A Data Mining approach for trip time prediction in mass transit companies

João M. Moreira[1], Alipio Jorge[2], Jorge Freire de Sousa[3], and Carlos Soares[4]

[1] Engineering Faculty, Porto University, Portugal jmoreira@fe.up.pt
[2] Economics Faculty, Porto University, Portugal amjorge@liacc.up.pt
[3] Engineering Faculty, Porto University, Portugal jfsousa@fe.up.pt
[4] Economics Faculty, Porto University, Portugal csoares@liacc.up.pt

Abstract. In this paper we discuss how trip time prediction can be useful for operational optimization in mass transit companies and how data mining techniques can be used to improve results. Firstly, we analyze which departments need trip time prediction and when. Secondly, we review related work and thirdly we present the analysis of trip time over a particular path. We proceed by presenting experimental results conducted on real data with the forecasting techniques we found most adequate, and conclude by discussing guidelines for future work.

1 Introduction

Our aim is to determine whether trip time prediction is a valuable management decision support tool for mass transit companies. As a case study we are using STCP - Sociedade de Transportes Colectivos do Porto, SA, the bus transport operator for Oporto - Portugal, but we expect that the results of this study can be generalized to other mass transit companies mainly in developed countries.

The estimation of trip time is required by different departments inside a mass transit company at different times. Trip time prediction can be useful, typically, in four different situations:

1. For the definition of timetables for trips: these predictions are required several months in advance and cover a long period, usually months.
2. For the definition of the crew's duties: this information is required by the operational managers at a time period prior to the trip. In the case of STCP, changes in the scheduled trip time are made at least three days in advance.
3. For real time adjustments: to have an up to the minute prediction of what will happen at any given moment on the current day. This is very important in a situation where there are, necessarily, just in time management policies of unforeseen circumstances.
4. For client information: short term (few hours of anticipation) trip time prediction can also be used for marketing activities such as the information service by sms - short message service, or bus stop information system.

The trip time prediction for the definition of the crew's duties is our goal. With more accurate predictions, the operational managers can organize the duties of the crews better, which enables a better usage of company resources. In particular, it reduces costs in terms of extra time payed to the crews.

2 Related work

The INRETS - Institut National de Recherche sur les Transports et leur Securité, has conducted a study of traffic flow prediction for the period one or two days in advance of the trip ([6]). In addition they have developed traffic flow prediction algorithms for horizon periods of one week and one year. The techniques used for travel time prediction and for traffic flow are not necessarily the same as the ones used for trip time prediction. Travel time is a generic term to refer the expected time to travel between two points while trip time refers to one specific journey. In the latter case, there is, usually, past data obtained in comparable conditions. In fact, while travel time is, usually, obtained by calculus from measures such as volume, occupancy and speed, trip time is measured directly.

Factors that can explain trip time, such as the number of dwells or the total number of passengers alighting the bus, are discussed in [1].

3 A case study: data analysis of one path

The STCP company has an Operational Control System that includes, among others, an automatic vehicle location by GPS - Global Positioning System. It reports the start and the end of the trip, and the vehicle's position every 30 seconds, as well as other pieces of relevant information about the trip. The data for this study covers a period from January to August 2004. It includes 7166 trips, all of them from the same path and direction. For each trip, we have collected the start and end times, date, vehicle model, driver and duty number.

To gain insights into the factors determining trip time we have analyzed the data by visual inspection ([7]). Several seasonal / impact components were identified: the seasonality by period of the day; the day of the week seasonality; the day of the month seasonality; the week of the year seasonality; the national holidays impact; and the school breaks impact.

Other explanatory factors can be important to explain trip time, namely, weather conditions, occasional events (a football match, the annual students party, the visit of an international leader, etc.), road works in a particular stretch of the path, type of path, driver's behavior, etc. We expect that these factors may be identified by the proposed approach.

4 Analysis of results

The problem we are dealing with is one of regression, and in particular, of time series forecasting. Our goal is to evaluate whether a data mining approach can be

useful in an operational decision support environment in a mass transit company. A variety of different techniques can be applied. An overview is provided by, among others, [3] and [5].

We present results using random forests ([2]), projection pursuit regression ([4]) with three different smoother methods (super smoother, spline and GCV), and support vector machines ([8]) with two different kernels (linear and radial). Tests were done using data from January 1st 2004 to March 31st 2004, i. e., 2646 trip records. These tests are a first approach to the selection of the best forecasting techniques for the trip time prediction problem. In this first approach we use just 3 months data to reduce test time. The training is done on a sliding window one month long and the test data is one day long. The lag between the train and test sets is three days long. The explanatory variables used are: (1) start trip time (in seconds); (2) day type (holiday, normal, ...); (3) weekday; and (4) day of the year. The target variable is the trip time duration (in seconds). The series is irregularly time spaced with 29,5 trips per day in average.

The best result for each one of the tested models is presented in Table 1. Each result is the variation index of the trip time calculated as the ratio between the squared root of the mean squared error, for all the predictions with the same parameter set, and the average of the corresponding real values of the trip time.

RF	PPR			SVM	
	supsmu	spline	gcvspline	linear	radial
9.92%	10.60%	11.15%	11.80%	14.50%	12.97%

Table 1. Variation index

It is clear that the best overall result is obtained by Random Forests, followed by Projection Pursuit Regression and Support Vector Machines.

To obtain a more detailed perspective of these results, we have partitioned the data using a regression tree. Then we evaluated the predictions obtained in the experiments described above, grouping the examples based on the leaf nodes of the tree. Table 2 illustrates that the best algorithm varies significantly for the examples in different nodes. We also observed that, if we were able to pick the best model for each leaf node, the overall results would improve significantly.

This is a relevant result that confirms the need to conduct such tests in order to get information about how different techniques act under different conditions and, consequently, to get the most of them. It also shows that we can improve predictions once we are able to choose the most adequate model for the value we want to predict.

5 Future work

Future work will include the following points: (1) to test if splitting data in the learning phase may improve results; (2) to study the features selection; (3) to

	rf	ppr.supsmu	ppr.spline	ppr.gcvspline	svm.linear	svm.radial	Best	Size
1	21,84%	7,01%	**5,83%**	7,49%	7,46%	6,97%	5,83%	11
2	9,52%	9,46%	9,47%	9,62%	**9,39%**	9,49%	9,39%	649
3	16,76%	9,46%	6,93%	**6,76%**	8,52%	7,43%	6,76%	15
4	10,30%	**7,09%**	9,92%	11,01%	8,93%	8,69%	7,09%	28
5	**7,44%**	9,64%	8,03%	7,71%	7,80%	8,54%	7,44%	217
6	**6,15%**	8,09%	10,18%	10,66%	25,73%	13,96%	6,15%	104
7	11,10%	**8,24%**	11,98%	12,10%	13,01%	11,17%	8,24%	45
8	13,45%	**11,02%**	12,88%	14,69%	16,54%	15,96%	11,02%	265
9	**6,66%**	12,73%	7,48%	6,81%	7,66%	7,83%	6,66%	335
10	**5,69%**	7,28%	11,19%	12,67%	21,69%	10,89%	5,69%	82
11	7,10%	9,05%	**6,47%**	8,05%	10,59%	8,48%	6,47%	37
							8,36%	1.788

Table 2. Variation index of trip time obtained by the algorithms on the leaf nodes of the regression tree. Column "Size" represents the number of examples in each node.

expand these tests to other techniques, namely, neural networks, local regression and exponential smoothing in order to get a wider technique perspective; (4) to enrich the data with external features that are known to have an impact on trip times; (5) to quantify how a better trip time prediction accuracy may improve the operational results of a transport company.

Acknowledgments

This work was partially supported by FCT - Fundação para a Ciência e a Tecnologia, project reference: POCT/TRA/61001/2004 and FEDER e Programa de Financiamento Plurianual de Unidades de I&D.

References

1. R. L. Bertini and A. M. El-Geneidy. Modelling transit trip time using archived bus dispatch system data. *Journal of transportation engineering*, 130(1):56–67, 2004.
2. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
3. V. Cherkassky and F. Mulier. *Learning from data: Concepts, theory, and methods*. John Wiley and Sons, Inc., 1998.
4. J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of american statistical regression*, 76(376):817–823, 1981.
5. T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data minig, inference, and prediction*. Spinger series in statistics. Springer, 2001.
6. S. V. Iseghem and M. Danech-Pajouh. Prevision du trafic a j+1 (j+2) une approache intermodale (in french). *Recherche Transports Securite*, (65):79–97, 1999.
7. J. M. Moreira, A. Jorge, J. F. d. Sousa, and C. Soares. Trip time prediction in mass transit companies. a machine learning approach. In *16th Mini-EURO Conference: "Artificial Intelligence in Transportation"*, Poznan, 2005. Submitted to.
8. A. J. Smola and B. Scholkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, 1998.