Transportation (2005) 32: 649–669
DOI: 10.1007/s11116-005-0219-y

Short-term prediction of travel time using neural networks on an interurban highway

SATU INNAMAA
VTT Technical Research Centre of Finland, P.O. Box 1800, 02044 VTT, Finland
(E-mail: satu.innamaa@vtt.fi)

Key words: neural networks, prediction, travel time, two-lane highway

Abstract. The main purpose of this study was to investigate the predictability of travel time with a model based on travel time data measured in the field on an interurban highway. Another purpose was to determine whether the forecasts would be accurate enough to implement the model in an actual online travel time information service. The study was carried out on a 28-kilometre-long rural two-lane road section where traffic congestion was a problem during weekend peak hours. The section was equipped with an automatic travel time monitoring and information system. The prediction models were made as feedforward multilayer perceptron neural networks. The main results showed that the majority of the forecasts were close to the actual measured values. Consequently, use of the prediction model would improve the quality of travel time information based directly on the sum of the latest measured travel times.

1. Introduction

1.1. *Background*

Traditionally, traffic flow has been monitored by point measurements. As traffic control systems develop from local control towards area control, the extent of the monitored area increases and implementation of monitoring by point measurements turns costly – especially when installing detectors in rural areas which lack power and communication infrastructure on the road side. A large network can be covered with fewer detectors when applying section-based monitoring than with point-based monitoring to get the same coverage. In addition, for example, the average travel time of a section gives a good picture of the flow status in the section (Haugen 1996).

Travel time is also one of the most important pieces of information that road users need from the road operator. Road users will benefit more from accurate travel time information where there is great variability in travel times. Therefore, road users expect information to be up-to-date if the actual travel time varies substantially. Travel time information based directly on the sum of the latest measured travel times is always outdated and the longer the section is, the more outdated the information is. The reason for this is that by definition a vehicle has to drive the whole section before its travel time can be determined. Thus, the vehicles used for measuring the travel time are different from the vehicles whose driver sees the information on the road based on those particular measurements (Figure 1). Without short-term prediction, real-time information on travel time cannot be given.

Much research has been done over the past decade in the field of travel time prediction. Many studies are based on simulated, unbiased data, which leads to well-performing models (Yasui et al. 1995; Suzuki et al. 2000; You & Kim 2000; Chen & Chien 2001; van Lint et al. 2002, 2003; Nanthawichit et al. 2003). However, these models cannot cope equally well with imperfect, real-life data. Real-life applications should be robust with respect to faulty and incomplete input (van Lint et al. 2002).

Automatic travel time monitoring systems are not common, and although the whole road network cannot be covered completely with loop detectors, the traffic information collected by inductive loops is used as input for many models that predict travel time (Saito & Watanabe 1995; Lee & Choi 1998; Lee et al. 1998; Al-Deek 2003; D'Angelo et al. 1999; van Grol et al. 1999a, b; Kwon et al. 2000; Lindveld et al. 2000; van Lint 2003;

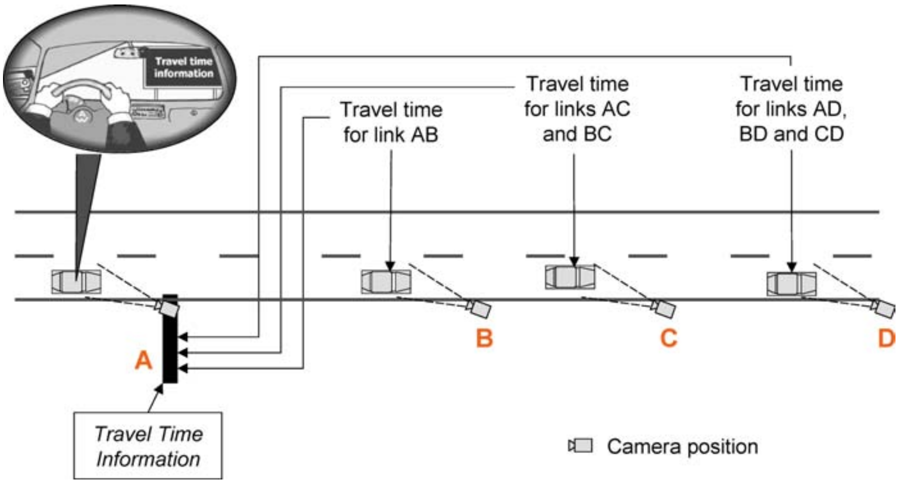

Figure 1. The vehicle whose driver sees the travel time information on the variable message sign, and the vehicles on whose travel time the information is based. The travel time of a section is determined as the difference between the passing times at two camera stations.

Matsui & Fujita 1998; McFadden et al. 2001; Paterson & Rose 1999; Zhang & Rice 2003). Even though the general relations between travel time and traffic volume, occupancy and point speed have been widely explored, these relations might not apply during saturated flow conditions (Chien and Kuchipudi 2002). However, in those conditions the travel time information is most valuable.

Studies in which travel time forecasts are based on abundant field measurements of highway travel time are few. Chien and Kuchipudi (2002) predicted travel time with a Kalman filtering algorithm and Ohba et al. (2000) with pattern recognition based on travel time data provided by tagged vehicles on an urban motorway. Park and Rilett (1998), Park et al. (1999) and Rilett and Park (2001) predicted travel time with neural networks based on travel time data provided by an AVI system on an urban motorway. To our knowledge, the current literature does not include prediction models based on field measurements of travel times made for two-lane $(1+1$ lanes) two-way highways.

Interurban two-lane two-way highways with relatively small capacity and at-grade junctions differ from motorways, and they are more sensitive to the impacts of the incidents. Consequently, the results obtained from them cannot be generalised to two-lane highways as such. However, two-lane highways are important because they carry most of the traffic in most countries. Thus, there is a lack of knowledge in the field of predicting travel time on interurban two-lane highways based on real-time field measurements of travel time.

1.2. *Objectives*

The purpose of this study was to investigate, first, the predictability of travel time with a model based on travel time data measured in the field on an interurban two-lane two-way highway. Second, the purpose was to determine whether the forecasts would be accurate enough to implement the model in an actual travel time information service. Specifically, a target was set to get 90% of the forecasts within a 10-percent error margin (±10%). Toppen and Wunderlich (2003) found that, when the accuracy of the travel time information system drops below a critical point, one is better off not using the information but relying on the experience with historical traffic patterns. According to their study in Los Angeles, that point is in the range of 13–21% error. Thus, the limit we have chosen for the target was tighter than that and corresponded well to an everyday way of evaluating travel times.

2. Method

2.1. *Study site*

The research was carried out on the main road 4 between the cities of Lahti and Heinola in Southern Finland. The study site was a 28-kilometre-long two-lane two-way highway section with alternating passing lanes. Because the site was located between two motorways, traffic congestion was a problem during weekend peak hours. Specifically, the section was congested on Friday and Sunday afternoons and evenings almost every weekend between the beginning of May and the end of October. The free-flow travel speed on the section was around 100 km/h. In congested conditions, the travel time might be up to three times the normal – especially in the northbound direction on Fridays.

The study section was equipped with an automatic travel time monitoring system. The system was based on an image processing and neural network application, which automatically reads licence plates at four locations in both directions (Finnra 2000). The system was capable of reading on average 40% of all licence plates on the monitored lane at a single point in good conditions when the camera was clean.

There were two types of monitoring stations within or nearby the study site: camera stations and inductive loop detectors. The study section was divided into three sub-sections with four camera stations (marked A, B, C and D in Figure 2) used to measure travel time. The distance between two consecutive camera stations varied between 8.7 and 10.3 km. The travel time measurement system covered one lane in each direction (Figure 2), i.e. the passing lanes were excluded. Thus, there was an unmonitored passing lane at every measurement point, expect for camera stations B and C in the northbound direction. The inductive loop detectors were installed at location C and 11.9 km south of location A on the other side of the nearby city that gathered information about traffic volumes and point speeds.

Drivers were informed of the travel time with variable message signs (VMS) at both ends of the road section. The goal of the system was to inform drivers about congestion and to offer an estimate of the expected travel time. The underlying rationale was that congestion is more tolerable when drivers are aware of the expected traffic conditions, as was shown by Luoma (1998). In addition, travel time information was provided on the Internet.

The travel time displayed on VMSs was not a forecast but an estimate of the travel time. This estimate was based on a sum of the latest measured travel times on each sub-link – or on a combination of sub-links in case one or two camera stations were not operating along the section. No matter

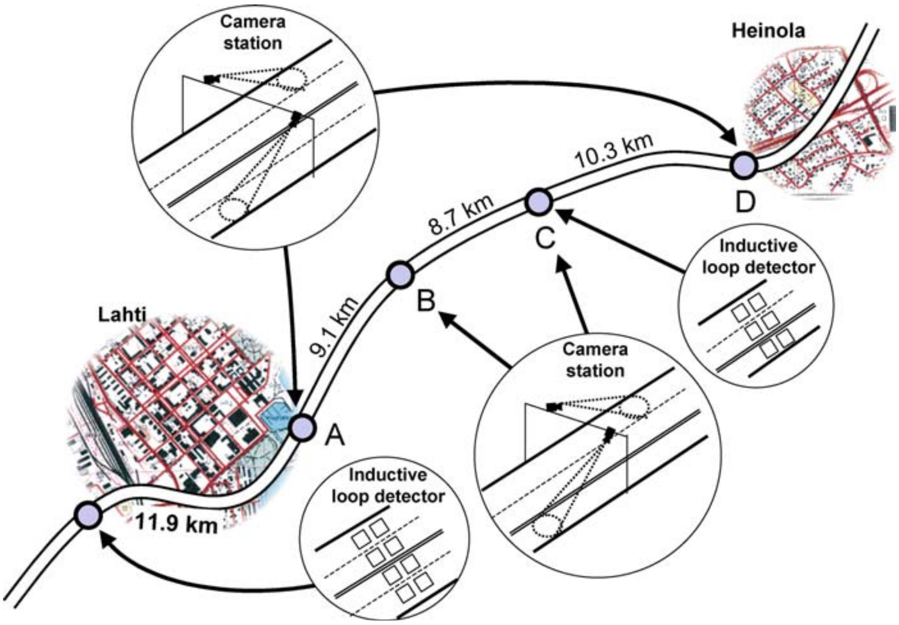

Figure 2. Study site with number of lanes, link lengths and traffic monitoring equipment. The section contained four camera stations (A–D) used to measure travel time and inductive loop detectors at location C and south of location A on the other side of the nearby city (outside the section).

how often the update was made, the problem was that the information was always outdated (Figure 3). In conditions when the mean travel speed was lower than 75% of the free flow speed, the VMS in the northbound direction gave correct travel time information (measured value was between the upper and lower limit shown on the VMS) 32.9% of the time and in the opposite direction 49.7% of the time.

If the road has an alternative route, the proportion of vehicles taking the detour as a function of the information shown on the VMS has to be considered. However, on the section of this study there was no real route alternative to which large traffic volumes could be directed. Consequently, this consideration could be ignored.

2.2. *Data*

The study was based on data collected during approximately 4 months of summertime. The data was collected 24 h per day, 7 days per week. The raw data of individual vehicles produced by the travel time monitoring system as well as inductive loops near the study site were included.

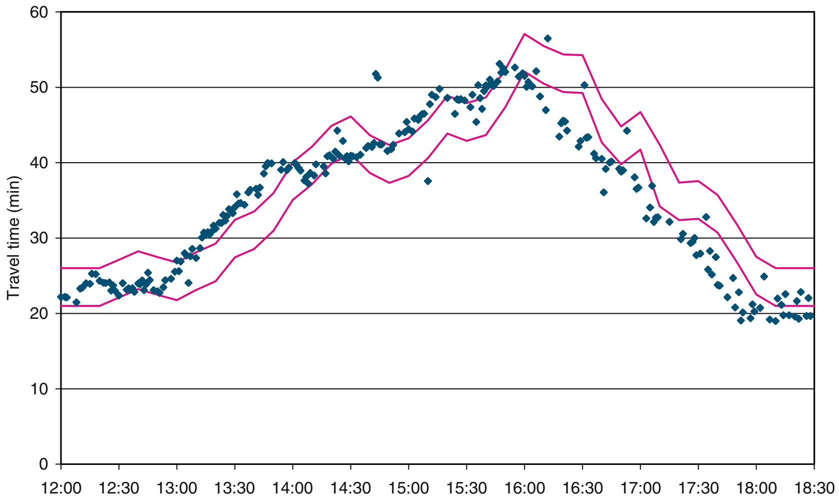

Figure 3. The travel time estimate (upper and lower limit) given by the non-predictive system based directly on the sum of the latest measured travel times lagged behind the correct travel time (dots). The example was from the section AD. The update frequency of the estimate was 10 min.

Most of the data was from free-flowing traffic (Table 1). Traffic was defined as congested if the mean travel speed was lower than 75% of the free-flow speed, i.e. slow, queuing or stopped traffic according to Kiljunen and Summala (1996).

The sample sizes were small in the travel time monitoring system (assessed value: 6–19% of travel times were measured, only sub-links CA,

Table 1. Traffic conditions in the data. The definitions were obtained from Kiljunen and Summala (1996).

Traffic	Free flowing (%)	Heavy (%)	Slow (%)	Queuing (%)	Stopped (%)	Number of obs.
Definition (% of v_{free})	>90	75–90	25–75	10–25	<10	
AD	77	16	7	0	0	30,323
AC	76	16	8	0	0	42,540
AB	83	13	4	0	0	43,694
BD	78	14	8	0	0	39,747
BC	72	20	6	2	0	51,690
CD	89	10	0	0	0	50,105
DA	74	19	6	0	0	12,712
DB	74	17	9	0	0	9,982
DC	77	15	8	0	0	18,579
CA	77	22	1	0	0	15,681
CB	71	29	1	0	0	13,855
BA	83	17	0	0	0	11,526

BC and CD having the proportion above 10% – leading for example on section DA with 73% of periods in free flowing traffic and 56% in slow traffic with only one vehicle sample) and the number of travel time observations was not equal to the flow. There were two reasons for this: first, the travel time measurement system covered only one lane per direction – as explained earlier – and the travel time of the vehicles driving on the unmonitored lane at either end of the section could not be measured. Second, the travel time monitoring system did not detect all the vehicles driving on the monitored lanes due to faults in the detection of a moving vehicle or in the automatic recognition of the licence plate. Besides the small sample size, the samples were also probably slightly biased towards longer travel times because the unmonitored lane was the overtaking lane. However, the bias did not matter in this case because although by monitoring also the overtaking lane the values would be different (i.e. lower travel times), the overall accuracy would probably be the same.

A moving average was used in filtering the individual vehicle travel time data in order to exclude deviating observations. Deviating observations mean observations of those vehicles that have stopped along the section or mismatches (i.e. too long or too short travel times). The moving average was calculated from seven consecutive travel time observations – the travel time under evaluation being in the middle. If the travel time under evaluation differed by more than 33% of the moving average, the situation was checked manually. If the large difference was not due to a rapid and dramatic change in travel time pattern, for example, then the observation was removed from the database.

The filtering procedure was developed based on trial and error. It seemed to detect the deviating observations well but left natural variation to the data. Partly manual checking was chosen because of the sensitiveness of an educated human eye. It is hard to replace it with a simple algorithm without losing also some extra data. As the number of observations measured in congested conditions was limited, it was in our interest to be able to use as much of it as possible. Therefore, an objective and rigorous alarm system (the moving average) was set to identify the data periods that may have some problems; but as congestion develops sometimes fast, they also may be samples of a true quickly increasing congestion. The educated human eye was used in a systematic way to resolve which the case was.

After filtering, the data was aggregated. This data included one-minute average and median travel times, and the median travel time and standard deviation of the observations from the latest five minutes or the latest 10 or 20 vehicles. In addition, it included the one-minute flow, mean point speed and standard deviation of the point speed at the inductive loop detectors.

2.3. *Prediction models*

The models were made as feedforward multilayer perceptron (MLP) neural networks as the model was to be kept simple but effective. MLP-neural networks are easy to implement and there have been encouraging results in previous studies using the same method (Park & Rilett 1999; McFadden et al. 2001; Shao et al. 2002). MLP-neural networks have also proven to be good in predicting other measures that describe the traffic situation like flow rate (Smith & Demetsky 1994, 1997; Lee et al. 1998; Innamaa & Pursula 2000).

A separate neural network was trained to predict the travel time of each sub-link. As output, the models gave the average travel time for vehicles entering the section within the following minute. The one-minute aggregation period was chosen to ensure fast detection of the changes in travel times. However, it was acknowledged that the one-minute average was less stable than, for example, a five-minute average as a drawback for fast detection of tendencies. The sample sizes would also have been greater for a longer aggregation period, but as this model was made for study purposes to run offline, and the raw data could be checked manually, the problems caused by the small sample size could be partly overcome.

The models obtained as input the traffic information based on the latest measurements. The input parameters are described later in more detail. The number of input neurons was equal to the number of input parameters; the number of output neurons was one, since there was just one output parameter. The input parameters were normalised to have a zero mean and standard deviation of one.

The structure of the neural network was to be kept simple. Neural networks were chosen to have one hidden layer. The number of hidden neurons was chosen so that the number of training samples was at least ten times the number of parameters to be estimated. However, the number of hidden neurons was limited to be at most 20 in order to keep the training process fast.

The activation function of the hidden layer was chosen to be a hyperbolic tangent and for the output layer a linear function. Innamaa and Pursula (2000) found that this combination provided good results.

Neural networks were trained with the Fletcher–Reeves update (Demuth & Beale 1998), which is one of the conjugate gradient algorithms. In those algorithms, the search is performed along conjugate directions. This produces generally faster convergence than the steepest descent direction, which is a common method in basic back propagation algorithms.

If the neural network learns the training data too well, it memorises the data and cannot generalise. This can be avoided by ensuring that the training set is large enough and by setting some stopping criteria for the training process (Demuth & Beale 1998). In this study, several stopping conditions were given. These criteria were the maximum number of training epochs, the minimum values of the gradient and of the mean squared error, and the situation when the mean squared error of the calibration data stopped decreasing. In practice, the training was stopped most of the time because of the last criterion.

For this last mentioned calibration data criterion, the original training data set was divided into three sub-sets: training, calibration and validation set. The division was done sequentially from the chronological database: the first observation for the training set, the second for the calibration set, the third for the training set, the fourth for the validation set and so forth – leading the training set to become double in size compared to the other two sets. The training was performed with the training set. The calibration set was used to track the point at which the model started to learn the peculiarities of the training set and no longer the general features of the modelled phenomenon. The validation set was not used in the training process in any way. It was used, however, after the training to check how the model performed with new data.

2.4. *Selection of input parameters*

The output of the prediction model was designed for a system that updates the display (and the forecast) once a minute. In order to present the expected travel time of the vehicles entering the section at the current time, the output for a certain section was the one-minute average travel time of the vehicles entering this particular section during the following minute. The forecasts were based on the latest information about travel times on different sub-links and about traffic volumes and point speeds at location C and south of location A (Figure 2). By using sub-links, the delay caused by data collection could be kept to a minimum, and information on the development of the traffic situation was more detailed than without them.

The input parameters of the models were selected according to their correlation with the travel time to be predicted (hereinafter: the prediction travel time) and to the mutual correlation of the input parameter candidates. The input parameters of a prediction model need to correlate highly with the prediction travel time to be able to provide sufficient information for the model. However, it may be that different input parameters have high mutual correlation and thereby the additional information that they give to the model may be small.

The correlation between different parameters and the prediction travel time was examined, as well as how these input parameter candidates correlated mutually. All the traffic data available (listed at the end of the chapter entitled Data) was used as input parameter candidates. The one-minute averages were given as time series of varying length.

The time series of the average travel times were chosen to form the basis for the input of the prediction models. Thus, all the time series of one-minute average travel time that had a correlation coefficient of at least 0.20 with the prediction travel time were selected for the input. The criterion for the correlation coefficient between the parameter and the prediction travel time was the same for other parameters as well. However, if two input parameter candidates had a high mutual correlation (coefficient of at least 0.95), the parameter that correlated highest with the prediction travel time was chosen and the other parameter was omitted.

The boundary values 0.20 and 0.95 were selected in order not to limit the input set too much according to the trial and error method, i.e. to keep the input set diversified. The minimum value of 0.20 kept practically all the mean and median travel times in the input set. As the standard deviation may be biased when the sample size is small, only the best correlating deviations were kept. Similarly, the boundary value of 0.95 seemed to find the evident cases but was not too strict.

The correlation coefficients could not answer the question of how long an optimal time series of each input parameter should be. Thus, several lengths (3, 4 and 5 min) were used.

Although the input sets of different models were similar, there were slight differences between the sets. The number of input parameters varied between 18 and 54 for different models, the model for the last sub-link having the smallest number of inputs in both directions (18–23 for the model CD and 22–27 for the model BA). As an example, the contents of the input data set for the model of section AD is presented here.

AD (total 36–54 inputs depending on the length of the time series): three- to five-minute-long time series of one-minute average travel times from all sections; median travel time from the latest 10 or 20 vehicles from sections AD, BD and CD; standard deviation of the travel time from the latest 5 min from sections AC, AD, BC and BD and standard deviation from the latest 20 vehicles from sections AB and CD; three- to five-minute-long time series of one-minute average flows at location C and south of location A, and similar time series of one-minute average point speeds at location C.

The raw data did not include observations for each sub-link for every minute. Therefore, the input data set was made according to the principle that the value of an input parameter was assumed to be invariant until a new observation was obtained. However, the maximum time between updates was set to be 30 min, i.e. if the information on some of the input parameters was older than 30 min, the sample was dropped from the training set. This updating rule was not applied to the output samples of the training set, i.e. all the samples without a new value for the output parameter were excluded from the set.

The boundary value of 30 min was selected to exclude only those periods when one or several detectors had been totally out of order. Thirty minutes might seem like a long time, but neural networks are not very sensitive to incomplete inputs, and several detectors seldom suffer from momentary breaks (e.g. sun shining towards the camera) at the same time. Thus, it was decided that the model had to rely on the correlation between different variables, when updates for certain input variables were not received, in order to gain better reliability of operation for forecasting.

2.5. *Procedure*

The order of superiority of the models depended on which measure of effectiveness was used. If the purpose was to select the statistically best model, the model with the smallest error terms or the best goodness of fit should be chosen. However, from the point of view of the travel time information system, the statistically most accurate model may not always be the best. Road users want information to be sufficiently accurate as frequently as possible. It does not matter if the model makes slight mistakes.

The effectiveness of the models was examined both statistically and from the point of view of the information system. The statistical examination was performed with different error terms: the mean error and relative error, the mean absolute value of error and relative error, and the mean squared error. The first two error terms measure whether the model tended to underestimate or overestimate the travel time, while the last three measured how the errors were distributed around correct values.

Let us assume that the VMS informed the road users that the travel time was going to be the predicted travel time ±10%. However, the absolute minimum travel times shown on the VMS would be those based on the speed limit, i.e. predicted travel times shorter than the travel time based on the speed limit would not have been shown. Therefore, the displayed travel time was not considered erroneous if vehicles travelled faster than the limit, if the VMS showed the minimum travel times allowed. Thus, the travel time information was correct in two ways: the measured travel time was between the upper and lower limits shown on the VMS or it was shorter than the lower limit if the VMS showed the minimum limits.

The erroneous information was divided into two categories: the travel time information was either pessimistic (travel times on the VMS were too long) or optimistic (travel times on the VMS were too short). For the road user, over-optimistic travel time information is worse than over-pessimistic. However, the information should be as exact as possible in order to maintain road users' confidence in the system.

The correctness of the forecasts was investigated for both uncongested and congested traffic. In uncongested flow, the travel time does not vary much and is therefore easy to predict; in addition, the relevance of the information is less important than in congested conditions. The width of the error margin was also evaluated.

3. Results

3.1. *Statistical examination*

Statistical examination of the effectiveness of model was performed with different error terms that compared predicted average values with measured averages. The length of the time series was chosen for each model so that the error terms were minimised. The results showed that the models were very good at prediction over all time periods and that the majority of the forecasts were close to the measured values (Table 2). However, on average the models tended to slightly overestimate the forecasts rather than underestimate them.

False forecasts could be seen as observations of situations where the models predicted travel time to be less than 20 min, but in fact the mea-

Table 2. Different statistical error terms of travel time forecasts on different sections. The predicted average travel times were compared to the measured averages. When a range of the length of time series is given, the length within that range did not affect the results.

Section	Length (km)	Time series (min)	Mean squared error (min^2)	Mean error (min)	Mean abs. value of error (min)	Mean relative error (%)	Mean abs. value of relative error (%)
AD	28.1	3–5	2.5	0.0	1.1	0.6	6.0
AC	17.8	5	1.3	0.0	0.8	0.7	5.9
AB	9.1	4	0.4	0.0	0.4	0.7	6.5
BD	19.0	5	1.0	0.0	0.7	0.5	5.6
BC	8.7	5	0.5	0.0	0.4	0.6	6.1
CD	10.3	5	0.2	0.0	0.4	0.5	5.5
DA	28.1	4	2.1	0.0	1.1	0.6	6.0
DB	19.0	4	1.0	0.0	0.8	0.5	5.9
DC	10.3	3	0.4	0.0	0.5	0.6	6.9
CA	17.8	4–5	0.7	0.0	0.6	0.5	5.6
CB	8.7	5	0.2	0.0	0.3	0.5	5.8
BA	9.1	5	0.2	0.0	0.3	0.6	5.9

sured travel time was 22–25 min (Figure 4). On section DA, all these observations were from situations where the model missed an isolated longer travel time (an isolated peak). Besides missing isolated peaks, the model for section AD was also occasionally delayed from the start of congestion.

Isolated peaks were either local and quickly-resolved incidents or situations where filtering of the raw data had been too coarse. After filtering, there remained observations from which it was hard to tell whether the observation was deviant or not. If the mean value is based on a few observations only, an individual deviating observation may have a significant effect on it.

The first type of error (missing a short isolated peak) was not serious – or could even be considered beneficial – but errors of the second type (being delayed from the start of congestion) should be avoided. However, if the first signs of an unusual situation cannot be measured until after making the forecast, it is hard to come up with an analysis technique that could resolve the problem.

The congestions from which the model for section AD was delayed were typically located between camera stations B and C and therefore the congestion began clearly after the forecast was made (the first signs of congestion could be measured approximately 10–20 min after making the forecast). When random incidents are considered, the delay probably cannot be avoided. However, congestion due to over-demand should be better anticipated based on information about incoming flows.

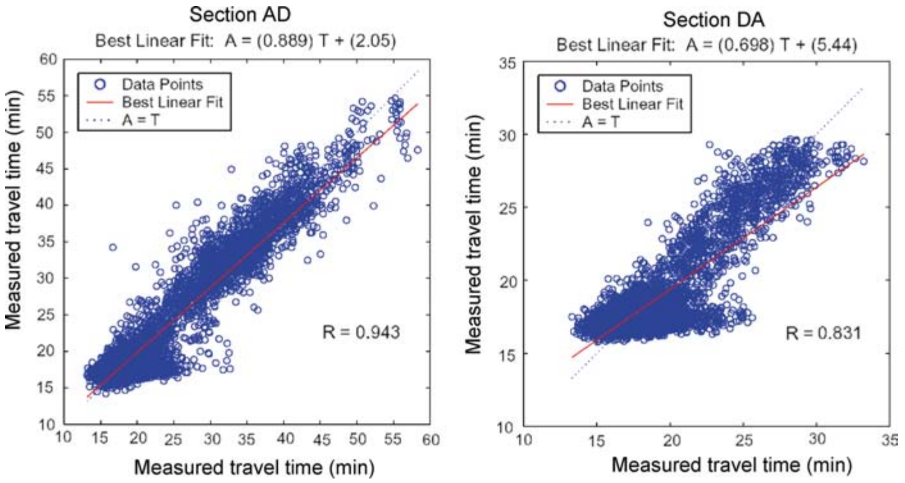

Figure 4. Predicted and measured travel times. The model made for section AD was based on a 5-minute time series and the model for section DA on a 4-minute time series.

Along this particular study section, the problem was that information on flow was available only at location C in the middle of the section, and south of location A on the other side of the nearby city. The traffic flow at location C was not problematic in either direction and therefore, in most cases, could not indicate problems on the section. On this road, the inductive loop detector station outside the section did not indicate traffic-related problems directly either, because the vehicles going to the nearby city were also included in the data and there was no information about vehicles coming from there.

3.2. *Effectiveness from the point of view of the information system*

First, the correctness of the forecasts was investigated for both uncongested and congested traffic all together. The same analysis was also made for samples in which the average travel speed was less than 75% of the free flow speed (congested conditions). These proportions could not be defined for all the sub-links because they had almost no congestion at all during the data collection period (Table 1).

All the models in direction DA gave correct travel time information on average more than 97 percent of the time and in the opposite direction more than 95% of the time (Table 3). If the forecasts for sections AD and DA had been presented on VMSs, the proportion of correct forecasts would have been 71% of the time for section AD and 79% for section DA in congested conditions (Table 4).

A limited examination was performed on the width of the accepted error margin (Figure 5). A decreasing number of forecasts lay within the mar-

Table 3. Effectiveness of prediction models from the point of view of the information system. The numbers are based on both uncongested and congested traffic all together.

Section	Length (km)	Time series (min)	Correct (%)	Too short (%)	Too long (%)
AD	28.1	3–5	97.4	1.5	1.1
AC	17.8	4	96.4	1.9	1.7
AB	9.1	3	96.5	1.8	1.7
BD	19.0	5	97.8	1.2	1.0
BC	8.7	5	95.5	2.1	2.4
CD	10.3	5	99.8	0.2	0.0
DA	28.1	4	98.4	1.1	0.5
DB	19.0	4	97.2	1.8	1.1
DC	10.3	3	97.1	1.5	1.5
CA	17.8	3–5	99.5	0.5	0.0
CB	8.7	4–5	99.5	0.4	0.1
BA	9.1	3–5	100.0	0.0	0.0

gins, as they were set further away from the zero percent error. When the width of the accepted margin was changed from 5.0% to 7.5% the share of correct forecasts improved by 15.4% units on average, whereas the improvement was only 5.8% units when the width was changed from 12.5% to 15.0%.

4. Discussion

This study was designed to investigate, first, the predictability of travel time when the forecast was based on travel time data measured in the field on an interurban two-lane two-way highway. Second, the purpose was to determine whether the forecasts would be accurate enough to implement the model in an actual travel time information service.

The target was to predict the travel time correctly more than 90% of the time. This was on average achieved. However, as the sections had slow, queuing or stopped traffic only 0–9% of the time, the proportion of correct forecasts in congested conditions should be chosen as the main criterion for measuring the effectiveness of the models. Although the target was not achieved in congested conditions, it must be taken into consideration that some of these "false" forecasts were undetected isolated peaks, which cannot be considered serious errors.

The effectiveness of the models was determined as the proportion of forecasts that lay within an accepted error margin. Hence, the width of this margin had a great effect on the absolute values that measured the effectiveness. An accepted error margin wider than 13% would not have been

Table 4. Effectiveness of prediction models from the point of view of the information system in congested conditions only (the average travel speed being less than 75% of the free flow speed).

Section	Length (km)	Time series (min)	Correct (%)	Too short (%)	Too long (%)
AD	28.1	3	70.8*	20.8	8.4
AC	17.8	3	64.2	24.2	11.6
AB	9.1	5	39.3	44.5	16.3
BD	19.0	5	76.5	15.4	8.2
BC	8.7	3	62.8	25.0	12.3
DA	28.1	5	78.8*	18.8	2.5
DB	19.0	4	72.8	20.8	6.4
DC	10.3	3	68.4	19.5	12.1

*The proportion was 32.9% for the section AD and 49.7% for the section DA with non-predictive system.

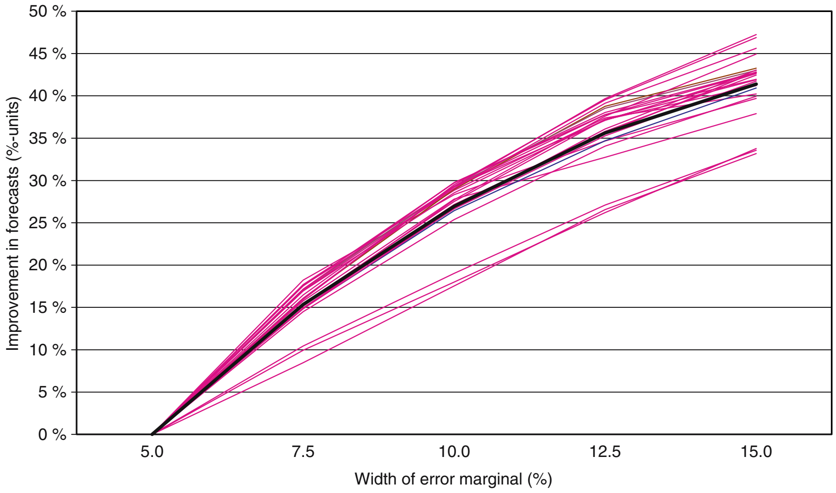

Figure 5. Improvement in forecasts, measured as the proportion of correct forecasts in congested conditions as a function of change in the width of the accepted error margin when comparing with the 5% error margin. The heavy black line is the average of all the models.

justified based on the results of Toppen and Wunderlich (2003). As a great number of forecasts still lay between the 7.5% and 10.0% margins, a margin narrower than 10.0% would not have been practical either. Thus, the width of the 10-percent accepted error margin seemed justified.

The prediction models of Chien and Kuchipudi (2002) and Rilett and Park (2001) had an equal mean absolute value of relative error value as the models presented above, while the models of Park and Rilett (1998) and Park et al. (1999) performed a little worse. However, the degree and type of congestion as well as the quality and the amount of input information varies from one study to another, thus direct comparison of mean errors is inappropriate. One relevant point of comparison, though, is the non-predictive model on the study section. The prediction model described above outperformed the non-predictive model and fulfilled the target in this respect.

In addition to the travel time information, it would be beneficial if the input of the model included flows of all the (at least major) incoming and outgoing flows and the flows at problematic locations (bottlenecks etc.). This way, information would be obtained about flows that are entering the section in question and their effect on travel time could be estimated as a result. Comprehensive flow information could resolve the problem of the model being late from the start of the congestion caused by over-demand.

This kind of measurement system would also better allow the faster detection of an increase in travel time caused by random incidents (i.e. the cause being other than over-demand). In practice, a random incident would show as an increase in the difference in incoming and outgoing flows (incident detection algorithms). Another important aspect in which the flow information could be helpful is the evaluation of whether the measured travel times make sense. If there is no information on flow rates, it is impossible to say whether long travel times are likely due to over-demand, a random incident, or a failure in the measurement system.

It would be desirable if the flow information could be measured directly by the travel time measurement system. This way, overlapping systems could be avoided and costs would be lower. However, this would require an improvement in the current monitoring system such that a larger proportion of vehicles could be detected and all the lanes monitored. The current system was capable of reading on average 40% of all the licence plates on the monitored lane at a single point in good conditions when the cameras were clean. However, the actual sample size of travel time observations of all the potential vehicles was significantly smaller as the conditions were not always good, all the lanes were not monitored, and the vehicles that were detected at one camera point may not have been detected at the next one.

A small sample size is not a problem if we merely want to estimate an average travel time – assuming that an effective filtering of deviating observations can be carried out. An increase in standard deviation could be used as a sign of incidents or changes in the traffic situation. However, the standard deviation of the travel time cannot be estimated from small samples.

Besides getting better information about traffic, a modular approach could also improve the model's ability to predict travel time in different circumstances. A separate neural network trained for a special task only gives better results than one single neural network predicting all kinds of situations. However, if a prediction model is run online, there must anyway be several neural networks to predict the travel time of one single section to ensure the model's ability to work also with imperfect input information. If the model has already been divided into several parts predicting different kinds of traffic situations, this leads to a rather complicated model. Thus the price of improved ability to make forecasts is – as usual – increased complexity.

5. Conclusions

In conclusion, the results of the travel time prediction model were promising, and even a simple prediction model could improve the quality of travel time information especially in congested conditions. The findings suggest that the forecasts could be improved by setting up an adequate monitoring system.

The structure of the monitoring system has effects on the forecasts. Additional camera stations and inductive loop detectors can offer information that improves the model's ability to react to changes in a traffic situation. Consequently, the effects of the section length and the location of different additional measurement stations should be investigated. This information will be important when new, but relatively similar systems are implemented and the structure of the monitoring system is designed.

The online filtering of data should be studied. Periods when there are problems with the monitoring system could be omitted from the data set when working with an offline model. In addition, filtering of the raw data can be done manually if needed. However, when the model is working online this is not possible. The model should work smoothly despite imperfect input information.

Another matter for future research could be the training method. Such a training method should be developed that would weight more heavily errors in congested conditions more than during free flow conditions.

Acknowledgements

The research was funded by the Ministry of Transport and Communications and by the Finnish Road Administration. The study formed part of the Finnish National Research and Development Program on Transport Telematics Infrastructure (TETRA) 1998–2000 and the Finnish Research and Development Programme on ITS Infrastructures and Services (FITS) 2001–2004.

The author would like to thank Risto Kulmala, Juha Luoma and Pirkko Rämä from the Technical Research Centre of Finland, Matti Pursula from Helsinki University of Technology and the reviewers of Transportation for their helpful suggestions on earlier drafts of this article. The neural network computing was done with computers from CSC – Center for Scientific Computing Ltd. in Finland.

References

Al-Deek H (2003) The impact of real-time and predictive traffic information on travelers' behavior in the I-4 corridor. Final report, University of Central Florida. 117 p.

Chen M & Chien S (2001) Dynamic motorway travel time prediction using probe vehicle data: link-based vs. path-based. *Transportation Research Record* 1768, Transportation Data and Information Technology, 157–161.

Chien S & Kuchipudi C (2002) Dynamic travel time prediction with real-time and historical data. Transportation Research Board, Washington DC. Remarks: Paper 02–2548 prepared for presentation at the 81st annual meeting of the Transportation Research Board, Washington, D.C. 26 p.

D'Angelo M, Al-Deek H & Wang M (1999) Travel time prediction for motorway corridors. *Transportation Research Record* 1676, Travel Behavior and Passenger Travel Demand Forecasting, 184–191.

Demuth H & Beale M (1998) Neural networks toolbox for use with Matlab. *User's Guide*, Version 3. The Math Works Inc.: 5-1–5-58.

Finnra (2000) Vt 4 Lahti–Heinola matka-ajan seuranta- ja informaatiojärjestelmän toiminnan arviointi (Main road 4 Lahti–Heinola journey time monitoring and information system functional analysis). *Finnra Reports* 58/2000. Häme District of Finnish National Road Administration, Tampere. 46 p. + app. 61 p.

van Grol H, Danech-Pajouh M, Manfredi S & Whittaker J (1999a) DACCORD: Online travel time prediction. In: Meersman H, van de Voorde E & Winkelmans W (eds.) *World Transport Research*, Selected proceedings of the 8th World Conference on Transport Research, Vol 2: planning, operation, management and control, 455–467.

van Grol R, Lindveld K, Manfredi S & Danech-Pajouh M (1999b) DACCORD: Online travel time estimation/prediction results. *Proceedings of 6th World Congress on Intelligent Transport Systems (ITS)*, held Toronto, Canada, November 8–12, 1999. 12 p.

Haugen T (1996) *Section Data. Possibilities and Experiences*. SINTEF Civil and Environmental Engineering, Transport Engineering, Norway. 16 p.

Innamaa S & Pursula M (2000) Liikennemäärän ja nopeuden lyhyen aikavälin ennustaminen (Short-term prediction of flow and speed). *Finnra Reports* 54/2000. Finnish National Road Administration, Helsinki. 101 p. + app. 3 p.

Kiljunen M & Summala H (1996) Ruuhkaisuuden kokeminen ja liikennetilanne-tiedottaminen. Tienkäyttäjätutkimus kaksikaistaisilla teillä. (Perception of traffic conditions, and traffic information – a road user survey on two lane roads). *Finnra Reports* 25/1996, Finnish National Road Administration, Helsinki. 77 p. + app. 5 p.

Kwon J, Coifman B & Bickel P (2000) Day-to-day travel time trends and travel time prediction from loop detector data. *Transportation Research Record* 1717, Highway and Traffic Safety: Crash Data, Analysis Tools, and Statistical Methods, 120–129.

Lee S, Kim D, Kim J & Cho B (1998) Comparison of models for predicting short-term travel speeds. *Proceedings of 5th World Congress on Intelligent Transport Systems*, Seoul, Korea. 9 p.

Lee Y & Choi C (1998) Development of a link travel time prediction algorithm for urban expressway. *Proceedings of 5th World Congress on Intelligent Transport Systems*, Seoul, Korea. 8 p.

Lindveld C, Thijs R, Bovy P & van der Zijpp N (2000) Evaluation of online travel time estimators and predictors. *Transportation Research Record* 1719, 45–53.

van Lint H (2003) Confidence intervals for real-time motorway travel time prediction. IEEE Conference on Intelligent Transportation Systems, Shanghai, China. 6 p.

van Lint J, Hoogendoorn S & van Zuylen H (2002) Motorway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural networks. *Transportation Research Record* 1811, 30–39.

van Lint J, Hoogendoorn S & van Zuylen H (2003) Toward a robust framework for motorway travel time prediction: experiments with simple imputation and state-space neural networks. Transportation Research Board 82nd Annual Meeting, Compendium of papers CD-ROM, Washington D.C. 11 p.

Luoma S (1998) Liikenteen sujuvuus ja sen mittaaminen (Transport system efficiency and its estimation). *Finnra Reports* 21/1998, Finnish National Road Administration, Helsinki. 101 p. + app. 27 p.

Matsui H & Fujita M (1998) Travel time prediction for motorway traffic information by neural network driven fuzzy reasoning. In: *Neural Networks in transportation Applications*, 355–364.

McFadden J, Yang W.T & Durrans S (2001) Application of artificial neural networks to predict speeds on two-lane rural highways. *Transportation Research Record* 1751, Geometric Design and the Effects on Traffic Operations (2001), 9–17.

Nanthawichit C, Nakatsuji T & Suzuki H (2003) Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction on a motorway. Transportation Research Board 82nd Annual Meeting, Compendium of papers CD-ROM, Washington D.C. 16 p.

Ohba Y, Ueno H & Kuwahara M (2000) Travel time prediction method for expressway using toll collection system data. *Proceedings of the 7th World Congress on Intelligent Systems*, held in Turin, Italy. 8 p.

Park D & Rilett L (1998) Forecasting multiple-period motorway link travel times using modular neural networks. *Transportation Research Record* 1617, Land Use and Transportation Planning and Programming Applications, 163–170.

Park D & Rilett L (1999) Forecasting motorway link travel times with a feedforward neural network. *Computer-Aided Civil and Infrastructure Engineering* 1999/09, 14(5), 357–367.

Park D, Rilett L & Han G (1999) Spectral basis neural networks for real-time travel time forecasting. *Journal of Transportation Engineering* November/December 1999, 515–523.

Paterson D & Rose G (1999) Dynamic travel time estimation on instrumented motorways. *Proceedings of the 6th World Congress on Intelligent Transport Systems (ITS)*, held in Toronto. 11 p.

Rilett L & Park D (2001) Direct forecasting of motorway corridor travel times using spectral basis neural networks. *Transportation Research Record* 1752, Travel Patterns and Behavior; Effects of Communications Technology, 140–147.

Saito M & Watanabe T (1995) Prediction and dissemination systems for travel time utilizing vehicle detectors. Steps Forward, Intelligent Transport Systems World Congress, Yokohama, Japan. Proceedings, Vol. I, 106–111.

Shao C, Gu Y & Zhang K (2002) A study on dynamic travel time forecast with neural networks. In: Wang K, Xiao G, Nie L & Yang H (Eds.) *Traffic and Transportation Studies*. Proceedings of ICTTS 2002, Vol. 1, 716–721.

Smith B & Demetsky M (1994) Short-term traffic flow prediction: neural network approach. *Transportation Research Record* 1453, 98–104.

Smith B & Demetsky M (1997) Traffic flow forecasting: comparison of modeling approaches. *Journal of Transportation Engineering*, Vol. 123, No. 4, July/August 1997: 261–266.

Suzuki H, Nakatsuji T, Tanaboriboon Y & Takahashi K (2000) Dynamic estimation of origin-destination travel time and flow on a long motorway corridor. *Transportation Research Record* 1739, Evaluating Intelligent Transportation Systems, Advanced Traveler Information Systems, and Other Artificial Intelligence Applications, 67–75.

Toppen A & Wunderlich K (2003) Travel time data collection for measurement of advanced traveler information systems accuracy. Federal Highway Administration, Project No. 0900610-D1. 20 p.

Yasui K, Ikenoue K & Takeuchi H (1995) Use of AVI information linked up with detector output in travel time prediction and O–D flow estimation. Steps Forward, Intelligent Transport Systems World Congress, Yokohama, Japan. Proceedings, Vol. I, 94–99.

You J & Kim T (2000) Development and evaluation of a hybrid travel time forecasting model. *Transportation Research*, Part C: Emerging Technologies 8, 231–256.

Zhang X & Rice J (2003) Short-term travel time prediction. *Transport Research* Part C 11, 187–210.

About the author

Satu Innamaa obtained her Master of Science degree in Civil Engineering in 1997 at Helsinki University of Technology. After graduation, Satu Innamaa worked at the University in the Laboratory of Transportation Engineering as a Research Scientist until 2001. In 2002, she became a Research Scientist at VTT Technical Research Centre of Finland.