TRIP TIME PREDICTION IN MASS TRANSIT COMPANIES. A MACHINE LEARNING APPROACH.

João M. MOREIRA*, Alípio JORGE[†], Jorge Freire de SOUSA[‡], Carlos SOARES[§]

Abstract. In this paper we discuss how trip time prediction can be useful for operational optimization in mass transit companies and which machine learning 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: this prediction is made several months in advance and covers 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.

[*] Engineering Faculty, Porto University, Portugal, jmoreira@fe.up.pt
[†] Economics Faculty, Porto University, Portugal, amjorge@liacc.up.pt
[‡] Engineering Faculty, Porto University, Portugal, jfsousa@fe.up.pt
[§] Economics Faculty, Porto University, Portugal, csoares@liacc.up.pt

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.

2. Related work

There has been great interest in the study of short-term (few hours of anticipation) travel time prediction in recent years. The reason for this is its importance for ATIS - Advanced Traveler Information Systems. Several techniques, or combinations of techniques, have been used including: multiple regression ([9]), the local linear regression model ([14]), neural networks ([8]), the ARIMA model ([16]), Kalman filters ([3]), support vector regression ([17]), simulation models ([3]), k-nearest neighbors ([14]), principal component analysis ([1]) or spectral analysis.

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 ([7]). 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 last 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.

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 all the 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.

The aim of this section is to gain insights into the factors determining trip time, by visual inspection of the data. Several seasonal / impact components can be identified:

- The seasonality by period of the day: analyzing figure 1 (left-hand side) one can identify this type of seasonality.
- The day of the week seasonality: trip time throughout the day looks differently according to the day of the week (figure 1). Figure 2 (left-hand side) shows the differences between average daily trip time for different week days. The days of the week with the shortest average trip times are Sundays or national holidays, whilst the

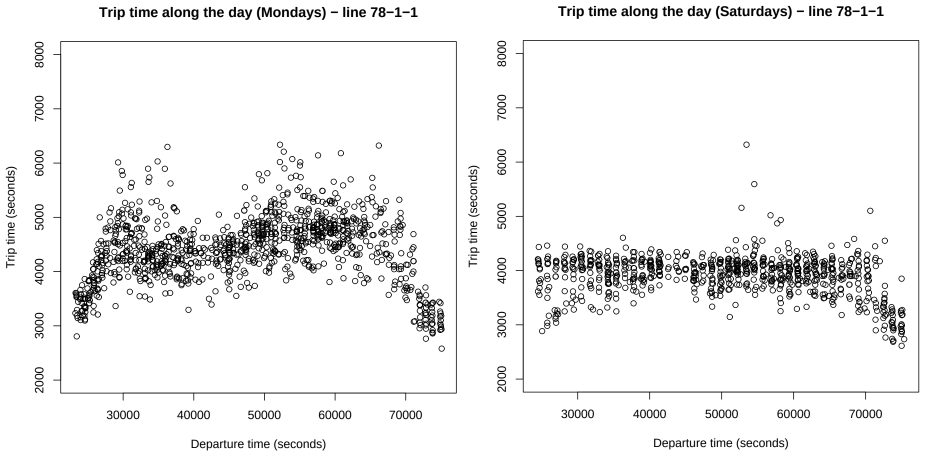

Figure 1. Trip time along the day on Mondays and on Saturdays

days of the week with the longest average trip times are working days. Average trip times for Saturdays fall between these two extremes.

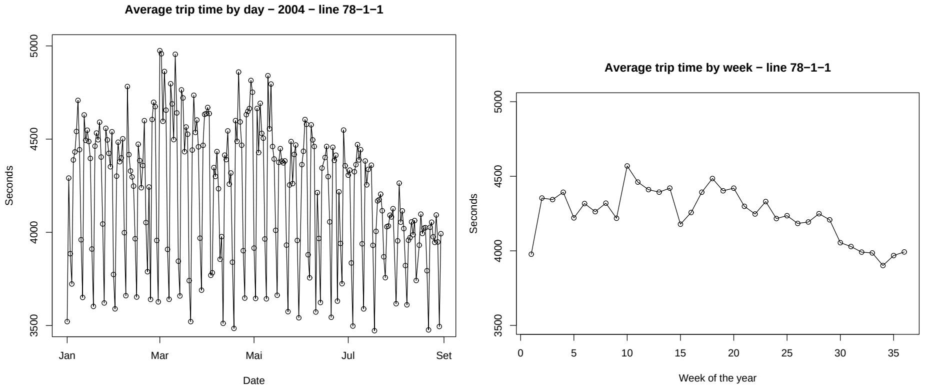

Figure 2. Daily and weekly average of trip time

- The day of the month seasonality: not easy to identify. However, on Sundays previous to a typical pay day, the trip times are shorter. This can be explained from the fact that at this period of the month people have less money and also because on Sundays, the traffic is mainly for family purposes and not for work.

- The week of the year seasonality: we do not have enough data to make meaningful conclusions from a statistical point of view, however the expected variations throughout the year are apparent.

- The national holidays impact: this impact changes according to the day of the week. If the holiday is on a Tuesday or Thursday, it is likely that workers take Monday or

Friday off. This potentially increases the impact of holidays. From the period of time covered by this study it is not possible to obtain meaningful statistical conclusions regarding the extent of this impact.

- The school breaks impact: there are four main school vacations per year in Portugal. These are the Carnival, Easter, summer and Christmas breaks. From figure 3 it can be seen there is a differing trip time throughout the day for school vacations (right-hand side) and for normal working days (left-hand side). In figure 2 (right-hand side) shorter trip times during weeks 15 - 16 (Easter) and after week 30 (summer holidays) are evident.

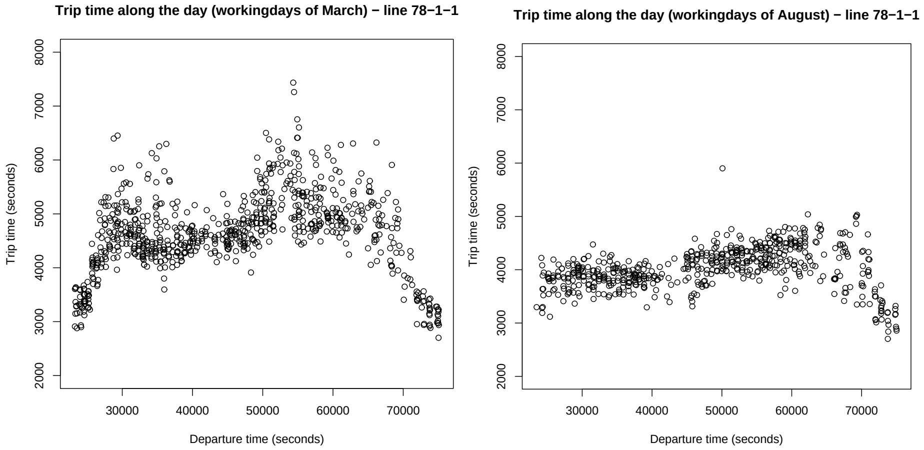

Figure 3. Trip time along the day (working days - March and August)

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. Experimental setup

The problem we are dealing with is one of regression, and in particular, of time series forecasting. Our goal is to evaluate whether a machine learning approach can be useful in an operational decision decision support environment in a mass transit company. A variety of different techniques can be applied. An overview is provided by, among others, [4], [6] and [10].

We present results using random forests ([2]), projection pursuit regression ([5]) with three different smoother methods (super smoother, spline and GCV), and support vector machines ([13]) with two different kernels (linear and radial). We used SVM μ-regression instead of ϵ-regression once [12] reports similar results for both and the μ-regression has lower and upper limits. For these tests we have used the R statistical package ([15]).

Tests were done using data from January 1st 2004 to March 31st 2004, i. e., 2646 trip records from the path previously analyzed. The input variables used are: (1) trip start time, (2) day type (normal, bank holiday, ...), (3) week day, and (4) day of the year. The target variable is the trip time. The variables trip start time, day of the year and the target variable are numeric while day type and week day are symbolic. We used the sliding window strategy with 30 days as training set and one day as test set. The test set is three days ahead from the training set. For each set of parameters we have 59 training sets and 59 test sets. Our evaluation criterion is the mean squared error.

The selection of parameters was done using a grid search. This search had, in some cases, several iterations to expand and/or to refine the grid. Initially, the grid search was done according, mainly, to [11].

5. Analysis of results

The best result for each one of the tested models is presented in table 1. Each result is the mean squared error for all the predictions with the same parameter set. Figure 4 presents the distribution of the differences between those predictions and the real values for each case.

RF	PPR			SVM	
	supsmu	spline	gcvspline	linear	radial
186 565	212 875	235 542	263 717	398 737	319 083

Table 1. Mean squared error (in squared seconds)

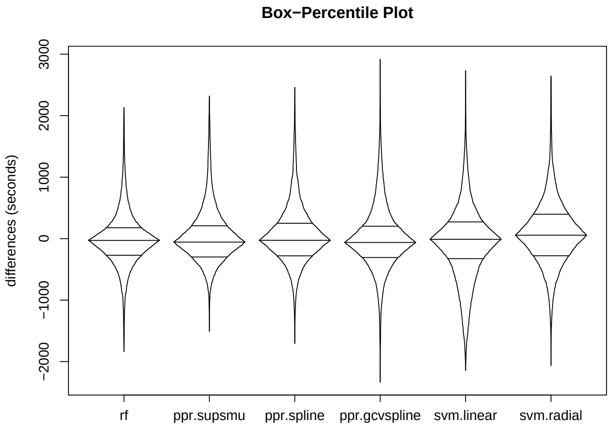

Figure 4. Differences box-percentile plot

It is clear that the best overall result is obtained by Random Forests, followed by Projection Pursuit Regression and Support Vector Machines. If we split results on Saturdays, Sundays and working days (table 2), the best results will still be the ones of Random Forests. However, the best parameter set is not necessarily the same for each one of the mean squared error results presented in both tables. This means that there are differences inter and intra groups of cells formed by each technique. This is a relevant result that confirms the reasonability of conducting such tests in order to get information about how different techniques act under different conditions and, consequently, to get the most of them.

	RF	PPR			SVM	
		supsmu	spline	gcvspline	linear	radial
W.D.	211 681	229 557	251 891	281 357	447 105	352 186
Sat.	67 659	84 699	93 179	94 636	136 886	124 258
Sun.	107 990	133 020	132 184	144 248	163 967	164 206

Table 2. Mean squared error (in squared seconds)

6. Conclusions and future work

We have presented preliminary research on how trip time prediction can be used for operational optimization in mass transit companies. We have also analyzed a data set by visual inspection in order to obtain insights into trip time seasonalities. Finally, we have presented experimental results for Random Forests, Projection Pursuit Regression and Support Vector Machines. We have shown that the best model from the three we have tested is random Forests.

Future work will include the following points: (1) to verify if by splitting data in the evaluation phase by other features rather than weekdays will improve results; (2) to test if splitting data by weekdays or other features in the learning phase may improve results; (3) to study the features selection; (4) to expand these tests to other techniques, namely, neural networks, local regression and exponential smoothing in order to get a wider technique perspective; (5) to enrich the data with external features that are known to have an impact on trip times; (6) 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] P. Bickel, C. Chen, J. Kwon, J. A. Rice, P. Varaiya, and E. V. Zwet. Traffic flow on a freeway network. In *Workshop on Nonlinear Estimation and Classification*, 2001.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] M. Chen and S. Chien. Dynamic freeway travel time prediction using probe vehicle data: link-based vs path-based. In *Transportation Research Board*, 2001.

[4] V. Cherkassky and F. Mulier. *Learning from data: Concepts, theory, and methods*. John Wiley and Sons, Inc., 1998.

[5] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of american statistical regression*, 76(376):817–823, 1981.

[6] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data minig, inference, and prediction*. Spinger series in statistics. Springer, 2001.

[7] 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.

[8] L. Kisgyorgy and L. R. Rilett. Travel time prediction by advanced neural network. *Periodica Polytechnica Ser. Civ. Eng.*, 46(1):15–32, 2002.

[9] J. Kwon, B. Coifman, and P. Bickel. Day-to-day travel time trends and travel time prediction from loop detector data. *Transportation Research Record 1717*, pages 120–129, 2000.

[10] S. Makridakis, S. C. Wheelwright, and V. E. McGee. *Forecasting: methods and applications, 2nd edition*. Wiley, 1983.

[11] D. Meyer, F. Leisch, and K. Hornik. Benchmarking support vector machines. Technical Report 78, Vienna University of Economics, 2002.

[12] B. Scholkopf, A. J. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. Technical Report NC2-TR-1998-031, 1998.

[13] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, 1998.

[14] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran. Short term traffic forecasting using the local linear regression model. In *Transportation Research Board annual meeting*, 2003.

[15] R. D. C. Team. R: A language and environment for statistical computing. Technical report, R Foundation for Statistical Computing, 2004. ISBN 3-900051-07-0.

[16] M. V. d. Voort, M. S. Dougherty, and M. V. Maarseveen. Combining kohonen maps with arima time series models to forecast traffic flow. *Transportation Research Part C: Emerging Technologies*, 4(5):307–318, 1996.

[17] C.-H. Wu, C.-C. Wei, D.-C. Su, M.-H. Chang, and J.-M. Ho. Travel time prediction with support vector regression. In *IEEE Intelligent Transportation Systems Conference*, volume 2, 2003.