ELSEVIER

International Journal of Forecasting 13 (1997) 51–61

international journal of forecasting

Tracking and predicting a network traffic process

Joe Whittaker[a,*], Simon Garside[a], Karel Lindveld[b]

[a]Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

[b]Hague Consulting Group, Den Haag, Netherlands

Abstract

This article deals with the problem of real-time modelling and prediction of motorway traffic. Conditional independence relationships and ideas of Bayesian forecasting are proposed leading to the employment of dynamic state-space models, with optimal state estimation coming from the Kalman filter. Models, based on classical differential equations, which incorporate representations of the network topology are derived and are implemented in a state-space framework. The model is applied to several road networks in The Netherlands from which encouraging preliminary results are obtained.

Keywords: Motorway networks; Traffic dynamics; State-space model; Kalman filter; Independence graph

1. Introduction

Increasing traffic volumes, especially on motorways, are putting pressure on road systems throughout all industrial economies. Expected growth in demand is liable to cause drastic increases in travel delays, reductions in safety standards, increased energy consumption and pollutant emissions. In order to reduce these undesirable effects traffic management measures are required, and the challenge is to build models for traffic data processing that enable efficient management.

The electronics revolution has led to advances in transportation technology. In particular induction loops implanted in the road surface taking real-time data from the traffic process are now being wired to local and distant processors. This enables on-line monitoring and management of whole networks, never previously possible, and naturally require automatic methods of information processing.

This article presents a dynamic state-space model. Dynamic models transform on-line roadside measurements of traffic flow into explicit assessments of the current and future state of an inter-urban network; and thus provide transport management with a tool for monitoring, prediction and potentially control. The stochastic traffic network process and the measurement process are jointly modelled within a common framework.

Bayesian methodology naturally provides online estimates of the current state of the network and predictions of its future trajectory. The apparently large quantity of data (both in time and space) is simplified by investigation of the

* Corresponding author. Tel.: +44 1524 593956; fax: +44 1524 592681.

PII S0169-2070(96)00700-5

conditional independence structure of the process. This structure in part reflects the given spatial topology of the network, and is incorporated in the state transition equations. The transition and observation equations lead to an associated Kalman filter which provides the relative discounting of past data in proportion to the underlying variability of the traffic process. Prediction ahead is an instance of Bayesian forecasting. Basic ideas of traffic flow, such as conservation of vehicles and the so-called fundamental diagram, are built into the state transition equations. However, some complications ensue because of non-linearities in speed–flow relationships and resulting non-linearity of the equations.

The objective of the article is to sketch out the skeleton of the traffic model, indicate why the state-space format is so appropriate, and to elucidate the conditional independence statements inherent in its structure. In Section 2 the application is discussed, the networks, the measurement process and some simple ideas of traffic dynamics. Section 3 lays out the state-space model, emphasising the conditional independence aspects, and summarising the filtering and prediction procedures. In Section 4 this model is tailored to motorway networks, by appropriate choices of the states, transitions and observation equations. Some preliminary work on the evaluation of this state traffic model is reported in Section 5.

2. The application

A typical motorway network is represented in Fig. 1.

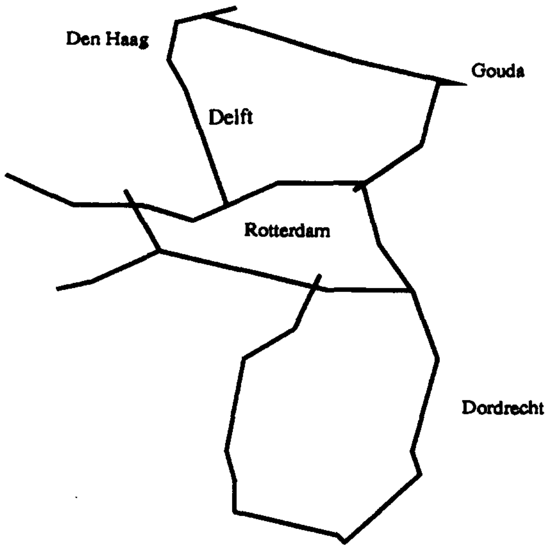


Fig. 1. The Rotterdam pilot network.

The central loop is the motorway ring around Rotterdam. There are recurrent bottlenecks at some of the intersections. The triple loop allows an element of route choice along monitored motorway. There are about 50 on–off-ramps onto the otherwise closed system. So far there are around 500 monitoring stations installed at intervals of between 0.5 and 10 km. This leads to about 500 sections of motorway, or links, between monitoring stations. The information from the stations is fed to the Rotterdam central control centre.

Fig. 2 shows the daily flow and velocity from a monitoring station on the A13 to Delft from a day in 1993. The daily cycle is prominent, stochastic variation is noticeable and there is evidence of serial correlation. Fig. 3 plots the stacked output of 16 adjacent monitoring stations for a $2\frac{1}{2}$-hour period during the morning rush-hour. The furthest upstream station is represented by the lowest line with subsequent stations being indexed by the y-axis. There is clear evidence of feed forward interaction of flows (but not velocities) in free flow conditions, as seen by the lower left to upper right waves (e.g. commencing on Station 1 at 20 min), and of feed-back interaction of both flows and velocities in congested flow conditions, evidenced by the upper left to lower right (e.g. commencing on Station 12 at 25 min).

2.1. Monitoring station measurement process

The measurement process is based on a pair of embedded conductance loops

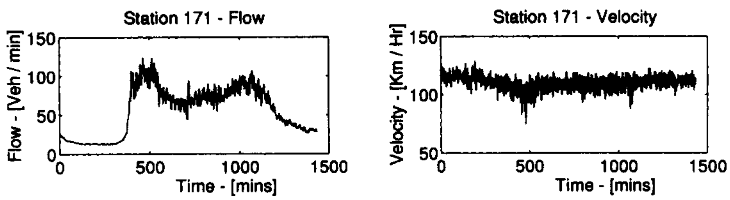
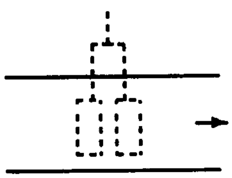


Fig. 2. Daily flow and velocity, A13 Delft, 1993.



in each lane of the motorway. As a vehicle passes over the loops, in principle four times are recorded and the vehicle passage, the vehicle speed and type can be recorded. The information from the loop pair is processed at the station site with that from the other lanes to give just the flow and the velocity by 1-min aggregates or averages. In fact some of the older stations give exponentially weighted moving averages of the measures. The newer stations, aggregate but also report % standard deviations and other information such as occupancy and vehicle composition. The pre-processed data are then transmitted to the traffic control centre. Certain difficulties can occur: one is synchronisation of the measures from the different stations, another is missing observations from dead or malfunctioning loops.

2.2. Elements of traffic dynamics

Relatively little specialised knowledge is required to understand this application. The essential traffic variables (q, r, v) are flow, density and velocity. They vary in space and time (s, t) and, under stationary conditions, satisfy the identity $q = rv$. So for instance on a three-lane motorway, if $v = 120\,\mathrm{km\,hr^{-1}}$, $r = 50$ vehicles km^{-1} then $q = 100$ vehicles min^{-1} and the distance headway between vehicles in the same lane is 60 m with time headway 1.8 sec.

As vehicles are conserved in a closed system one essential part of the dynamics of the process is an accounting equation that keeps track of the number of vehicles on the different links of the network. A second component is the 'fundamental traffic diagram' which describes the way in which traffic congestion leads to a reduction in

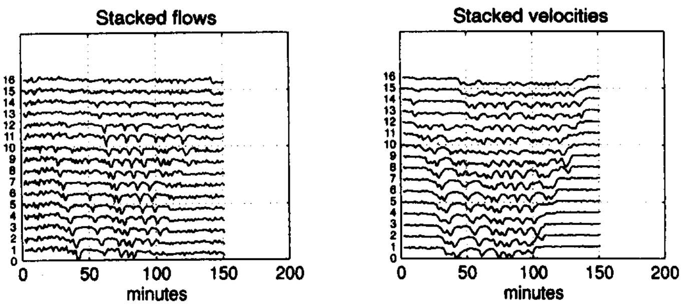


Fig. 3. Spatial interaction: motorway flows and velocities, A2 Amsterdam, 1990.

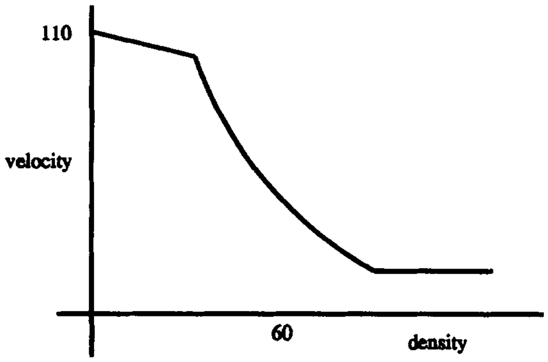


Fig. 4. A generic velocity–density function.

average velocity by a velocity–density function $v = V(r)$ (Fig. 4).

The free-flow capacity of the link is reflected in the level and the slope of the low density linear part of the function. Congestion occurs when a critical density is reached. A striking feature of empirical versions of this diagram is the enormous variability in velocity in congested periods relative to that in free flow periods.

3. The state-space model

The Kalman filter (Kalman, 1960) is an important algorithmic device in the theory of signal detection and automatic control (see, for example, Ljung and Söderstrom, 1983 or Young, 1984). It has many fields of application, including statistics (see West and Harrison, 1989), and economic time series modelling (see Harvey, 1989).

Underlying the filter is the state-space model, a multivariate stochastic process, for which the filter is an optimal Bayes solution to a prediction problem. The state-space model supposes that at each discrete time point t, a (vector-valued) observation y_t, is related to x_t, the states, through an observation equation. The state process is an unobserved first order Markov process and is specified by a transition equation. On the basis of past observations $y^t = (y_1, y_2, \ldots, y_t)$ and the current observation y_{t+1} it is required to best predict x_{t+1}. It turns out that the filter depends on the observations only through the previous estimate of the state x_t and the current prediction error, and so allows a recursive computation.

The basis of the model are the transition and observation equations

$$\begin{aligned} x_{t+1} &= A_t x_t + Bu_t + q_t, & \mathrm{var}(q_t) &= Q \\ y_t &= Cx_t \quad + r_t, & \mathrm{var}(r_t) &= R \end{aligned} \tag{1}$$

The structure of these equations is held in the conditional independence graph (Fig. 5) in which it is the missing edges that indicate independences (see Whittaker, 1990, for an introduction to conditional independence and graphical modelling; see Normand and Tritchler, 1989; Dempster, 1990; Smith, 1990, for the relationship between graphical modelling and Kalman filtering). The 'past' represents past values of the variables. That there is no edge to x_{t+1} from the 'past' or to y_t indicates that enough information is held in the current state x_t to render past information redundant in predicting its evolution. That there is no edge from the 'past' to u_t indicates that the exogenous variables are not determined by past values of the system variables; if there is feedback, as for example, in control application then this diagram requires modification.

3.1. Filtering and prediction

There are many ways to write the updating formula for estimating the current state of the process. Here we choose one expressed in terms of the conditional mean m_t and conditional variance p_t of the value of the state vector x_t at

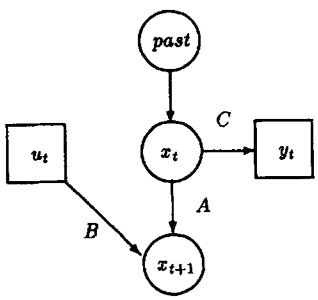


Fig. 5. Independence graph of a general state space model.

time t given all past and present observations y^t and u^t. We use the fairly standard notation that the past data vector is $y^t = (y_1, y_2, \ldots, y_t)$; $y^0 = (\cdot)$. Put $m_t = E(x_t \mid y^t, u^t)$ and $p_t = \mathrm{var}(x_t \mid y^t, u^t)$. The recursion that determines the Kalman filter is

- initialise: choose m_0, p_0
- recurse:

$$K_{t+1} = (A_t p_t A_t' + Q)C'[C(A_t p_t A_t' + Q)C' + R]^{-1} \tag{2}$$

$$m_{t+1} = A_t m_t + Bu_t + K_{t+1}(y_{t+1} - C(A_t m_t + Bu_t)) \tag{3}$$

$$p_{t+1} = (I - K_{t+1}C)(A_t p_t A_t' + Q) \tag{4}$$

The Kalman gain matrix is K_t.

The recursion for m_t in (3) can be re-expressed in three steps: the first is to calculate the best one-step ahead predictor of x_{t+1} from the best predictor of x_t alone, alone, $m_t^1 = E(x_{t+1} \mid y^t, u^t)$, given by $m_t^1 = A_t m_t + Bu_t$. The second is to calculate the prediction error of y_{t+1} from this best predictor of x_{t+1}, $y_{t+1} - Cm_t^1$; and the third is to error correct the one-step ahead predictor using the updated Kalman gain and so obtain the updated filtered estimate of the new state. If A_t depends on the states this is an instance of an extended Kalman filter, whereas A_t constant leads to a standard filter.

The k-step ahead predictor at time t is obtained by iterating this sequence through

$$m_t^k = A_{t+k-1} m_t^{k-1} + Bu_{t+k-1} \tag{5}$$

which depends on knowledge or forecasts of $u_{t+1}, u_{t+2}, \ldots, u_{t+k-1}$ and of $A_{t+1}, \ldots, A_{t+k-1}$.

4. State-space models for network traffic applications

Various salient features of a state-space model make it appropriate for traffic network applications: it is a multivariate stochastic process that is both inherently random and contains explicit measurement error, which, above all, has a simple flexible structure to be tailored to the application. The structure allows:

1. Variables to exist at different levels (observed, latent, derived), so for instance link flow is observed, link density is latent, link travel time is derived, while on-ramp flow is exogenous.
2. Physical understanding: traffic dynamics is incorporated through the transition equation, so that
   - density is predictable through a dynamic accounting equation,
   - derived variables such as velocity are consequent on density via the fundamental traffic diagram, and
   - the level of flow is dependent on the state of present and downstream links expressed as a form of spatial interaction.
3. The temporal and spatial Markov structure of the physical process to be mapped onto the Markov structure of the state-space model; this both allows recursive processing and allows decomposition into subnetworks.
4. Flexible observation equations to allow for the vagaries of the measurement process.

4.1. Choosing the states x_t

The model under development consists of the following components, with illustrative dimensions taken from the Rotterdam application:

- a latent state vector x_t of dimension approximately (1000×1) that holds two variables per link, traffic densities and flows, and is ordered by corresponding measurement station, i.e.

$$x_{s,t} = \begin{pmatrix} r_{s,t} \\ q_{s,t} \end{pmatrix}, \quad x_t = \begin{pmatrix} \vdots \\ x_{s-1,t} \\ x_{s,t} \\ x_{s+1,t} \\ \vdots \end{pmatrix}$$

- an observation vector y_t approximately (500×1) that holds the information from the monitoring stations ordered by station,
- a vector of exogenous input u_t approximately

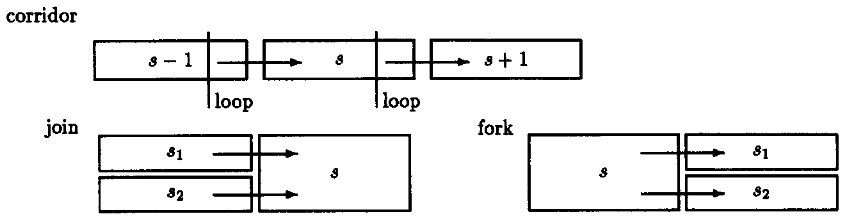


Fig. 6. Block components of the traffic model.

(50 × 1) of flows from monitoring stations on the on-ramp entrances, and

- variables that specify the behaviour of the system, which include measures of link length, link capacity and fork turning fractions.

The lack of an edge between y_t and the 'past' in the independence graph of the general state-space model above, Fig. 5, indicates that given the state of the network x_t any further variation is due to independent measurement error. Importantly there is no arrow from the past or from x_t to the input, the on-ramp flows u_t. This assumes that traffic demand is not dependent on past state of the network (and of course this may be unrealistic and require further modelling).

4.2. Choosing the transition equations A_t

The transition equations and the transition matrix A_t carries both the network topology (which is time invariant) and the traffic dynamics.

4.2.1. Network topology

The network is a set of stations and links, the section of motorway between two adjacent stations. Motorway networks are relatively simple because there are at most two upstream and two downstream links to any station. Consequently the sections of the motorway fit together in the following way (Fig. 6).

A block is a combination of links: either a single section, or the two sections that form a join, or the two sections that form a fork. An on-ramp is a join block, an off-ramp is a fork block; intersections are combinations of forks and joins. Roundabouts and cross-roads are prohibited. The whole motorway network may be decomposed into these three types of blocks. For example, a simple network might have the following adjacency graphy composed from these three building blocks (Fig. 7).

If the state-space vector holds just the density and the exit flow(s) for each block then the section and join blocks have dimension 2, while the fork blocks have dimension 3 as there are two exiting flows. The block adjacency graph in part determines the spatial structure of the conditional independence graph of the state space blocks.

4.2.2. Transitions: time varying traffic dynamics

The transition matrix $A_t = A(x_t)$ depends on traffic dynamics. Classically these follow differential equations analogous to those of fluid dynamics (for a review see the articles in

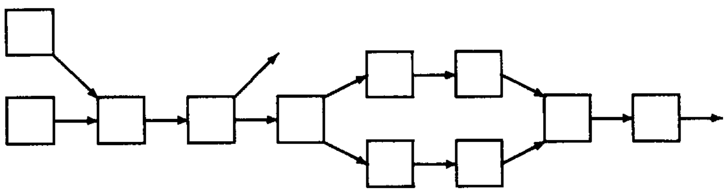


Fig. 7. Adjacency graph for an illustrative network.

Papageorgiu, 1991). There are three types of equations which comprise the model:

- a balance or accounting equation,
- a flow equation based on the identity $q = vr$, and
- an equation determining velocity from the fundamental traffic diagram $v = V(r)$ in Fig. 4 above.

There are three types of block transitions to consider section to section, sections to join and fork to sections. The independence graph of a section to section transition is constructed by replicating the graph of Fig. 8 over s. There are no other directed arrows into block s.

The detailed specifications that lies behind these arrows is: the accounting (or mass-balance or conservation) equation for the density of the form

$$r_{s,t+1} = r_{s,t} + k_s^{-1}(q_{s-1,t} - q_{s,t}) \tag{6}$$

where k_s is the area (lanes × length) of link s, which gives the edge between nodes $(s-1, t)$ and $(s, t+1)$ (feed forward spatial interaction) and part of the edge between nodes (s, t) and $(s, t+1)$; together with one of several possible flow equations, of which the simplest is

$$q_{s,t+1} = \min(v_{s,t}, v_{s+1,t}) r_{s,t} \tag{7}$$

and also gives part of the edge between nodes (s, t) and $(s, t+1)$ and the edge between nodes $(s+1, t)$ and $(s, t+1)$ (the feedback interaction). The flow equation incorporates velocities $v_{s,t}$, $v_{s+1,t}$, which are computed from the fundamental diagram above using $r_{s,t}$, respectively $r_{s+1,t}$, and so entail a non-linear function of the state-space reflecting the physical reality of traffic congestion.

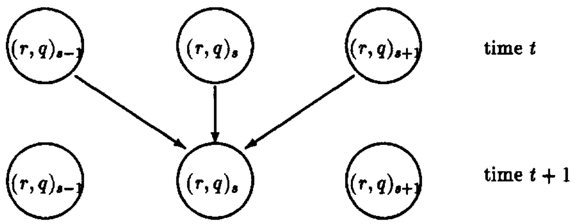


Fig. 8. Independence graph of state transitions.

Similar equations hold for sections to joins; fork to sections are slightly more complicated and depend on the turning fraction between the two downstream choices. The unknown parameters require estimation, which may be performed on-line from current measurements or off-line from an historical data base.

The observation equations for measured flow are straightforward, and graphically result in a single directed arrow from each traffic block to an independent observation y_s that holds the observed flow from the monitoring station exiting the block. Hence the C matrix is block diagonal. There are two observed flows in a fork block. The observation equation that incorporates measured velocity is slightly more involved but the graph is the same.

4.3. Sparse multiplication and subnetworks

For a network the size of the Rotterdam application the state-space is of order 1000 × 1000, and with 500 monitoring stations, computation of the Kalman gain requires inversion of matrices of the order 500 × 500. Evaluation, storage, multiplication, let alone inversion, are neither sufficiently accurate nor feasible in real time. This has led to on-going investigations by the authors into a decomposition of the full network into overlapping subnetworks which is viable due to conditional independence structure entailed by the modelling of near-neighbour interactions.

5. Model performance

An indication of the model's performance is shown on traffic measurement data from the A12 between the cities of Zoetermeer and The Hague in The Netherlands. This is a small network consisting of two carriageways which have each been decomposed into five blocks of lengths between 2.25 km and 4.5 km. All but one block is a fork block as all off-ramps exiting the network are monitored; hence there are two out-flows per block. This gives a total of 29 state variables within the state vector: ten occupancies (surrogate densities), ten main-carriageway out-flows and nine off-flows.

The model is run off-line for a specified time period and all predictions and estimated states are stored. Mean absolute errors for every state variable are then calculated and compared to the equivalent errors obtained when the fitted/predicted variables are replaced by the current measurement, the so-called naive predictor.

Some slight alterations to the state traffic model (STM) are necessary. Specifically, as we have measurements of occupancy available with this application, density is replaced by occupancy as a state variable. The fundamental structure of the model and the relationships between the variables are relatively unchanged as occupancy and density are linearly related. The impact on the model equations is to simply require extra parameters multiplying both the k_s terms and the $\min(v_{s,t}, v_{s+1,t})$ terms.

Approximate values of these extra parameters are known a priori; however calibration of the model against real data allows the validity of the model to be checked, as well as providing empirical estimates of the model parameters.

5.1. Simple parameter estimation

ARX (autoregressive exogenous input) routines are used for parameter estimation as, in the relatively simple case of the road traffic context, a simple relationship between an ARX format and a state-space format exists. A general form of an ARX model is

$$y_t + \Phi_1 y_{t-1} + \cdots + \Phi_{na} y_{t-na} = \Phi_{nk} u_{t-nk} + \cdots + \Phi_{nk+nb-1} u_{t-nk+1-nb} + e_t \quad (8)$$

otherwise expressed as

$$\Phi(q) y_t = \Theta(q) u_{t-nk} + e_t \quad (9)$$

where both y and u may be multivariate. By setting y_t to x_t and each of time lags na, nb and nk to one we are able to directly obtain estimates of the A_t and B matrices (see Ljung and Söderstrom, 1983, for further details).

Though it is accepted that y_t is not equal to x_t it is believed that the measurement noise is low and that initial parameter estimates obtained this way are sufficiently accurate.

To obtain parameter estimates the ARX routines are applied to data in a 60-min window, and re-estimated by sliding the window through the whole day. This produces a set of time-varying estimates for each parameter with which to detect any violations of the model's assumption of constant parameters. If none are found, parameters are averaged to give a single set of non-time-varying estimates. The ARX routines produce encouraging results with parameter estimates being very similar in magnitude to their theoretical values.

5.2. Prediction results

Various periods are used for testing and several different parameterisations of the noise variance matrices Q and R are tried. The predictions obtained from the model compare favourably with those of the naive predictor.

Fig. 9 shows the five-step predictions for one of the state variables. It can be seen from the lower plot that there is strong correlation between predicted and observed variables, particularly at lower magnitudes. The upper plot shows that the prediction follows the observed data closely most of the time. These results are typical for other stations, other days and other variables.

5.3. Comparison of model and naive predictors

Fig. 10 details the mean absolute errors from the STM with various Q/R matrix settings in the 11 am to 1 pm time period with a prediction horizon of two time steps.

The plots show these mean absolute errors and it should be noted that the line consisting of dots shows the errors from the naive predictor which are, in general, worse than those from any of the models for this time period and this prediction horizon. It should also be noted that none of the various Q/R parameterisations was consistently

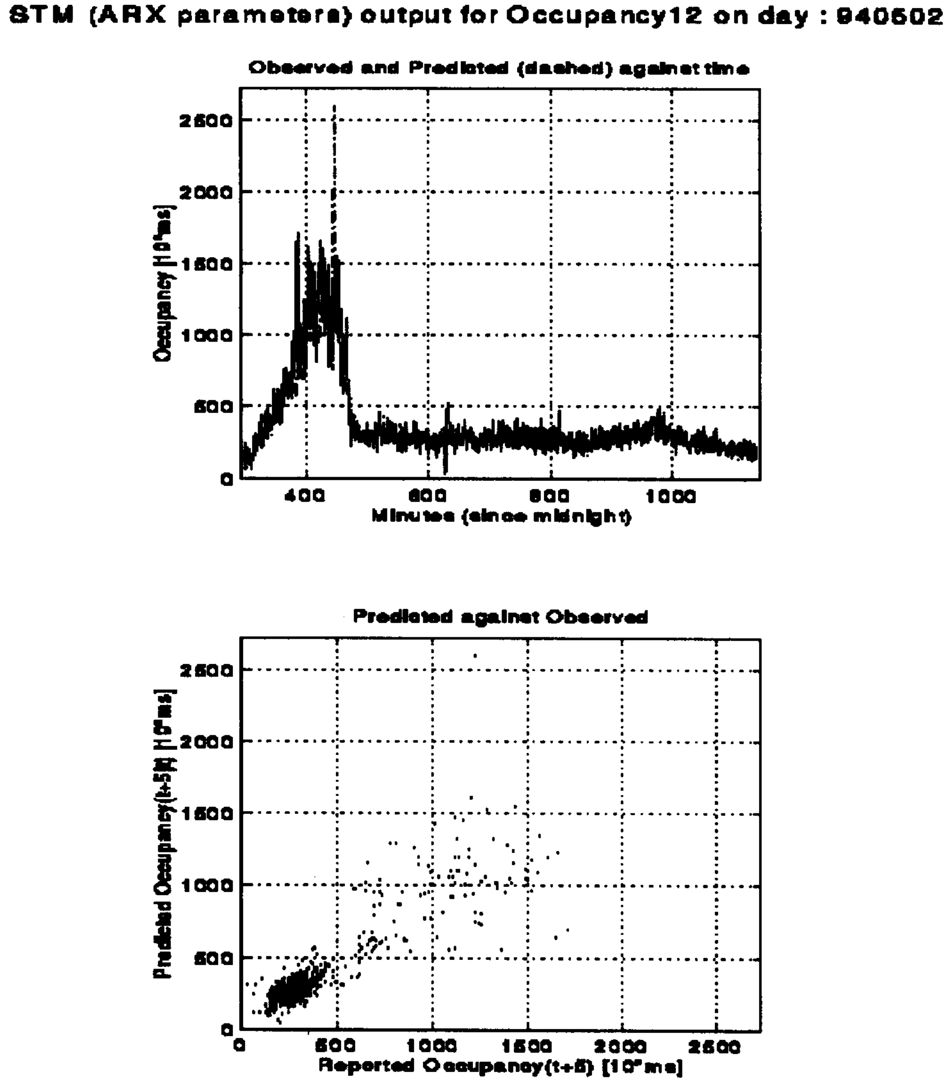


Fig. 9. Five-step ahead predictions for variable occupancy 12.

the best for every time period and every prediction horizon.

Other points to be noted from these preliminary results are: that the model performs best in the periods of free-flow; predictions of off out-flow are much better than the naive for just about all time intervals and prediction horizons; the quality of the occupancy variable predictions varies between time interval and prediction horizon but are in some cases worse than equivalent predictions from the naive predictor; there is evidence that only in free-flow the prediction of main out-flow is better, with other time-periods showing less conclusive results.

6. Concluding remarks

We have outlined above a partial description of the monitoring and prediction system that is under development, concentrating on the underlying graphical representation of the state-space

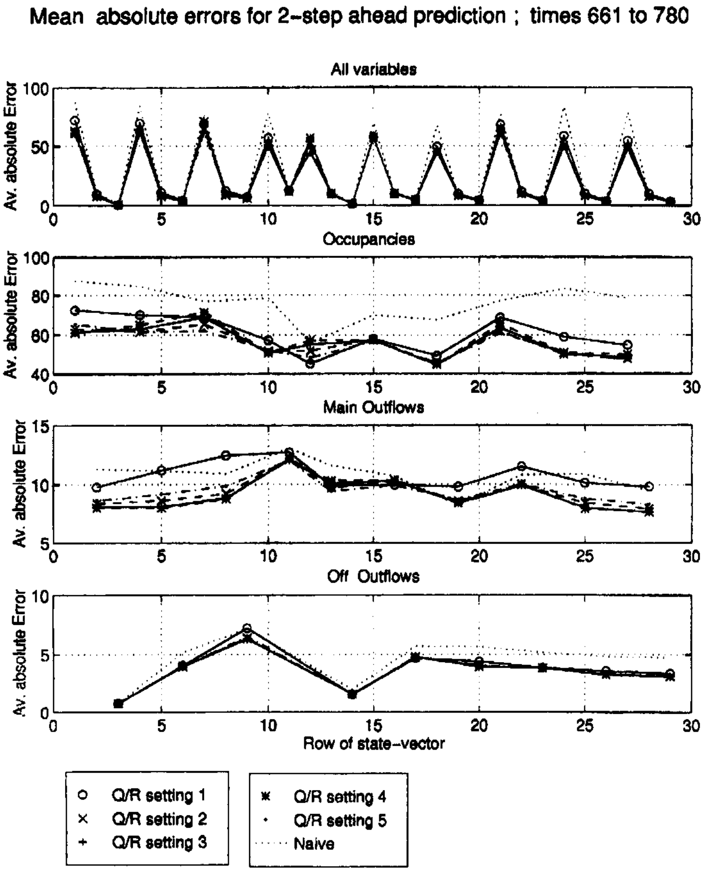


Fig. 10. Two-step ahead errors for all variables over period 11 am to 1 pm.

model. Many issues are not dealt with here, but we point out that the independence graph is suggesting decomposition techniques for simplifying networks, especially cycles in which path choice plays a part and which correspond to non-decomposable graphical models.

Acknowledgements

The article is based on project work underway at the University of Lancaster under the EC/DRIVE-II initiative in cooperation with partners from the Hague Consulting Group, DVK (from the Dutch Ministry of Transport), CSST and ELASIS (part of the FIAT group), and from the Universities of Brussels and Napoli. The overall aims of the project entitled Dynamic Traffic Modelling (DYNA) are to develop practical forecasting tools and to validate these in a full-scale pilot project on an inter-urban motorway network centred on Rotterdam. Other methodology investigated in this project includes demand modelling, dynamic OD-estimation and prediction, and dynamic assignment. The emphasis of this work is to provide practical working methods based on existing technology, but preliminary pilot studies have naturally raised some interesting statistical and theoretical questions.

References

Dempster, A.P., 1990, Construction and local computation aspects of network belief functions, in: R.M. Oliver and

J.Q. Smith, eds., *Influence Diagrams, Belief Nets and Decision Analysis* (Wiley, Chichester).

Harvey, A.C., 1989, *Forecasting, Structural Time Series Models and the Kalman Filter* (CUP, Cambridge).

Kalman, R.E., 1960, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering, Transactions ASME, Series D*, 82, 35–45.

Ljung, L. and T. Söderstrom, 1983, *Theory and Practice of Recursive Identification* (MIT Press, Cambridge, MA).

Normand, S. and D. Tritchler, 1989, Kalman filtering in a Bayesian network, in: C. Dempster (ed.), *Proc. of Stat'l. Computing Sect.* (American Statistical Association, Alexandria, VA), 259–264.

Papageorgiu, M., 1991, *Concise Encyclopedia of Traffic and Transportation Systems* (Pergamon Press, Oxford).

Smith, J.Q., 1990, Statistical principles on graphs, in: R.M. Oliver and J.Q. Smith, eds., *Influence Diagrams, Belief Nets and Decision Analysis* (Wiley, Chichester).

West, M. and P.J. Harrison, 1989, *Bayesian Forecasting and the Dynamic Linear Model* (Springer-Verlag).

Whittaker, J., 1990, *Graphical Models in Applied Multivariate Statistics* (Wiley, Chichester).

Young, P., 1984, *Recursive Estimation and Time-Series Analysis* (Berlin, Springer-Verlag).

Biographies: Joe WHITTAKER is a senior lecturer in the Department of Mathematics and Statistics, Lancaster University. A graduate of the London School of Economics, he has held positions at Colorado State University and the University of Chicago. His research interests are in graphical models, applied statistics and traffic prediction. He worked on several projects involving prediction including DYNA which was part of the European Community's DRIVE-II initiative for development of intelligent transportation systems, and its successor DACCORD.

Simon GARSIDE is a Ph.D. student in the Department of Mathematics and Statistics, Lancaster University. He received a B.Sc. in Mathematics from the University of Durham in 1991 and in 1992 gained a Diploma in Mathematical Statistics at the University of Cambridge. His research interests are in the field of Applied Statistics, in particular the modelling and forecasting of road networks through State Space methodology.

Karel LINDVELD is a consultant with Hague Consulting Group where he has been employed since 1989. He received a M.Sc. in applied Mathematics and Information Science from the University of Leiden in 1988. Within Hague Consulting Group he has worked on the application of discrete-choice models on transportation modelling, and on short-term traffic forecasting. He has worked as technical secretary within the DRIVE-II project DYNA.