations, therefore the need for five different pandects. This has the great advantage of eliminating one more attribute, furthermore, five different cutpoint values, and also reducing the size of the datasets, causing the computational delays to considerably drop. Implicitly, this is also equivalent to add an extra degree to all the patterns to be generated: the booking class attribute.

Many tests were made in order to explore the different combinations of homogeneity, prevalence and degree. When the homogeneity is too high (low) for the positive (negative) type of patterns, none are generated. The best results occur when using the highest (lowest) homogeneity possible that allows to generated patterns of both types, combined with the highest prevalence possible.

As for the maximum number of degrees, it was set that it had to be at least 3, but less than 5. In fact, the use of a single degree is equivalent to making simple linear correlations between the attribute and the outcome. With 2 degrees, the level of description is still not high enough, and very few patterns actually exist. The higher degrees such as 5 and 6 are not useful for our problem, because many patterns already appear when using 3 or 4

degrees. If we allow a fifth attribute on a valid 4 degree pattern, the software can create more patterns by adding each of the remaining attributes to the existing pattern, therefore creating several patterns of 5 degrees, all containing the same first four conditions, and covering the same passengers. All five of our models use 4 degrees, which we found was optimal for our problem.

For each of the five products, LAD proposes a list of patterns among which a subset should be selected. For simplicity sake, we skipped this phase, and chose to keep all patterns for the model. For the weights attached to each pattern in the discriminant equation, it was decided to use the relative prevalence. To illustrate this concept, let the number of positive patterns be 5, and the number of negative patterns be 15. A new passenger is evaluated and satisfies one pattern of each type. Each pattern for the positive type has a weight of  $+1/5$ , while each negative pattern has a weight of  $-1/15$ . Thus the total score for this passenger's discriminant is  $+2/15$ . Because we chose the critical values to be equal to zero for both types, this passenger is classified as positive. When the score is less than 0, the passenger is classified negative. If the passenger does not satisfy any pattern, his score will be 0, and he will remain unclassified.

#### **4. Results and discussion**

After the four phases, the LAD model includes the list of patterns, their weights and the critical values for both types. We tested the model on a new set of passengers, i.e. all travellers for the same market (Vancouver-Calgary) during April 2009. For each passenger from the April 2009 dataset, we test each pattern generated with the March 2009 dataset. According to the patterns satisfied by the passenger, his score is calculated, and the passenger is classified as show ( $>0$ ), no-show ( $<0$ ) or unknown ( $=0$ ). Then, the number of shows is multiplied with the show rate of the positive group (obtained from March 2009), the number of no-shows with the negative group show rate, and the unknown with the unclassified show rate. This weighted sum of the three groups gives the total number of expected presences. If the global show rate is needed, it can easily be calculated by dividing the sum of presences by the total number of bookings.

To illustrate how the show rates are estimated with the LAD model, the following is an example with one of the products (Tango), but it is the same for all five products.

Table 3: Optimal parameters for Tango product

| <b>Parameters</b>                    | <b>Positive</b> | <b>Negative</b> |
|--------------------------------------|-----------------|-----------------|
| <b>Homogeneity</b>                   | 99%             | 15%             |
| <b>Prevalence</b>                    | 3%              | 3%              |
| <b>Number of patterns in pandect</b> | 315             | 485             |

Table 4: Results for Tango model March 2009 classification

|                          | <b>Positive group</b> | <b>Negative group</b> | <b>Unclassified</b> |
|--------------------------|-----------------------|-----------------------|---------------------|
| <b>Total homogeneity</b> | 98,16%                | 22,22%                | 93,13%              |
| <b>Total prevalence</b>  | 23,62%                | 0,61%                 | 75,78%              |

These results are obtained from the classification of all passengers for March 2009, training and testing subsets altogether, which is why the homogeneity for the positive group is lower than the required 99%. The homogeneities from this table are equivalent to the show rate of the three groups. All Tango passengers extracted from April 2009 are classified as follows:

Table 5: Detailed results from the LAD model for Tango, April 2009

| <b>Tango</b>        | <b>Count</b> | <b>Rate</b> |
|---------------------|--------------|-------------|
| <b>Presences</b>    | 3 219        | 98,16%      |
| <b>Absences</b>     | 58           | 22,22%      |
| <b>Unknown</b>      | 10 083       | 93,13%      |
| <b>Weighted sum</b> | 12 563       |             |

Table 6: LAD comparison with actual global show rate, for Tango, April 2009

| <b>Tango</b>     | <b>Actual</b> | <b>LAD</b> |
|------------------|---------------|------------|
| <b>Presences</b> | 12 575        | 12 563     |
| <b>Total</b>     | 13 360        | 13 360     |
| <b>Rate</b>      | 94,03%        | 94,12%     |

Clearly, the LAD obtains global monthly show rates that are extremely close to reality. However, this is obviously not enough to assure that this method is accurate, because it is the plane capacity that determines whether they are empty seats at departure or denied boardings. This is why it is important to reassemble the flights, i.e. to gather all passengers on one flight, on a specific day, to count for each group, and calculate the weighted sum, just as it is done in Table 5.

In Table 7, the actual presences by flights are put in comparison with the LAD results, as described above, and also with the forecasts provided by Air Canada's commercial tool. It is important to take notice that the current tool produces forecasts by class, by flight number, and by day, all based on historical statistics of the same class, same flight, same day. In a given class, on a given flight and day, the number of bookings varies between none and a dozen, making comparisons class to class rather insignificant statistically. This is why we aggregated the predictions of each class to reconstruct the flights.

Table 7: LAD comparison with actual flights show rate, and with Air Canada's forecasts

| Flight | Date | Actual | Forecasts |         | LAD errors         |                | AC errors          |                | Best     |
|--------|------|--------|-----------|---------|--------------------|----------------|--------------------|----------------|----------|
|        |      |        | LAD       | AC      | Error <sup>2</sup> | Relative Error | Error <sup>2</sup> | Relative Error | AC / LAD |
| 202    | 1    | 47     | 75.916    | 81.405  | 836.147            | 61.5%          | 1183.698           | 73.2%          | L        |
| 210    | 1    | 59     | 60.574    | 60.203  | 2.478              | 2.7%           | 1.447              | 2.0%           | A        |
| 214    | 1    | 70     | 69.234    | 76.332  | 0.587              | 1.1%           | 40.098             | 9.0%           | L        |
| 224    | 1    | 110    | 104.588   | 113.073 | 29.288             | 4.9%           | 9.445              | 2.8%           | A        |
| 202    | 5    | 85     | 79.187    | 94.704  | 33.793             | 6.8%           | 94.161             | 11.4%          | L        |
| 210    | 5    | 121    | 120.634   | 125.072 | 0.134              | 0.3%           | 16.581             | 3.4%           | L        |
| 214    | 5    | 105    | 105.551   | 118.163 | 0.303              | 0.5%           | 173.275            | 12.5%          | L        |
| 224    | 5    | 91     | 91.025    | 92.106  | 0.001              | 0.0%           | 1.223              | 1.2%           | L        |
| 202    | 10   | 107    | 107.143   | 95.287  | 0.020              | 0.1%           | 137.194            | 10.9%          | L        |
| 210    | 10   | 114    | 112.664   | 114.635 | 1.784              | 1.2%           | 0.404              | 0.6%           | A        |
| 214    | 10   | 72     | 68.096    | 63.849  | 15.238             | 5.4%           | 66.444             | 11.3%          | L        |
| 224    | 10   | 44     | 53.466    | 71.361  | 89.599             | 21.5%          | 748.616            | 62.2%          | L        |
| 202    | 13   | 129    | 123.103   | 127.760 | 34.772             | 4.6%           | 1.539              | 1.0%           | A        |
| 210    | 13   | 122    | 118.295   | 133.100 | 13.727             | 3.0%           | 123.220            | 9.1%           | L        |
| 214    | 13   | 96     | 94.180    | 87.009  | 3.314              | 1.9%           | 80.833             | 9.4%           | L        |
| 224    | 13   | 111    | 111.521   | 120.747 | 0.272              | 0.5%           | 94.996             | 8.8%           | L        |
|        |      |        | Sum       |         | 1061.457           | 54.6%          | 2773.175           | 228.8%         |          |
|        |      |        | Average   |         | 66.341             | 3.6%           | 173.323            | 14.3%          |          |
|        |      |        | Std dev.  |         | 206.624            | 5.4%           | 324.115            | 21.4%          |          |

All indicators prove the LAD to be superior to the commercial tool. The square sum of errors is almost three times smaller with LAD, so are the average and standard deviation.

Also, when looking at the absolute number of times each method is closer to reality, LAD provides a more accurate forecast 12 out of 16 times, or 75%. We did the same analysis for each cabin, and also at an even more precise level: for each product. Each

time, the results were similar and very good. These results were calculated using a month of historical statistics, because the LAD was only using a month history of passengers for the patterns.

When using a longer history for Air Canada's commercial tool, a full year of historical flights, the results were a little less impressive, but still clearly in favour of LAD. We do not show them in this paper because we do think that this comparison is not significant if we are not using also a full year of passengers for the LAD predictions.

There are still some improvements that can be brought to our application of LAD in order to obtain even better results. First, a search for more relevant attributes highly correlated to the outcome could be done. We also recommend using a clustering algorithm for the discretization phase. This would allow grouping the values that have a small difference on their show rate, at the same time as separating the ones that are more distanced.

The literature suggests many pattern generation algorithms that would make the solving more efficient, therefore permitting to test even more combinations of parameters. The solving of a set covering problem allows to reduce significantly the number of patterns listed in the pandect, and also to eliminate the redundancy phenomenon that appears when using the higher degrees such as 5 and 6. This is highly recommended, and we do expect that it would trigger some great improvements in the results.

We also suggest that the choice of weights be explored more in depth. It would be possible to have a linear program to optimize these choices. Finally, the selection of the two critical values could be improved. For simplicity sake, we used only two, and they were both equal to 0, but it is possible to have more than two thresholds for the classification, allowing grading the risk of no-show from very high to very low, in as many intervals as the user chooses critical values.

## **Conclusion**

In conclusion, the LAD method appears to be very promising as it offers some better predictions than the current tool, even when not used to its full potential. The objectives of estimating show rates accurately are achieved, and in addition the LAD offers some explanation of the no-show causes when we look into the content of the patterns and the usage of each attribute.

The main improvements that are required for better results include the choice of the attributes, the discretization, the pattern generation, the selection of a subset of patterns for the model, their respective weights, and finally the thresholds for the classification.

As the LAD is a classification method that has proven to be accurate, robust, flexible and adaptable, there are many problems that it could be used for solving. To name only a few in the airline industry, we have thought of using LAD for classifying the flights instead of the passengers, for detecting probability of pilots absence and for the cargo freight.

## References

- Abramson, S. D., Alexe, G., Hammer, P. L., Kohn, J. (2005) A computational approach to predicting cell growth on polymeric biomaterials. *Wiley InterScience*. Consulted September 25th 2008, from <http://dx.doi.org/10.1002/jbm.a.30266>
- Alexe, G., Alexe, S., Axelrod, D., Hammer, P. L., Weissman, D. (2005). Logical analysis of diffuse large B-cell lymphomas. *Artificial Intelligence in Medicine*, 34, 235-267.
- Bennane, A., Yacout, S. (2009). LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance. *J Intell Manuf.* Consulted January 7th 2010, from <http://dx.doi.org/10.1007/s10845-009-0349-8>
- Bertoni, Steven (2009), America's Most Overbooked Airlines, <http://www.forbes.com/2009/04/16/airline-tickets-flights-lifestyle-travel-airlines-overbooked.html>
- Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A. (1997). Logical analysis of numerical data. *Mathematical Programming*, 79, 163-190.
- Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE transactions on knowledge and data engineering*, 12 (2), 292-306.
- Gorin, T., Brunger, W. G., White, M. (2006). No-show forecasting: A blended cost-based PNR-adjusted approach. *Journal of Revenue and Pricing Management* 5 (3), 188-206.
- Hammer, A. B., Hammer, P. L., Muchnik, I. (1999). Logical analysis of Chinese labor productivity patterns. *Annals of Operations Research* 87, 165-176.
- Hammer, P. L., Bonates, T. O. (2006). Logical analysis of data – An overview: From combinatorial optimization to medical applications. *Annals of Operations Research* 148, 203-225.
- Hillier F. S. (1998). A tutorial on optimization in the context of perishable-asset revenue management problems for the airline industry. In Saigal, R., Nagurney, A., Zhang, D., Padberg, M., Rijal, M., Vanderbei, R., Jaiswal, N., Gal, T., Greenberg, H., Prabhu, N.,

Fang, S. C., Rajasekera, J., Tsao, H., YU, G (éd.), Operations Research in the Airline Industry (pp. 68-98). Boston: Kluwer Academic Publishers.

Lawrence, R. D., Hong, S. J., Cherrier, J. (2003). Passenger-based predictive modeling of airline no-show rates. *SIGKDD* 3, 24-27.

Suzuki, Y. (2002). An empirical analysis of the optimal overbooking policies for US major airlines, *Transportation Research*, Part E, 38, pp 135-149.

Appendix 1: Withheld discretized attributes, values and descriptions

| <b>Attributes</b>      | <b>Values</b> | <b>Descriptions</b>                       |
|------------------------|---------------|---|
| Presence of ticket     | 0             | No ticket number in PNR                   |
|                        | 1             | Presence of ticket number in PNR          |
| Day of the week        | 1             | Sunday                                    |
|                        | 2             | Monday, Tuesday and Wednesday             |
|                        | 3             | Thursday                                  |
|                        | 4             | Friday                                    |
|                        | 5             | Saturday                                  |
| Segment number         | 1             | First segment in full itinerary           |
|                        | 2             | Second segment                            |
|                        | 3             | Third, fourth or fifth segment            |
|                        | 4             | Sixth segment and following               |
| Advance                | 1             | Booked 60 days or less prior to departure |
|                        | 0             | Booked 61 days or more prior to departure |
| Options                | 0             | None                                      |
|                        | 1             | Choice of one of any offered «go-options» |
| Gender                 | 0             | None in PNR (incomplete)                  |
|                        | 1             | Male or female                            |
| Frequent flyer program | 0             | None                                      |
|                        | 1             | Member of frequent flyer points program   |
| Booking class          | 1             | Tango (T, E, P, G, N, K, R)               |
|                        | 2             | Tango plus (B, H, V, Q, A, L, S)          |
|                        | 3             | Latitude (Y, M, U)                        |
|                        | 4             | Executive (J, C, Z, I)                    |
|                        | 5             | Aeroplan (W, D)                           |

|                               |         |   |
|-------------------------------|---------|---|
| Origin of full itinerary      | 1 to 10 | According to geographic location                |
| Destination of full itinerary | 1 to 10 | According to geographic location                |
| Point of sale                 | 1 to 10 | According to geographic location                |
| Electronic ticket             | 0       | No  |
|                               | 1       | Presence of electronic ticket number in PNR     |
| Round-trip booking            | 0       | No  |
|                               | 1       | Full itinerary origin and destination are equal |
| OD_type                       | 4       | Flights not included in AC Inc.                 |
|                               | 6       | Flights included in AC Inc. but not AC only     |
|                               | 7       | AC only flights (also included in AC Inc.)      |
| Departure time                | 1       | From 6:00 to 7:59                               |
|                               | 2       | From 8:00 to 9:59                               |
|                               | 3       | From 10:00 to 12:59                             |
|                               | 4       | From 13:00 to 15:59                             |
|                               | 5       | From 16:00 to 17:59                             |
|                               | 6       | From 18:00 to 19:59                             |
|                               | 7       | From 20:00 to 21:59                             |
|                               | 8       | From 22:00 to 5:59                              |
| Number in party               | 1       | Single passenger in PNR                         |
|                               | 2       | Two passengers                                  |
|                               | 3       | Three passengers or more                        |
| Boarding status               | 0       | No (« no-show »)                                |
|                               | 1       | Yes   |