Modeling Traffic Information using Bayesian Networks

W.P. van den Haak*, L.J.M Rothkrantz*, P. Wiggers
Department of Computational Intelligence, Delft University of Technology, Delft, the Netherlands
** Corresponding authors: paul.vandenhaak@tno.nl, L.J.M.Rothkrantz@tudelft.nl*

B.M.R. Heijligers, T. Bakri, D. Vukovic
TNO, Intelligent Transport Systems, Delft, The Netherlands

DOI: 10.2478/v10158-010-0018-09

ABSTRACT: Dutch freeways suffer from severe congestion during rush hours or incidents. Traffic congestion increases travel time, resulting in a delay for travelers. To avoid these delays, rerouting traffic around congested areas is an option. Reliable travel time predictions are essential for dynamic routing and travel information. Travel time can be calculated from vehicle speed measurements (van Lint, 2004). These speed measurements are acquired from dual inductive loop detectors collected by the Dutch Monitoring Casco (MONICA) data system. In this paper, the predictability of average vehicle speed by Bayesian Networks is investigated in a case study. We propose a general Bayesian Network model and evaluate several simplified versions of this model on a well known traffic bottleneck in the Netherlands. We show that our Bayesian Network is capable of predicting the start and end of a congestion for a prediction horizon of 30 minutes with an accuracy of 14%. Furthermore, we present a prediction model based on historical data which is evaluated on the same bottleneck. This prediction model based only on historical data and our Bayesian Network are combined in a hybrid model, where we evaluate performance as well. This hybrid model is able to predict congestion with an accuracy of 85% for a rather long prediction horizon of 2.5 hours in our case study.

KEY WORDS: Bayesian Networks, prediction, vehicle speed, inductive loop detector data.

1 INTRODUCTION

Many highways in the Netherlands suffer from severe congestion at some times of the day. Regular congestions can be anticipated, but incidental congestions will increase people's travel time and raise frustration. Travel times are important for travelers, as well as for road managers. Accurate travel time predictions give travelers or road managers the ability to anticipate road congestion and to estimate their arrival time.

Travel time τ for a certain route can be computed as follows:

$$\tau = \sum_{i=1}^{N} \frac{d_i}{V_i},$$

where d_i denotes the distance and V_i denotes the travel speed on subsection i respectively and $1 \le i \le N$. The fact that speed V_i is actually a function of time is disregarded

for convenience purposes at this moment. There are different approaches for traffic prediction: an instantaneous approach, a model based approach, or data driven approaches. Instantaneous approaches assume that traffic conditions remain stationary for the time the prediction holds (van Lint, 2004). In general, traffic does not remain stationary if time passes (Lindveld & Thijs, 1999).

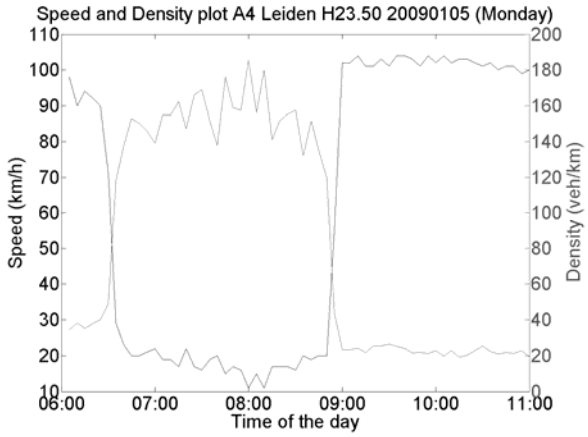

Figure 1a: Combined speed and density graph.

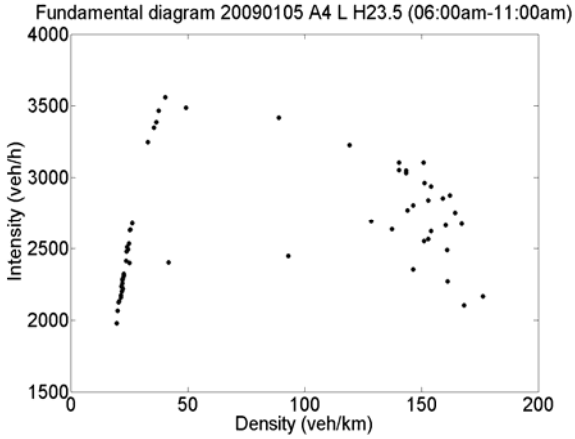

Figure 1b: Combined intensity and density graph.

Most congestion occurs during the morning or evening rush hours. Therefore, accurate speed predictions for these periods are essential. About 64% of the traffic during congestion consists of commuter traffic (Rijkswaterstaat, 2006). This group of travelers certainly benefits from accurate travel time predictions to plan their day.

Traffic propagates in space and time and is influenced by its context, such as weather, events, accidents, road works, or other incidents which makes traffic an uncertain domain. Bayesian Networks allow us to reason about an uncertain domain (Korb & Nicholson, 2004). We believe that there are strong correlations between traffic measurements such as speed or density at different locations. Furthermore, we have to deal with missing values, which are a common problem in traffic research (van Lint, 2004). Bayesian Networks can handle incomplete datasets and allow one to learn about causal relationships (Heckerman, 1995).

Traffic speed, density, and intensity represent traffic situations. Only two of these variables need to be known to be able to calculate the other. The Bayesian Network we propose in this paper incorporates density and speed measurements. Figure 1a shows a combined graph of speed and density from a morning rush hour at one location of the A4 highway in the Netherlands. This figure shows a switch point, where the speed decreases and the density increases. Figure 1b shows a combined data plot of intensity and density measurements on the same location at the A4 motorway. This figure shows that the intensity (flow) decreases when the density increases. There is an unstable traffic situation just before the congestion. We expect that congestion cannot be seen in these graphs a few hours beforehand, and therefore we combine our Bayesian Network with a historical data model in a hybrid model (Van Den Haak et al. 2010).

The outline of this paper is as follows. Section 2 details related work. In Section 3 we present our models. A description of the data is given in Section 4, and this data is used in our experiment which is described in Section 5. In Section 6 we present the results of our experiment and in Section 7 we present our conclusions and discussion.

2 RELATED WORK

Different approaches on modeling traffic with Bayesian Networks can be found in literature (Sun et al. 2004), (Sun et al. 2005), (Sun et al. 2006), (Yu & Cho, 2008). These approaches only include one traffic variable in their network, either traffic flow or traffic speed. Figure 2 shows a road network and the corresponding Bayesian Network model developed by Sun et al. (Sun et al. 2004) for predicting traffic flow with incomplete data. Each circle is a node where travelers can change direction. This model predicts traffic flow (intensity) at D_d and it is assumed that flow measurements at D_d for previous times influence the prediction. Furthermore, upstream links C_e, C_g, and C_h for different time instants also influence the prediction. Yu et al. (Yu et al. 2008) propose a Bayesian Model which includes upstream links as well as downstream links.

Our Bayesian Model differs since it includes speed as well as density measurements combined with historical data predictions in a hybrid model.

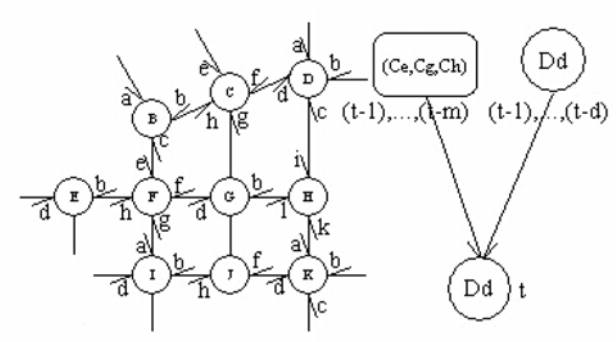

Figure 2: Bayesian Model proposed by Sun et al. (Sun et al. 2005).

3 PREDICTION MODELS

We propose a Bayesian Network which incorporates speed and density measurements for multiple locations and multiple time instants. Figure 3 shows a graphical representation of this model for k locations and i time instants. We have used GeNIe & SMILE from the Decision Systems Laboratory of the University of Pittsburgh to model, train and test our Bayesian Model. If multiple locations and time instants are incorporated, the model quickly becomes complex and computationally expensive to train. To cope with this problem, our Bayesian Network is kept simple and is connected to an adaptive prediction model which is based on historical data in a hybrid model.

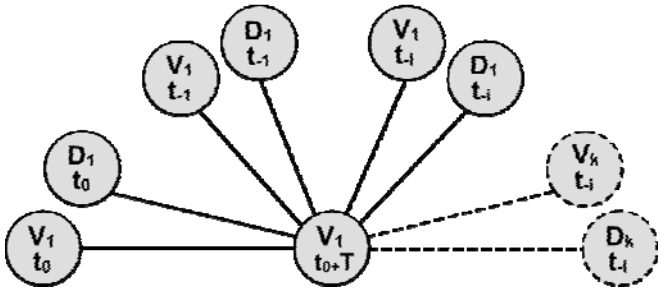

Figure 3: General Bayesian Network.

a. Bayesian model

At this stage of research, we propose a simplified version of our general Bayesian Model for our experiment. This model incorporates speed and density measurements for three different locations at the A4 highway in the Netherlands. The causal nodes take speed and density measurements at time t_0. This model is trained for different prediction horizons T, where $T \geq t_0$. We expect that this Bayesian model is able to predict short term congestion. In this paper, short term predictions lie in the range from 5 to 30 minutes.

b. Historical model

We propose an adaptive learning prediction model based on historical data. This model searches a database with traffic models of speed and density for every location. The model database contains mean patterns of four different clusters: (1) Monday-Thursday, (2) Friday, (3) Weekend and (4) Holiday, and the raw data which is not clustered. These clusters have been found during an explorative data analysis for homogeneous clusters in the traffic data. Figure 4a and Figure 4b show examples of the models for speed and density respectively.

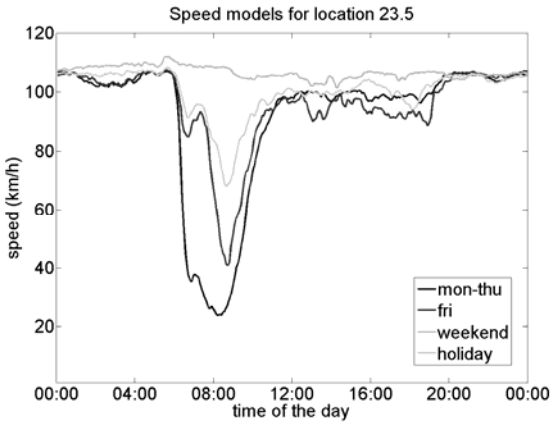

Figure 4a: Speed models.

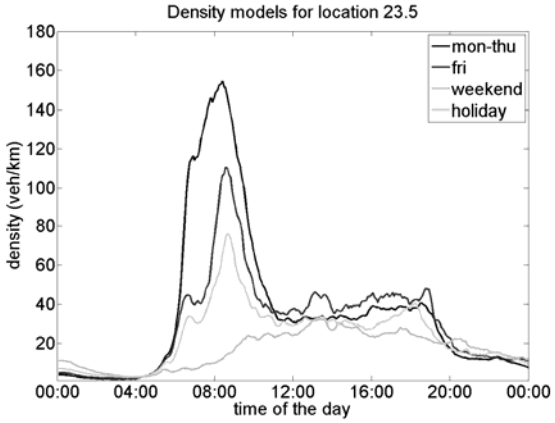

Figure 4b: Density models.

Figure 5a shows an example cluster of speed measurements from a location at the A4 in the Netherlands for a total dataset containing 365 days of 2009. Figure 5b shows the cluster containing only Tuesdays for the same location. This figure shows that the data becomes more homogeneous after clustering.

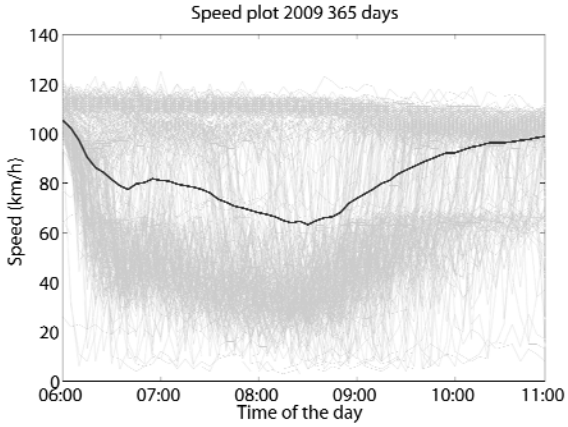

Figure 5a: Cluster for 365 days.

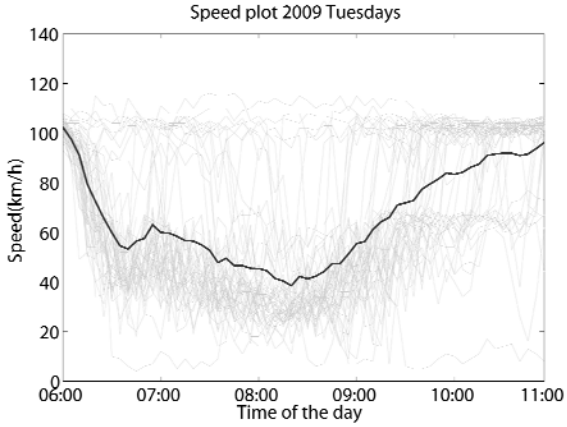

Figure 5b: Cluster for Tuesdays.

Our prediction model based on historical data is adaptive since it is able to find the best fitting clustered model when new data becomes available. It also tries to find a best fitting day in the total data which is not clustered. This dataset is likely to contain typical days containing accidents, extreme weather, or other incidents. Therefore, our model is able to adapt to new situations continuously. We expect that our adaptive prediction model based only on historical data is able to predict congestion for longer prediction horizons.

c. Hybrid model

We propose a hybrid model, in which we use historical data in our adaptive prediction model to compute initial predictions for speed and density for a prediction horizon T_h. These initial predictions are the input for the three locations in our Bayesian Model. Our Bayesian Model then predicts for a horizon of T_b, based on these initial predictions from the historical data of our adaptive learning prediction model. In total, we have a prediction horizon $T = T_h + T_b$. We expect that traffic congestion is not visible in the data a few hours ahead, and therefore we first conduct a prediction based on historical data for several locations. These initial predictions are given to our Bayesian Network which is able to model the relations between the locations on the road. We therefore expect our hybrid model to predict traffic congestion for long prediction horizons (from 60 to 150 minutes) accurately.

4 INDUCTIVE DUAL LOOP DETECTOR DATA

Real traffic data is one of the most important elements in analyzing and improving traffic systems. Real data gives us a better representation of the situation compared to simulated data. Traffic flow is often measured by inductive loop detectors (ILD) because of their widespread availability (Vanajakshi, 2004). The Dutch freeway system is equipped with inductive dual loop detectors, which are capable of measuring speed and intensity per lane. An example of an ILD can be found in Figure 6. The data of these detectors are collected in the Monitoring Casco (MONICA) data system, managed by the Dutch Ministry of Transport, Public Works, and Water Management.

In general, the MONICA data systems contains on average 12% missing values (van Lint, 2004), due to maintenance, accidents, and power or communication failures. To decrease the size of the dataset and the number of missing data points, our dataset has been aggregated from lane data to road data. This aggregation procedure takes the harmonic average over the lane measurements. We choose to take the harmonic mean, because speed measurements are used as a rate when calculating travel times. More details about computing the average of vehicle speed can be found in (van Lint, 2004). The intensity value for a road is the summation of the intensities per lane. The missing values in our data are filled in by the Treiber and Helbing filter (Treiber & Helbing, 2002), which reconstructs the spatio- and temporal traffic characteristics.

5 EXPERIMENT

We tested our models on an important traffic bottleneck in the Netherlands at the A4 from The Hague to Amsterdam around hectometer location 23.5 at the left side. The capacity of the road decreases from 3 lanes to 2 lanes which results in traffic congestion during rush hours or incidents. Figure 7 shows a schematic overview of this test location. Location A, B and C represent the locations in our Bayesian Network model.

A dataset has been extracted out of the MONICA data system for the year 2009 for locations A, B and C. A train and a test set has been randomly selected out of the 365 days. The train set contains 80% (292 days) and the test set contains 20% (73 days) from the total set.

Figure 6: Example picture of an ILD (Taale, 2006).

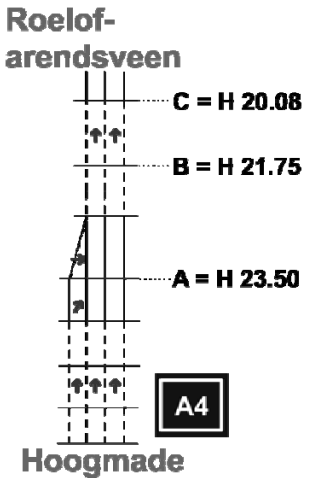

Figure 7: Schematic overview.

The Bayesian Network is trained for different time horizons, and evaluated for the corresponding time horizon. For this paper, all our models have been evaluated.

The sMAPE is used to compute the difference between the real data and the predicted data as follows:

$$sMAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_t - F_t}{A_t + F_t}\right|,$$

where A_t is the actual value and F_t is the forecast value at time t and $1 \le t \le N$. Since this measure only tells us the relative distance over all time instances, we need another measure to evaluate if our models are able to detect congestion.

Therefore, we introduce a measure for false positives and false negatives. False positives denote a situation in which the prediction model predicts a congestion, while there is none. False negatives denote a situation in which the prediction model predicts no congestion, while there is one. The general aim is the keep the false negatives low, as an error here would result in unexpected congestions which is not desirable for travelers.

For comparison purposes, we introduce another model: the Naïve model. This model only takes all speed values until the current time t_c and assumes that traffic remains stationary. This model can be expressed as follows: $V_{c+T} = V_c$ and will be referred to as the Naïve model.

6 RESULTS

The result of the sMAPE, the false positives and the false negatives percentage for the Bayesian Network model, are presented in Table 1. This table shows that our Bayesian Network has a slightly lower false negatives percentage, but is comparable to the Naïve Model.

Table 2 shows the results for our adaptive learning model based on historical data. The results show that our adaptive learning model outperforms the Naïve model when the prediction horizon increases. The results of the hybrid model are presented in Table

3. This table shows that our hybrid model clearly outperforms the Naïve model on the false negatives percentage. This means, that our hybrid model is clearly better in predicting congestion then the Naïve model.

Table 1: Experiment results for the Bayesian Network.

T	sMAPE Bayesian Network	sMAPE NAÏVE	FP Bayesian Network	FP Naïve	FN Bayesian Network	FN Naïve
5	3 %	3 %	4 %	4 %	5 %	7 %
10	5 %	5 %	8 %	6 %	4 %	12 %
15	6 %	6 %	7 %	7 %	7 %	15 %
20	7 %	7 %	12 %	9 %	10 %	18 %
25	7 %	7 %	13 %	10 %	12 %	20 %
30	8 %	8 %	14 %	11 %	14 %	21 %
35	8 %	9 %	15 %	12 %	15 %	22 %
40	9 %	9 %	15 %	13 %	20 %	24 %
60	12 %	11 %	19 %	18 %	29 %	29 %
120	11 %	15 %	11 %	28 %	35 %	47 %

Table 2: Experiment results for the adaptive prediction model based on historical data.

T	sMAPE Historical model	sMAPE Naïve	FP Historical model	FP Naïve	FN Historical model	FN Naïve
60	14 %	16 %	13 %	15 %	35 %	38 %
120	17 %	24 %	17 %	19 %	41 %	66 %

Table 3: Experiment results for the hybrid model.

T	sMAPE Hybrid	sMAPE Naïve	FP Hybrid	FP Naïve	FN Hybrid	FN Naïve
65	16 %	17 %	19 %	16 %	14 %	50 %
70	16 %	18 %	21 %	17 %	12 %	42 %
75	17 %	19 %	22 %	18 %	13 %	44 %
80	17 %	20 %	24 %	19 %	12 %	47 %
85	17 %	21 %	24 %	20 %	14%	48 %
90	18 %	21 %	24 %	21 %	15 %	50 %
125	16 %	25 %	25 %	20 %	17 %	68 %
130	16 %	25 %	27 %	21 %	15 %	70 %
135	17 %	26 %	28 %	22 %	14 %	72 %
140	17 %	26 %	29 %	23 %	12 %	74 %
145	18 %	27 %	29 %	24 %	14 %	76 %
150	18%	27 %	29 %	25 %	15 %	77 %

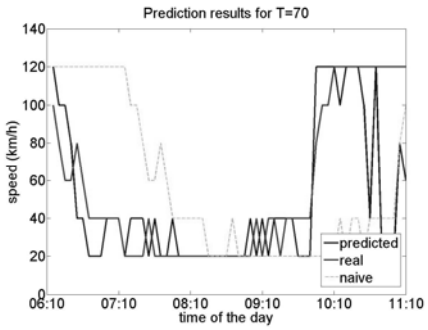

Figure 7a: Example result for T=70.

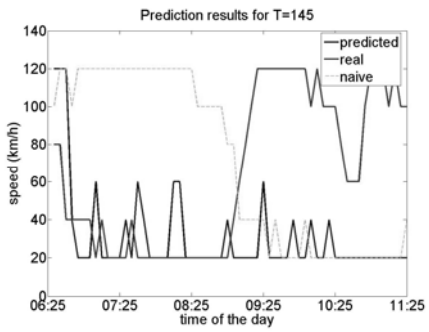

Figure 7b: Example result for T=145.

Figure 7a and Figure 7b show an example of a prediction plot for T=70 and T=145 respectively for the hybrid model. Figure 7a shows that our hybrid model is able to predict the start and end of the congestion accurately for 70 minutes ahead. In the middle of the congestion, the prediction shows stochastic behavior. Figure 7b shows that for longer prediction horizons it becomes more difficult to predict congestion.

7 CONCLUSIONS & DISCUSSION

In this paper we showed that our Bayesian Network is able to perform comparably to the Naïve model. The Naïve model only makes an error if the traffic changes, and therefore is shown to be a strong comparator. Our adaptive learning model is able to predict traffic congestions for longer horizons in which it outperforms the Naïve model. The combination of historical models and our Bayesian Model is shown to be a powerful one. Our hybrid model clearly outperforms the Naïve model.

We show that the results of our hybrid model are promising. It would be interesting to test our models on more locations for different datasets to see if this conclusion can be drawn for other situations as well.

REFERENCES

Lindveld, C.D.R., Thijs, R., 1999, *On-line travel time estimation using inductive loop data: The effect of peculiarities on travel time estimation quality.* Proceedings of the 6th ITS World Congress, Toronto, Canada.

Sun, S., Zhang, C., Yu, G., 2006, *A Bayesian Network approach to traffic flow forecasting.* IEEE Transactions on Intelligent Transportation Systems, vol. 7, pp. 124-132.

Sun, S., Zhang, C., Yu, G., Lu, N., Xiao, F., 2004, *Bayesian Network methods for traffic flow forecasting with incomplete data.* Lecture notes on computer science, Springer Berlin, vol. 3201, pp. 419-428.

Sun, S., Zhang, C., Zhang, Y., 2005, *Traffic flow forecasting using a spatio-temporal Bayesian Network predictor.* Lecture notes on computer science, Springer Berlin, vol. 3697, pp. 273-278.

Taale, H., 2006, *Analyzing loop data for quick evaluation of traffic management measures.* European Transport Conference, Strasbourg, France.

Treiber, M., Helbing, D., 2002, *Reconstructing the spatio- temporal dynamics from stationary detector data*. Cooperative Transportation Dynamics, vol. 1, 3.1-3.24.

Vanajakshi, L.D., 2004, *Estimation and prediction of travel time from loop detector data for intelligent transportation systems and applications.* PhD thesis, Texas A&M University.

Van Den Haak, W.P., 2010, *Modeling Traffic Information using Bayesian Networks*, Master Thesis, Delft University of Technology, Delft, The Netherlands.

Van Lint, J.W.C., 2004. *Reliable travel time prediction for freeways.* PhD Thesis, TRAIL Research School, the Netherlands.

Yu, Y.J., Cho, M., 2008, *A short-term prediction model for forecasting traffic information using Bayesian Networks.* Third international conference on Convergence and Hybrid Information Technology, vol. 1, pp. 242-247.