STATISTICAL MODELS FOR THE SUPPORT OF OVERBOOKING IN TRANSPORT SERVICE[5]

Vladimir Averkiev, Mikhail Alexandrov, Javier Tejada

***Abstract**: Overbooking is a strategy of extra ticket reservation, which needs precise models concerning the probability of ticket acquisition. Without such models a company will have losses related to unused seats or compensation for the absence of these seats. In the paper we consider several models for such a forecast and demonstrate their functionality on real data of one Peruvian railway company. The best model includes an original algorithm of revealing seasonal prevalence.*

***Keywords**: overbooking, transport service, statistics*

***ACM Classification Keywords**: 1.6.4. Model validation and analysis*

Introduction

Companies associated with transport services, often use overbooking strategy. This strategy consists in an redundant reservation of tickets when some passengers are expected to refuse their trip in future. Such a strategy must be supported by accurate predictive models, which could answer the question about the share of purchased tickets among all reserved tickets for a given train (or flight) in a given day. Here:

- pessimistic forecasts may lead to losses related to unused seats;
- optimistic forecasts may lead to losses related to compensation for passengers, which will not have seats.

In the latter case the company being responsible for ticket reservation may have a reputation loss.

The problem of determining virtual capacity of vehicles were considered in [Barnhart, 2003] and [Talluri, 2005]. In the paper [Mozgovaya, 2011] virtual capacity of transport were estimated on the basis of forecasts for the number of purchased tickets and the number of return tickets. In this paper we consider two forecasting models. The first model does not take into account any characteristics related to a given trip.. It is considered as a basic model. The second model takes into account the travel date and the interval of reservation. This is two-parameter model. It can be built on the basis of regularities concerning seasonality and reliability of preliminary reservation. To reveal these regularities we use two original algorithms. The first algorithm realizes time series decomposition. It was described in [Averkiev, 2012]. The second algorithm determines the intervals of constancy for the mean values of time series. Here the method of discrete dynamic programming is used [Bellman, 1962]. In both cases the visual presentations of ticket sales are good helpers for decision making.

The paper is structured by the following way. In section 2 we describe source data. In section 3 we present algorithms and results of experiments. Section 4 contains conclusions.

[5] The work done under support of the British Petroleum grant (BP-RANEPA 2013).

Data and models

Data description

Initial information for modeling is data of one Peruvian railway company. It is results of daily ticket sales during 2012 from January 1 to December 31. Table 1 shows parameters of various reservations.

Table 1. Initial data (example).

N	*D1*	*D2*	*D3*	*D4*
1000567	16.12.2011	06.04.2012	1	1
1033112	22.03.2012	25.04.2012	4	0
2034566	22.03.2012	25.04.2012	2	2
………	…	…	…	…

Here: *N* is the order number, *D1* is the date of reservation, *D2* is the travel date, *D3* is the number of reserved seats, *D4* is the number of used seats.

Conditions of booking and purchase of tickets as the follows:

- Several reservations can be made in the same day and for the same date of trip;
- Reservation interval is from 0 to 365 days (one year);
- Ticket sales can be done in any day including the day of the trip.

Models under consideration

The aim of the work is to forecast ticket sales on a given day. For the simplicity we assume that there is only one train per day related to a given route. This forecast concerns the share of purchased tickets among reserved tickets. *Hereinafter we will associate this share with the probability that a given reserved ticket will be really bought.* So, if we know this probability *p* and the number of reserved tickets Q_R then the number of purchased (used) tickets Q_U can be calculated by a simple formula $Q_U = p\ Q_R$. We consider several models. Each model is based on one of the hypotheses:

Hypothesis 0. There are no any regularities related to probability *p*

Hypothesis 1a. There is a stable dependence between the probability *p* and the date of trip

Hypothesis 1b. There is a stable dependence between the probability *p* and the interval of reservation.

Hypothesis 2. There are stable dependences between the probability *p*, the date of trip and the reservation interval.

List of all models presented in Table 2:

Table 2. Forecast models with their parameters

Model	***Day of trip***	***Interval of Reservation***
Basic model	-	-
Intermediate decomposition model	+	-
Intermediate interval model	-	+
Two-parametric model	+	+

The more detailed models can be considered. For example, a three-parametric model can take into account additionally the number of reserved tickets. Usually, the more number of tickets are reserved the less probability of purchased tickets is. In this paper we consider only the models mentioned in the table.

To build models we use data of the first 270 days for training models and the last 90 days for testing models. To evaluate model quality MAPE index is used. MAPE stands for Mean Absolute Percent Error.It is calculated by the formula:

$$MAPE = \frac{1}{90}\sum_{t=271}^{360}\left|\frac{\widehat{p_t}-p_t}{p_t}\right| * 100\%$$

where $\widehat{p_t}, p_t$ are data of modeling and data of experiments.

Preprocessing

To construct the models listed in the table, preprocessing is performed. This preprocessing consists in 2 operations:

1) Check of values. All data should be positive; the number of purchased tickets must be no more than the number of reserved tickets, etc.

2) Compression of data. All reservations with the same date of trip and the same date of reservations are combined: the number of reserved tickets and the number of purchased tickets are summed. Besides the reservation interval is calculated. Table 3 is a result of preprocessing:

Table 3. Data after preprocessing (example).

D	***T***	***R***
97	110	0.38
97	112	0.32
116	33	0.29
…	…	…

Here: *D* is a day of trip (from January 1, 2012), *T* is a reservation interval measured in days; *R* is a share of purchased tickets.

Algorithms and experiments

Basic model

The basic model does not consider any regularities in ticket reservation concerning time. It is assume that the ratio of the number of purchased tickets to the number of reserved tickets is a constant. According to the data of Table 3 we have:

$$p_D = \bar{p} = \frac{\sum_{i=1}^{270} R_i}{270}$$

Dependence R_D on day D is shown in Figure 1.

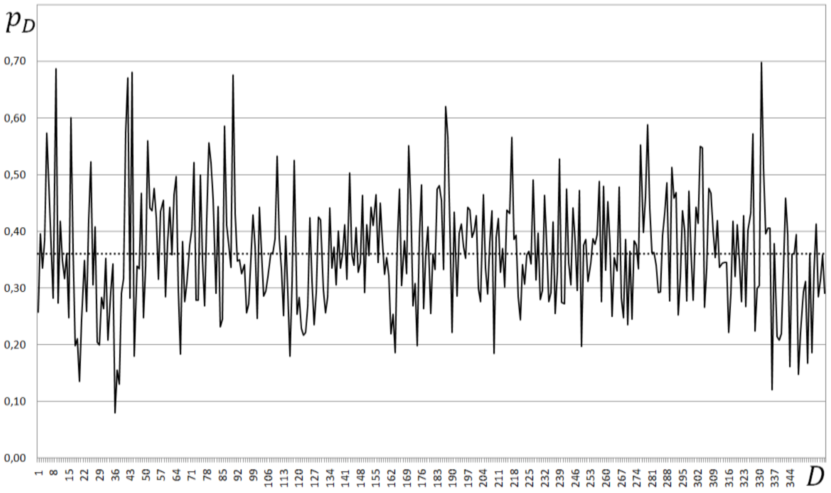

Fig.1. Dependence of ratio R_D on day D

The experimental results are as follows:

- The probability of purchasing ticket p_D = *0.36;*
- The accuracy of forecast MAPE = *26,7%.*

Seasonality identification

Seasonality allows to take into account the date of the trip. To reveal the seasonality we need to decompose a given time series into the components, i.e. we need to build such a model $p_t = T_t + S_t + N_t$. Here:

- T_t is the trend, the main component;
- S_t is the seasonal component, which gives information about the periodic oscillations in time series;
- N_t is the random component of time series.

The method of curvature minimization is used to evaluate the trend and seasonality. The corresponding algorithm is described in [Averkiev, 2012]. The forecast model based on the decomposition is built by the following way:

1) First, the local decomposition in a current step t of the time series is implemented. Thus, we find seasonal coefficients k_t.

2) Second, we calculate the local time series average value using a moving average:

$$p_{t+1} = \frac{\sum_{i=1}^{16} p_{t+1-i}}{16}$$

3) And finally, it is determined the forecast value of probability like the sum of time series average value and seasonal coefficients k_t:

$$\hat{p}_{t+1} = \frac{\sum_{i=1}^{16} p_{t+1-i}}{16} + k_t$$

Identification of patterns in reservation period

To reveal the regularities related to the length of reservation period let see on Figure 2 Axis X is the reservation period, and axis Y is the ratio of the number of purchased tickets to the number of reserved tickets. As have mentioned above we associate this value with a probability.

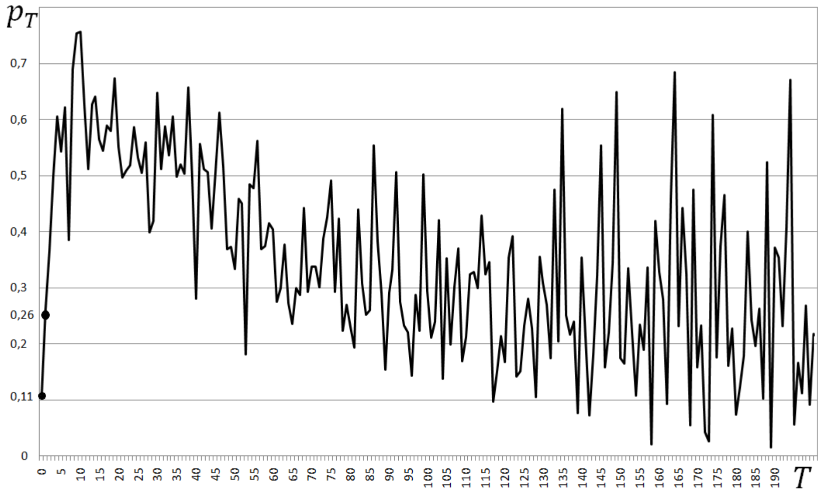

Fig. 2. Dependence of ratio R_T on interval of reservation T.

The figure clearly shows that a) the values p_T for bookings in the day of departure and one day before the departure (i.e. T=0 and T=1) are significantly smaller than p_T for other days (we mean here the stable values) b) there are no data on graph concerning intervals $T \geq 200$ because that data are not representative. With the method of discrete dynamic programming [Bellman, 1962] we revealed two intervals related to mean value of p_T on these intervals (the number of intervals were fixed but the boundary between them were unknown). Just the first two points proved to belong to the first interval. The forecast model based on the selected interval is built by the following way:

1) Each booking gets its weight. Weight of booking is equal to the probability of its realization. Bookings made in the day of departure and one day before the departure get weights 0.11 and 0.26, respectively. All other bookings get weight 0.36.

2) All weighted bookings for the departure day are summarized. As a result, we obtain the desired ratio $\hat{p}_t$.

The values 0.11 and 0.26 are taken from the Figure 2. The value 0.36 corresponds to the basic model. One should say that the selection of two intervals is enough subjective. In future it is necessary to consider more objective procedures.

Two-parametric model

In two-parametric model we use decomposition of a given time series together with taking into account interval of reservation. The model is built by the following way:

1) Trend and seasonal component are estimated

2) Correction is made using interval of reservation.

3) The final probability is the composition of the results of previous steps. We calculate the probability by the formula:

$$\hat{p}_D = \frac{\sum_{i=1}^{N} I_D R_i}{\sum_{i=1}^{N} I_D} + \frac{\sum_{j=1}^{16} p_{D-j}}{16} + k_D^{seasonal}, where\ I_D = \begin{cases} 1, when\ D_i = D \\ 0, when\ D_i \neq D \end{cases}$$

Here *D* is the day of departure, and *N* is the total number of bookings related to day *D*. In this expression the first term is the correction related to the interval of reservation, the second term and the third terms take into account the local mean value and seasonality.

Testing on real data gives MAPE = 23.7%. Figure 3 shows the experimental data (thick grey line) and the forecast data (thin black line). One should remind that we use the interval 270 days for learning model and the last 90 days for testing model.

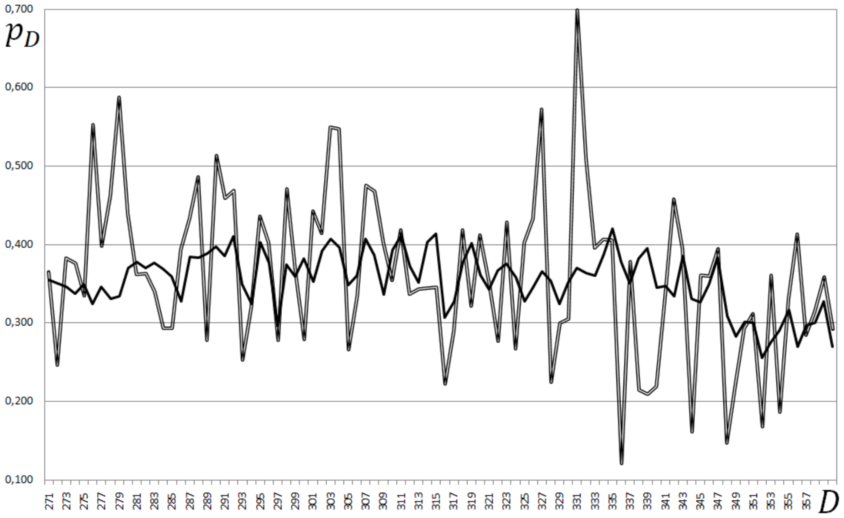

Fig. 3. Real data and forecast on the examination period

This figure shows that the series p_D is still strongly volatile but the most of variations are explained by an irregular component. It is also seen that the forecast well predicts general fluctuations of the trend

Conclusions

In the paper we build several models to predict the probability for purchase of reserved tickets. The best two-parametric model takes into account the day of departure and the interval of reservation.

We tested the basic and two-parametric models on real data related to the activity of one Peruvian railway company. The two-parametric model provides the accuracy 23.7%. In the comparison with the basic model with its accuracy 26.7% we have the absolute improvement 3% and the relative improvement 11,5%. This result shows the possibility to increase the efficiency of overbooking strategy.

Bibliography

[Averkiev, 2012] A. Kovaldji, V. Averkiev, M. Sarkissyan. Smoothing and prognosis of multi-factor time series of economical data by means of local procedures (regression and curvature evaluation) // Artificial intelligence driven solutions to business and engineering problems. ITHEA Publ, vol. 27, 2012, pp. 27-31

[Barnhart, 2003] Barnhart C., Belobaba P., Odoni A. Transportation Science // Applications of Operations Research in the Air Transport Industry, No.4, vol.37, 2003 , pp. 368-391

[Bellman, 1962] Bellman, R., Dreyfus, S. Applied dynamic programming // Amazon, 1962

[Mozgovaya, 2011] Mozgovaya K., Yablochkina M., Friedman G. Numerical analysis of the influence of predictive accuracy of the passenger demand on the efficiency of ticket sales with the overbooking // Scientific and Technical Bulletin Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, №6

[Talluri, 2005] Talluri K., van Ryzin G. The Theory and Practice of Revenue Management // Springer, 2005, pp. 129-160

Authors' Information

Vladimir Averkiev – M.Sc. student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State Research University); Institutskii per 9., Dolgoprudny, MoscowRegion, 141700, Russia e-mail: *vlaverkiev@gmail.com*

Major Fields of Scientific Research: mathematical modeling, system analysis.

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: malexandrov@ mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Javier Tejada – *Professor of Computer Science Department, San Pablo Catholic University; Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Perú; e-mail:* jtejada@ itgrupo.net

Major Fields of Scientific Research: Natural Language Processing, Business Intelligence