SHORT TERM TRAFFIC PREDICTION USING CLUSTER ANALYSIS BASED ON FLOATING CAR DATA

Alexander Sohr and Peter Wagner
DLR - Institute of Transport Systems
Rutherfordstr. 2, 12489 Berlin – Germany
+49-30-670-55-458, alexander.sohr@dlr.de
+49-30-670-55-237, peter.wagner@dlr.de

ABSTRACT

Precise short-term prediction of traffic parameters such as flow and travel-time is a necessary component for many ITS applications. This work describes the research on a novel, fast, and robust algorithm which is based on a partitioning cluster analysis. It is able to calculate travel times from Floating Car Data (FCD) for a whole city, even for minor roads.
A potential problem with FCD is the insufficient penetration rate of smaller taxi fleets and the resulting noisy and/or missing data (4). The new approach accounts for this by smoothing the data by a local fit method based on polynomials with the help of a Singular Value Decomposition (SVD). Numerical experiments confirm the high efficiency of the algorithm and a promising quality of the prediction.

KEYWORDS

Prediction, traffic, floating car data, cluster analysis, smoothing, singular value decomposition

INTRODUCTION

Floating Car Data (FCD) collected from vehicle fleets are an excellent technology for the traffic surveillance needed to support the various mobility services of Intelligent Transport Systems. Data from an FCD-fleet result in maps of travel time of the area under consideration. In the following, the taxi FCD system developed by DLR ITS, Berlin (1) is briefly reviewed. Every car of a taxi-fleet sends its current GPS-position via digital mobile radio at prescribed time intervals (10 s to 180 s) to its taxi-headquarter. After using for fleet disposition the data is stored and converted into XML-format. At the DLR the XML-formatted datasets are retrieved via File Transfer Protocol (FTP) and then the data is stored in a database. These GPS positions are map-matched. The resulting data is the current travel speed and a coverage-value for each edge, where data had been captured. Coverage is the distance the vehicle has travelled on the current link, divided by the length of the link. That way, traffic-condition maps for an entire road-network of urban areas are being generated.

PREPARING THE DATA

Unfortunately, the raw FCD data is very noisy. Hence the data must be aggregated or smoothed to make it comparable. Here we use a novel approach based on Singular Value Decomposition (SVD) to smooth the data. With this approach it is possible to create a daily-course with only 100 values. We pick a window of 10 data points and fit the data in this window by a 3rd order polynomial $p(t)$. Therefore, within this window the following error is minimized:

$$e = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{p(t_i) - v(t_i)}{\sigma_i}\right)^2$$

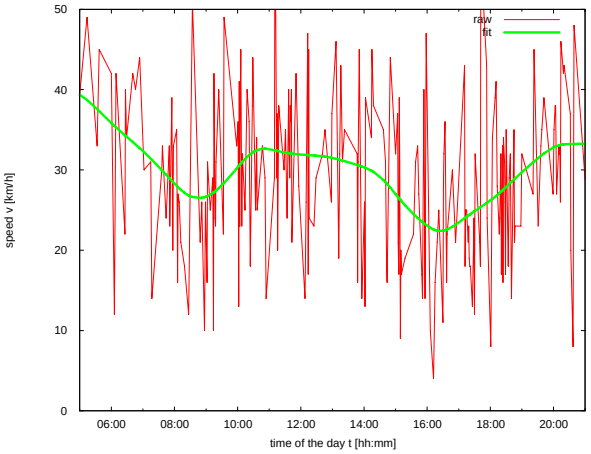

Fig.1 Daily course of raw FCD-data vs. SVD-fitted

The window is moved along the time-series from data-point to data-point, and at each data-point a new fit is being performed. This is necessary, since the data-points are not equidistant in time, so simpler approaches like Savitzky-Golay (3) filtering do not work. In this equation, $p(t)$ is the local polynomial, $v(t)$ is the measured speed and σ_i is an approximation of the (local) standard deviation of the data. In addition to this local approximation, the data are further aggregated into 20 minutes bins; furthermore, the data from the last four weeks are used. This results in 72 aggregated values for each day of the week. The aggregation is done with the harmonic mean, weighted by the number of data-points. Remaining periods which still do not have data are filled by linear approximation. This results in a daily course for each edge. To make them more comparable (even minor roads to free way), every daily course is scaled and squeezed into the interval [0,1]. Before that, its average speed is saved for later reconstruction.

Clustering

For the clustering we use the program “cluster” as described in (2). We use a partitioning clustering with simple k-means cluster algorithm and Euclidean distance as space function. Extensive investigations showed that six clusters is the best number to distinguish different behaviors of the weekdays and, at the same time, also to merge related days. The associated hydrograph curve (cluster nr) is reconstructed with the average speed of all cluster members.

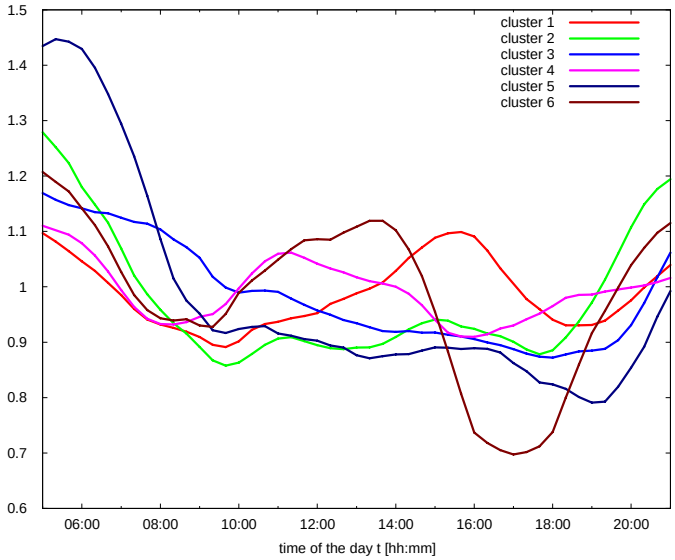

Fig.2: Clusters for Wednesdays in Berlin in January 2008

PREDICTION

At time t the prediction is done for the time $t+n$, where n is the time which will be looked into future. For every edge, that had a current value $v(t)$ since the last prediction a new prediction is calculated. The predicted speed $\hat{v}(t+n)$ consists of three parts:

1. $d_c(t+n)$, the relative speed from daily course, which is a result from the aggregation of all members of one cluster (see fig.2). This is from the cluster the edge belongs to.
2. the last measured value $v(t)$ of the edge.
3. the relative mean speed of the clustered daily course of the last N hours.

   $$\frac{1}{N}\sum_{i=t-N}^{t} d_c(i)$$

   These three are put together in the following equation to arrive at the predicted travel speed

$$\hat{v}(t+n) = \frac{d_c(t+n)}{\frac{1}{N+1}\sum_{i=t-N}^{t} d_c(i)}\, v(t)$$

Basically, the historical daily course is used to scale the current speed value by the factor historical speed at time t+n divided by the mean value of the historical speeds in a given past window.

EXPERIMENTAL RESULTS

The comparison to a naïve prediction, build from the last 4 weekdays is shown in fig. 3. The new prediction method is better than the linear prediction, indicated by the higher peek at ‘0’ and by the RMSE indicated in the Figure. While the naïve method yields an error of around 16%, the improved method gives an error of 12%.

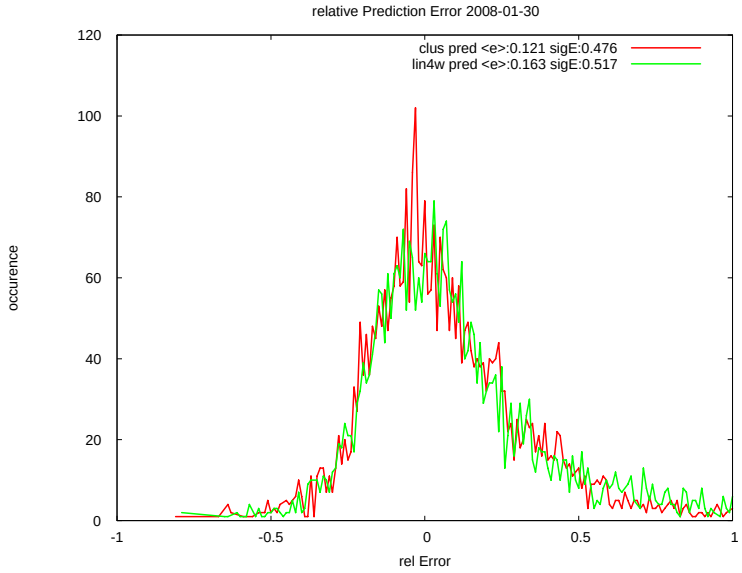

Fig. 3 Distribution of relative errors – cluster prediction vs. linear prediction

CONCLUSION

Solving the problem of noisy FCD data is the main improvement of this prediction method. With the smoothed daily course good short term travel time prediction results can be achieved. Furthermore a compact form for saving historic speeds is generated. The experiments that we conducted show that the approach described here has a better prediction quality than naïve methods.

REFERENCES

(1) Ralf-Peter Schäfer, Kai-Uwe Thiessenhusen, Elmar Brockfeld and Peter Wagner, “A traffic information system by means of real-time floating-car data”, ITS World Congress 2002, Chicago (USA)
(2) M. J. L. de Hoon, S. Imoto, J. Nolan and S. Miyano, ”Open Source Clustering Software“, Bioinformatics 20 (9), 2004, p.1453 – 1454
(3) William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, “The Art of Scientific Computing, Cambridge University”, Numerical Recipes 3rd Edition, Press 2007
(4) Hong-En Lin, Rocco Zito and Michael A P Taylor, “A review of travel-time prediction in transport and logistics”, Proceedings of the Eastern Asia Society for Transportation Studies Vol. 5, 2005, p.1433 – 1448