Estimation and Prediction of Time-Dependent Origin-Destination Flows

by

Kalidas Ashok

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Transportation Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1996

Author .....
Department of Civil and Environmental Engineering
August 9, 1996

Certified by .....
Moshe E. Ben-Akiva
Professor, Department of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Joseph M. Sussman
Chairman, Departmental Committee on Graduate Students

Estimation and Prediction of Time-Dependent Origin-Destination Flows

by

Kalidas Ashok

Submitted to the Department of Civil and Environmental Engineering
on August 9, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation Systems

Abstract

In this thesis, we present a comprehensive framework to estimate and predict time-dependent Origin-Destination (O-D) flows. The key feature of this framework is its ability to handle different types of information ("measurements") with different error characteristics and from different sources in a consistent and unified manner. The framework is used to address two types of problems – the *offline* estimation problem and the *real-time* estimation and prediction problem.

For the offline estimation problem, we enhance least squares based procedures developed by other researchers. To the real-time estimation and prediction problem, we apply state-space modeling techniques to obtain recursive estimation algorithms. Main features of our models are: (a) use of *deviations* of O-D flows from historical averages as unknown variables (b) modeling of originating trips and destination fractions separately to improve prediction efficiency and (c) introduction to the notion of a stochastic assignment matrix for mapping O-D flows to link counts.

The suite of models developed in this thesis is evaluated rigorously using actual traffic data from three different sources. Empirical results are promising and indicate the robustness of the proposed framework.

Thesis Supervisor: Moshe E. Ben-Akiva
Title: Professor, Department of Civil and Environmental Engineering

Acknowledgments

I would like to thank the following individuals for their involvement at various stages of this research:

My supervisor Professor Moshe Ben-Akiva for being a constant source of inspiration. I have learnt much from him, both academic and otherwise, and for this I am indebted.

Professor Haris Koutsopoulos for his continuous advise, encouragement and friendship, Professor Joseph Sussman for helping me place this research in broader perspective, and Professor Ennio Cascetta for his many insights that helped enhance the quality of this work.

The MIT ITS Program staff including Michel Bierlaire, Tony Hotz, and Rabi Mishalani for their valuable suggestions. Taka Morikawa and Megan Khoshyaran, for their helpful comments.

John Judge at the Massachusetts Turnpike Authority, Hisham Noeimi and Karl Petty at the University of California, Berkeley, and Nanne van der Zijpp at Delft University, Netherlands, for providing data used in this research.

The UPS Foundation and the MIT ITS program for their generous financial support through fellowships and the CA/T and DTA research projects.

All my friends and fellow students, too numerous to list, who made my MIT experience pleasant, as well as administrative staff at the Center for Transportation Studies, specially Lisa and Julie, for their help.

Radhika for her patience and support in the final stages of this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The need for effective management of traffic congestion has never been greater. In the United States alone, urban freeway delay exceeds 2 billion vehicles hours with the percentage of peak hour travel on urban interstates that occurred under congested conditions reaching 70% in 1989, up from 41% in 1975[42]. Rush-hour conditions in some metropolitan areas extend throughout the day. Furthermore, due to a variety of physical, environmental, and economic constraints, the traditional response of building more roads is no longer feasible.

To develop a coordinated strategy to address urban and suburban congestion, a transportation planning agency must be able to predict the consequences of alternative strategies. This in turn requires that it be able to abstract from a complex system a simplified representation – a *model* – that it can manipulate to analyze the options open to it. The planning applications of such a model could include, for example, evaluation or design of a traffic control system, study of the impact of construction activity on traffic flow distribution, evaluation of alternate incident management schemes, etc.

Regardless of either the model or the application, a necessary input into the planning process is the underlying *demand* for use of the transportation network. The usual way of expressing this demand is by way of an Origin-Destination (O-D) matrix. Each cell of this matrix represents the number of trips between a specific combination of an origin and destination.

In conventional state-of-the-practice planning applications, the O-D matrices that are used are *static* – in that they represent the number of trips between origins and destinations made over a relatively large period (such as an entire day or a morning peak period) within which conditions are assumed "homogeneous". Clearly, this is an approximation of reality, in practice one observes a definite temporal variation of O-D departure rates over the course of an analysis period. Applications that are based on these static O-D flows do not capture the dynamics of build up and dissipation of congestion, of time-varying link and path flows and hence travel times, or of changes in spatial distribution of congestion over time. For any short-term planning study therefore, a knowledge of time-varying or temporal O-D flows could be extremely useful. An entry in a time-dependent or temporal O-D matrix represents the volume of traffic departing from an origin i in time interval h and destined to j.

Estimation and prediction of such time-dependent O-D flows had gained further relevance with increasing attention being paid to Intelligent Transportation Systems (ITS)[1] as a means of alleviating urban and suburban traffic congestion. ITS is an umbrella term that embraces a variety of advanced technologies in the areas of communication, computers, information display, road infrastructure, and traffic control systems[2]. It envisages development of a Dynamic Traffic Management System that would, in real-time, attempt to improve capacity utilization by (a) providing both pre-trip and en-route information to motorists with respect to optimal paths to their destinations and (b) using advanced traffic control systems that are adaptive to rapidly changing traffic conditions in real-time.

It is widely believed (see for example, [29],[8]) that a desirable feature of such systems is the ability to *predict* future traffic, the rationale being that without projection of traffic conditions into the future, control or route guidance strategies are likely to be irrelevant and outdated by the time they take effect. van Toorenburg et al.[45] provide a list of scenarios under which it helps to base decision making (routing or traffic control) on basis of predicted conditions rather than current. They show

[1] formerly Intelligent Vehicle-Highway Systems (IVHS)
[2] For an overview of ITS, refer to Sussman[42].

that a traffic forecast over a few hours for a traveler information system is useful if re-routable traffic volumes are large or travelers get forecasts well in advance for them to have flexibility in their departure times. Forecasts over a much shorter term of a few minutes might be useful to prevent overshoot and unwanted oscillations between different control states. And finally, forecasts are useful if conditions change quickly over the analysis period.

At least two different approaches to Traffic Prediction have been proposed. One approach involves using Dynamic Traffic Assignment (DTA). This offers the advantage of providing the ability to model driver behavior and response to guidance. Another category of methods is the use of statistical techniques ([35],[40],[46]) or state-space based ([23],[47]). While these could offer computational advantages over a DTA based approach, they lack behavioral rules and tend to have a "local" outlook.

One of the most important components of the DTA is the dynamic O-D Estimation and Prediction module. The structure of a DTA is shown in Figure 1-1. Essentially, the DTA accepts inputs from traffic sensors and uses these to estimate and predict Origin-Destination flows. These are then loaded on to the network to generate predictions of network performance measures such as travel times, queue lengths, etc. The predictions, in turn, form the basis for generation of route guidance strategies. Thus, in a DTA based approach for traffic prediction, time-dependent O-D flows are critical for on-line prediction of network performance.

To obtain these matrices directly (for example from surveys) is extremely difficult and costly. The usual procedure hence is to estimate these indirectly from the traffic volumes they induce on the links of the network. The latter can be easily measured using standard surveillance equipment. Of course, the estimation procedure would also include any prior information that is available – in a dynamic context, this typically comes from results of previous estimations.

In this thesis, we present a comprehensive framework for estimation and prediction of these time-dependent O-D matrices. We now define the problem more rigorously.

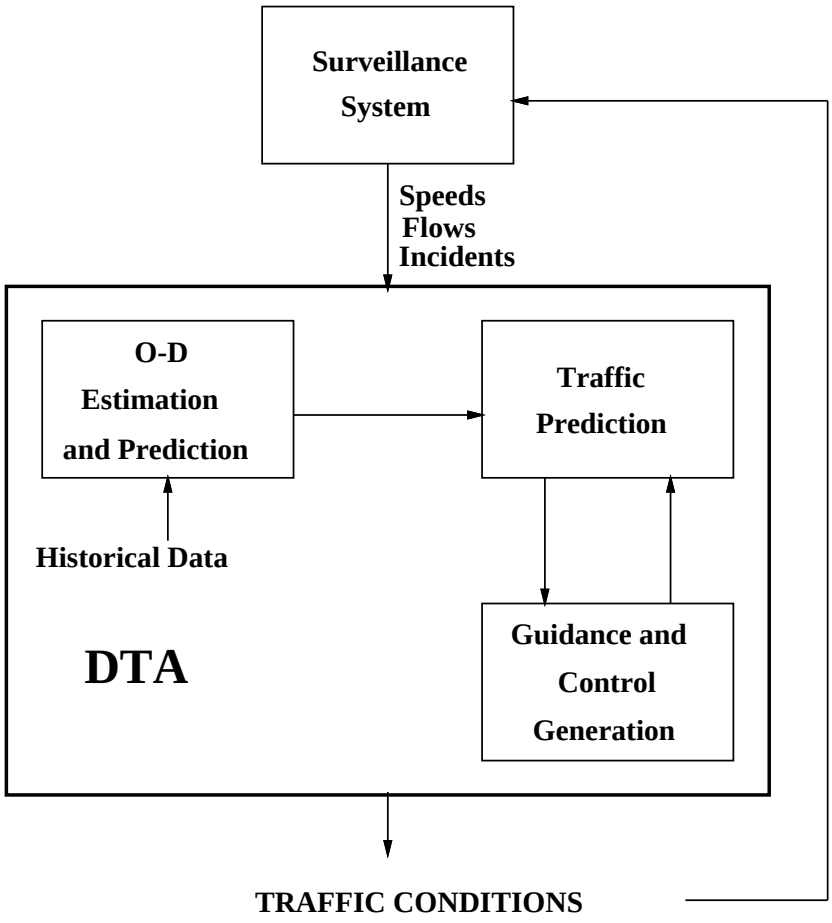

Figure 1-1: Structure of a DTA

1.1 Problem Definition and Thesis Objective

We distinguish between two types of dynamic or time-dependent O-D estimation problems. The first, which we refer to as the *offline* problem, involves estimation of a set of time-dependent O-D matrices given a time-series of link volumes (and other information such as travel times, historical O-D flows, etc). This is more relevant to planning or evaluation studies or as we will see, in the construction of a historical database of time-dependent O-D flows. The *real-time* problem on the other hand, involves O-D estimation in tandem with a DTA within a real-time traffic management system. An additional issue of relevance in the real-time problem is *prediction* of future O-D matrices.

More rigorously, we define the problem as follows:

Offline Estimation

Given a time-series of link traffic counts for h = 1,2,...,N and historical O-D flows, estimate O-D flows for each departure interval.

Real-time Estimation and Prediction

Given link traffic counts in time interval h and historical O-D flows,

1. Estimate O-D matrix for vehicles departing during interval h (and re-estimate O-D matrices for prior departure intervals).

2. Predict O-D matrices for future departure intervals.

The objective of this thesis is to develop a methodology to effectively address the offline and real-time O-D estimation and prediction problems. The models developed here should be capable of handling the most general networks, modeling incomplete/erroneous information and fusing in a consistent manner diverse types of information such as historical O-D flows, link volumes, travel times and speeds, etc. Empirical case-studies should be conducted to evaluate the performance of the proposed model system.

1.2 Literature Review

The *static* O-D matrix estimation problem has been paid considerable attention by several researchers (See [9], [10] for a review and bibliography). These models estimate "average" O-D flows given aggregate link volumes over a relatively long period. Thus, the time varying nature of link and O-D flows over the analysis period is not modeled.

Compared to static estimation, there exists little work on the dynamic front. Research on dynamic estimation and prediction can broadly be categorized into two groups: (a) Pertaining to closed networks and (b) Pertaining to open networks. In this context, a closed network implies that complete information is available on all the entry and exit counts of the network at all points in time.

1.2.1 Closed Networks

All the techniques in this section attempt to find the O-D *split fractions*, i.e., the proportion of each entering volume headed towards each destination. The simplest type of closed network is an isolated intersection for which the O-D flows correspond to turning movements. A variety of methods have been proposed to identify these turning movements based on measurements of entry and exit flows([18],[33]). In one of the earliest works by Cremer and Keller[18], the sequences of short-time exit flow counts are hypothesized to depend upon the time-variable sequences of entry flow counts through a linear relationship. Ordinary Least Squares is used to estimate splitting fractions. This procedure involves computation of autocorrelation and cross-correlation matrices for entrance and exit flows over the period of estimation. A constrained version of this problem has also been proposed to accommodate conditions of non-negativity and the requirement that the turning fractions sum up to unity.

While the above technique can be considered a *batch-processing* technique in that it considers a time series of counts simultaneously to estimate unknown splitting fractions, there have also been a class of *recursive* estimation methods proposed for isolated intersections. Cremer and Keller[18] propose that for each estimation interval, two steps be carried out. In the first step, deviations of exit counts from an

average are predicted using the deviations in the entry counts measured for that interval and the turning fractions of the previous interval. In the second step, these exit count deviations are compared with the deviations actually measured by the surveillance system. Based on the discrepancy between the predicted and observed values, the turning fractions estimated in the previous interval are updated. The recursive updating formula uses a "gain" factor that is pre-specified by the analyst. Cremer and Keller give some recommendations for choosing this factor and show that under some assumptions which they claim, are fairly easily satisfied in practice, the recursive estimates are asymptotically unbiased.

Another recursive estimation procedure suggested by Cremer and Keller for the single intersection is Kalman Filtering. The state vector for each subsystem (pertaining to each exit) is the vector of splitting fractions specific to that exit[3]. As in other approaches, the exiting volumes are represented as a linear combination of the entering inflows. In addition, it is assumed that the splitting fractions follow a random walk process over time. The splitting fractions obtained after each iteration could be normalized (as in [33]) in order to incorporate the constraints mentioned earlier.

For problems of more meaningful size, Ploss and Keller[30] suggest an alternate approach. Essentially, they try to exploit the fact that analysis of a time-series of traffic volumes at entries and exits of a network provides information about the relative frequency of trips between the respective counting sites. A matrix of weights is constructed by them to reflect this correlation. This approach however requires a simplifying assumption about travel times between entry and exit counting stations being equal to a constant integral multiple of the length of each time interval.

A host of methods have been proposed for freeway networks where an O-D flow corresponds to the flow between entry and exit ramps. The relationship between link counts and O-D flows is much more complicated in this case. The earliest approaches to the problem assumed that the travel time required to traverse such networks was

[3]Formally, if b_{ij} is the proportion of traffic from entrance i destined for exit j, then the column j of the matrix $\mathbf{B} = \{b_{ij}\}$ denotes the state vector for j and the row i of $\mathbf{B}$ contains the splitting fractions for the flow entering at i.

either negligible compared to the size of the estimation interval or was equal to a constant number of time intervals. Nihan and Davis[33] and Nihan and Hamad[34] report some experience with estimating freeway O-D matrices under these restrictive assumptions. Nihan and Hamad[34] also conduct some simulation studies on effect of errors in the link counts on the performance of the above models and conclude that such errors severely affect the precision of O-D estimates. However, their network size is extremely small, much of their data is obtained from simulations, and the assumption of constant travel times is completely unrealistic in the presence of congestion.

Recent work in this area comes from Bell[3], van der Zijpp[44], Chang and Wu[16], and Chang and Tao[15]. All of these approaches assume complete information on entering volumes at on-ramps and exiting volumes at off-ramps and try to estimate the destination split fractions in each time interval. In some form or the other, these techniques attempt to model the dynamic interaction between O-D flows and link volumes in a more realistic manner. We discuss below the main features of each.

Bell's approach

Bell[3] suggests two approaches for estimating the split fractions. In the first approach, which might be suitable for single intersections or small networks, it is assumed that the fastest vehicle between each O-D pair reaches its exit within one time interval and the slowest vehicle does not stay on the network for more than one interval. Each exit volume for an estimation interval h is then expressed as a convex combination of two terms. The first term represents the exit volume measured in the previous interval h-1. The second term represents the sum of all the O-D flows departing during interval h that are headed towards the exit under consideration. Formally,

$$y_j(h) = (1 - \alpha_j)y_j(h-1) + \alpha_j \Sigma_i b_{ij} q_i(h) \tag{1.1}$$

where $y_j(h)$ is the volume measured at exit j during interval h, $q_i(h)$ the entering volume at origin i during interval h, b_{ij} the unknown proportion of $q_i(h)$ headed

towards exit j and α_j a travel time "dispersion" parameter also to be estimated.

The second approach allows for vehicles to stay on the network for a pre-fixed number of intervals m with the fastest vehicle still taking less than one interval to traverse the network. For m=3, each exit volume is specified as follows:

$$y_j(h) = \Sigma_i b_{ij0} q_i(h) + \Sigma_i b_{ij1} q_i(h-1) + \Sigma_i b_{ij2} q_i(h-2) \tag{1.2}$$

where b_{ijk} is the (unknown) proportion of traffic from entrance i destined for exit j with a travel time, when truncated, of exactly k intervals.

In both approaches, constrained weighted least squares is used to obtain estimates of the split fractions and additional travel-time dispersion parameters. While Bell's models offer some relaxation of the simplifying assumptions regarding travel times that characterize earlier approaches, a key limitation is that both split fractions and travel time dispersion parameters are assumed to be constant over time. The model therefore does not estimate a time varying split fraction, instead, it attempts to "refine" its previous estimate of the split fractions during each interval.

van der Zijpp's approach

Another approach comes from van der Zijpp[44]. In this approach, boundaries between consecutive time periods are not given by fixed points on the time axis but by time-space trajectories instead. For each departure interval, trajectories of the first and last departing vehicle from the upstream end of the study section to the downstream end are computed (it is assumed that vehicle speeds are known). It is assumed that the trajectories of all other departing vehicles lie in between these and that trajectories of no two vehicles intersect. The trajectories are then used to match measured link counts at various locations with the correct set of (lagged) O-D flows. Split fractions are assumed to follow a truncated multivariate normal distribution (TMVN) to take into account the inequality constraints they are subject to and are updated during each interval using a Bayesian updating formula. A practical difficulty encountered in this approach is computation of the mean of a TMVN during

each interval, for which no closed form solution exists. Another problem is that no analytical expression exists for computation of the prior distribution. Finally, constructing vehicle trajectories for departure interval h requires knowledge of speeds not only for interval h but also potentially for future intervals. Since vehicle trajectories assume a critical role in the model formulation, a mechanism for accurate prediction of speeds is essential in order to use this model in real-time.

Chang and Wu's approach

The next approach in this category comes from Chang and Wu[16]. The problem is formulated as an Extended Kalman filter with the state vector defined as a combination of split fractions and assignment parameters[4]. The measurement equation relates the link counts to the unobserved split fractions, the unknown assignment fractions and the observed entering volumes. An assumption made in constructing this equation is that vehicles belonging to a particular O-D pair and departure interval spend at most two time intervals on each link. To reduce the number of unknown assignment fractions to be estimated, the authors attempt to set selected elements of this matrix to zero based on computed travel times. Travel times for each interval h are computed from mainline counts based on a two-step procedure as follows:

- Compute segment density: The number of vehicles n_l^h present in each mainline segment l of the network at the end of interval h is first estimated using the following flow conservation equation.

$$n_l^h = n_l^{h-1} + I_l^h - O_l^h \tag{1.3}$$

  where I_l^h and O_l^h represent the inflow and outflow into segment l during interval h respectively[5]. The number of vehicles n_l^h is then divided by the length of segment l to get the segment density ρ_l^h.

[4]The assignment parameters define the dynamic mapping between link counts and O-D flows.
[5]The initial state n_l^0 is assumed to be known.

- Compute segment speed and travel time: Segment speed is obtained from segment density using the following relationship:

$$v_l^h = ((I_l^h + O_l^h)/2H)/\rho_l^h \tag{1.4}$$

where H is the length of the time interval. In other words, they assume that the speed within a segment during a given time interval can be approximated by the ratio of the average flow over the interval and the segment density at the end of the interval. Travel time is calculated from the above speed.

The transition equation for their approach is a simple random walk in split fractions and assignment parameters.

While the Chang and Wu approach overcomes several shortcomings of earlier approaches, it still has significant limitations. Firstly – and this is true for all the above approaches – all entering flows are assumed to be known and hence, using the model for prediction requires that incoming flows be predicted as well. Secondly, the procedure advocated for computing travel times is simplistic and unlikely to be accurate in congested networks. Thirdly, estimated travel times are used exclusively for the purpose of selecting elements of the assignment matrix to be zeroed out while they could also be used to generate (perhaps approximate) values for the non-zero elements of the matrix. Finally, the assumption of vehicles sharing an O-D pair and departure interval spending at most two time intervals on each link might be violated for congested freeway corridors with relatively short estimation intervals.

Chang and Tao's approach

The defining characteristic of the approach by Chang and Tao[15] is use of *cordonlines*. A cordonline is defined as a hypothetical closed curve that intersects with a set of links. The basic idea of this approach is to supplement existing sensor based measurements with additional measurement equations that describe the flow across cordonlines. The measurement equation is identical to that in Chang and Wu[16] except that the assignment parameters are assumed to be known. They assume that the split fractions

follow an auto-regressive process in time and use a Kalman Filter to estimate the state during each time interval. They show – based on a case study with synthetic data – that up to 20% improvement in estimation errors can be achieved by using cordonline based measurements in addition to conventional sensor based measurements.

Their model has serious shortcomings:

- The model purports to address urban networks. A key issue in O-D estimation for urban networks is specification of the assignment parameters. The authors do not provide any guidelines on how this might be obtained.
- As all other models for closed networks do, it assumes complete information on entry and exit counts. An even less realistic assumption is that entry and exit flows are completely observable for all the cordonlines used in modeling (the cordonlines themselves are arbitrarily constructed).
- The model cannot be used for prediction by itself since the entry volumes are assumed as known inputs and are not modeled.
- The need for cordonlines is not adequately motivated. Their case study compares a base model that uses only entry/exit counts with another model that uses both entry/exit counts as well as counts across cordonlines. It is no surprise that the second model performs better. It is not clear, however, why they could not have obtained a similar or better improvement by a third model that uses entry/exit counts and individual counts for all the sensors constituting the cordonline.

Conclusion

A general comment about the above techniques for closed systems is in order here. In large and complicated urban and suburban environments, it is very difficult to envisage availability of complete and accurate information at all exit and entry locations. This severely hampers the potential application of these techniques to realistic

situations[6]. They however remain useful starting points for development of more complicated models. An important challenge in developing more sophisticated models is in capturing accurately the dynamic mapping between O-D flows and link volumes in the presence of route choice.

1.2.2 Open Networks

Methods for open networks involve an extension of the static matrix estimation problem. The dynamic formulation of the matrix estimation process would have at its core, an equation of this form:

$$y_{lh} = \sum_{p} \sum_{r} a_{lh}^{rp} x_{rp} \tag{1.5}$$

where y_{lh} is the flow crossing sensor l in time interval h, x_{rp} is the flow between O-D pair r that departed its origin during time interval p and a_{lh}^{rp} is an assignment parameter reflecting the proportion of the demand x_{rp} crossing sensor l in time interval h.

The difference from the static case lies primarily in the dependence of the above parameters on p and h. The matrix of assignment fractions in the static case does not reflect the effects of O-D flows corresponding to *prior* intervals on the link counts observed in any interval since it lacks the granularity of time representation in dynamic implementations.

To our knowledge, only three approaches for dynamic estimation have been proposed for open networks. We discuss below the main features of each of these.

Cascetta et al.'s approach

Cascetta et al.[13] obtain estimates of dynamic O-D flows by optimizing a two part objective function. The first part seeks to minimize the difference between the estimated O-D matrix for an interval and an apriori estimate of the O-D matrix for

[6] One area in which the intersection turning fraction models could be useful is in optimizing signal timings within an adaptive traffic control system.

that interval. The second part seeks to minimize the difference between measured link volumes and those predicted by the model when the estimated O-D flows are assigned to the network. Two estimation procedures are presented – a "simultaneous" estimation procedure and a "sequential" estimation procedure. The simultaneous estimator gives in one step the entire set of O-D vectors for all the time-intervals of estimation using link traffic counts for all the intervals. The sequential estimator on the other hand gives in each step, the O-D matrix for one time-interval by using counts relating to that and previous intervals and possibly, O-D estimates of previous intervals. Apart from the obvious computational advantage, the sequential estimator can use estimated O-D flows obtained in a given interval as an apriori value for the next interval. Cascetta et al. apply a Generalized Least Squares (GLS) estimator and test the performance of their approach for an Italian freeway with encouraging results. This model was also implemented for a downtown Boston network with 692 O-D pairs and 1124 links by Khoshyaran[31]. Results for the Boston network were unsatisfactory and were attributed to inadequacy of available data – only 58 of the 1124 links on the network were instrumented. Moreover, on several key links, detectors provided only partial coverage of lanes and hence measured volumes and speeds were subject to huge errors.

The above model has some shortcomings. First, no formal procedure is developed for assigning weights (variances) to the two terms in the objective function. This feature will be explored in further detail in Chapter 2. Secondly, the model cannot be used for predicting future O-D flows. Nevertheless, this work is pioneering in that explicit equations were proposed for modeling the dynamic mapping between O-D flows and link volumes. Again, this aspect will be dealt with in detail in subsequent chapters.

An alternative model with predictive ability was developed as part of the DRIVE-II DYNA project(Inaudi et al.[26]). In this model, estimation and prediction are dealt with separately. For estimation, the sequential version of the model by Cascetta et al. is used. Values thus estimated are then used to generate predictions by a separate "filtering" approach that combines historical and estimated O-D information using

the concept of "deviations" proposed by Ashok and Ben-Akiva[2] (see below). The main disadvantage of this approach is that the prediction component is completely exogenous to the estimation resulting in a statistically inefficient estimator.

Okutani's approach

The state vector (set of decision variables) here is the vector of unknown O-D flows. This technique is based on equation (1.5) with an additional random error present. In addition, the model includes an autoregressive formulation in which the state vector for period h is related by correlation factors to state vectors for prior periods. The degree of lag is pre-specified by the analyst. Okutani uses standard linear Kalman Filter theory to get optimal estimates of the state vector in each time interval. No information, however, is provided about how the matrix of assignment fractions $\{a_{lh}^{rp}\}$ is computed. Though this model has predictive and updating elements and hence is amenable for use in real-time, there are serious problems with the autoregressive specification. These are discussed in detail in Section 2.3. Kachroo et al.[27] have extended this approach to account for serial correlation of errors in the autoregressive formulation. This is accomplished by augmenting the state vector with additional parameters. They report an improvement in results as a result of this modification.

Ashok and Ben-Akiva's approach

Ashok[1] and Ashok and Ben-Akiva[2] present a Kalman Filter based approach for real-time estimation and prediction[7]. In order to overcome the inadequacy of Okutani's autoregressive specification for O-D flows (See Section 2.3), they introduce the notion of *deviations* of O-D flows from historical estimates. The state-vector is hence defined in terms of O-D deviations that conform to an autoregressive process. The measurement equation is the same as Okutani's. The assignment fractions are sought to be obtained either using the equations derived by Cascetta et al.[13] or by using a DTA. The model is evaluated using actual traffic data from the Massachusetts

[7] This model is described in detail in Chapter 2.

Turnpike with encouraging results.

While the above model overcomes deficiencies of several of its predecessors, it still has some shortcomings. First, no attempt is made to capture errors in the assignment matrix. As explained in Chapter 4, this could induce a bias in the estimates. Secondly, the model requires augmenting the state for a given interval with states corresponding to several prior intervals. This has the effect of greatly increasing the computational load associated with the problem, thereby making a real-time application of the model a formidable task. Finally, the model does not investigate alternate forms of the transition equation that could improve its predictive performance – this is the subject of Chapter 3.

Conclusion

In conclusion, there exist very few techniques available for open networks. Each of the models discussed above has some shortcomings that precludes its use in specific situations. None of the models discussed above captures errors in the mapping between link volumes and O-D flows. Not all of them can be used for both estimation and prediction. Only a couple explicitly make use of historical information in the problem formulation. The framework proposed in this thesis attempts to address these issues.

1.3 Overall Framework for Dynamic O-D Estimation and Prediction

Data for dynamic O-D estimation and prediction come from diverse sources. The purpose of this section is to cast the O-D estimation and prediction problem in the form of a data fusion procedure that combines information from these diverse sources in a statistically efficient manner. We draw upon in this section, the work of Ben-Akiva[9] in the context of static O-D estimation.

For dynamic O-D estimation and prediction, the most commonly available data source is traffic counts at specific locations on the network. Unlike static estimation,

these counts are typically over short time intervals. Traffic counts are an example of *indirect* measurements[8] of the unknown O-D flows. Typically, these counts are expressed as linear combinations of the O-D flows. The counts could either be in the middle of a roadway segment, at entry or exit ramps on freeways, or across a cordon or screenline in an urban area. The mapping between the counts and the O-D flows is termed the *assignment* matrix[9].

For networks of non-trivial size, the number of counts available is far less than the number of O-D pairs. Thus, the indirect measurements as given by the link counts are inadequate for obtaining unique estimates of the O-D flows. This is a fundamental characteristic of the O-D estimation problem – whether static or dynamic. The usual way to supplement the link counts is by specifying additional information in the form of prior O-D matrices. For planning applications, these prior O-D matrices could come from various sources – a partial survey, an out-of-date database, etc. In implementations for dynamic traffic management, these would most likely come from results of previous estimations – either for a previous day or for an earlier departure interval the same day. The latter are particularly useful in light of the correlation between O-D flows across successive time intervals that one typically observes on a within-day basis. However obtained, these prior matrices constitute *direct* measurements of the unknown O-D flows.

A critical issue in using indirect measurements such as traffic counts is in defining the mapping represented by the assignment matrix. This matrix depends upon link travel times and path choice fractions[10]. A third source of (indirect) information, therefore, might be travel speeds on specific links or empirically estimated/calibrated path choice fractions. Instantaneous travel speeds could be obtained for example, from sensors (as in various case studies in this thesis) or from probe vehicles.

The above, by no means, constitute an exhaustive list of all possible types of

[8] As in Ben-Akiva[9]

[9] Note that for a congested network, this mapping would in turn depend on the underlying O-D flows themselves, resulting in a highly non-linear equation.

[10] A detailed discussion of the assignment matrix will be deferred until Chapter 4. Suffice to know here that the dependence of the assignment matrix on travel times (which in turn depend on underlying O-D flows) is extremely complicated.

information. For example, probe vehicles could, in addition to travel times, provide direct measurements on the O-D flows[11]. Intersection turning fractions if available could, in principle at least, provide useful information about network level O-D flows. Information about special events or incidents could help in modifying the prior matrix. Information about route guidance strategies being employed could influence future predictions. We also note that in general, similar type of information could be obtained from multiple sensors. For example, link speeds could be obtained both from probe vehicles and from sensors.

Associated with these different types of measurements are different degrees of error. Broadly, these errors can arise from three sources. The first source of error relates to the inexactness or approximation in the functional form that describes the relationships between the measurements and the unknown variables. For example, the serial correlation between O-D flows could be modeled by an autoregressive process (as in Okutani's work) that is likely to be highly inexact. A second source of error could be imprecision in the parameters that map the measurements onto the unknown variables. For example, the assignment matrix, by virtue of its dependence on possibly erroneous travel times, could be subject to severe errors. A third source of error could be recording or instrumentation error. An obvious example is when a particular sensor systematically underestimates or overestimates a link count due to malfunction.

The problem of O-D estimation may now be stated as one of combining and reconciling information of different types with different error characteristics and from multiple sources. Operationally, this involves jointly estimating the unknown parameters of a system of equations. The joint estimation of such a system of equations is known in econometrics as a *mixed estimation* problem (See for example, Theil[43]). If the variances of the error terms in the various equations are known or can be computed, generalized least squares (linear or non-linear) can be used to estimate the unknown parameters. If the parametric form of the distribution of the error terms can be specified, a maximum likelihood estimator can be employed. Of course if the

[11]For a detailed discussion on using probe vehicles as direct measurements, see the discussion associated with the paper by Hellinga and Van Aerde[4].

error terms are normal, both approaches are equivalent. In the context of a recursive estimation procedure, a generalized least squares based approach has definite computational advantages. The maximum likelihood estimator, however, provides full flexibility in model specification and has desirable statistical properties under general conditions.

1.4 Thesis Contributions

This thesis presents various advancements beyond previous research. Specifically,

- A comprehensive framework for the dynamic O-D estimation and prediction problem for open networks is developed. This framework encompasses a variety of models and has the following main features:
  - Provides a natural way of incorporating information of varying types, with different levels of accuracy, and from multiple sources.
  - Builds on earlier work ([1], [2]) by retaining the idea of *deviations* in the model formulation.
  - Exploits the differential temporal variation of originating trips and destination shares to improve predictions.
  - Incorporates a stochastic mapping between O-D flows and link volumes. In this aspect, it differs quite fundamentally from existing approaches.
- The framework and models are evaluated rigorously using a combination of actual and synthetic traffic data from three different networks. These case studies provide insights into the advantages and disadvantages of specific modeling strategies for different types of networks, extent and quality of available data and level of computational resources.

1.5 Thesis Outline

This thesis is organized as follows. Chapter 2 consists of the basic mathematical framework. This is followed by presentation of an alternate formulation based on originating trips and destination shares in Chapter 3. We then present a fundamental extension in Chapter 4 by allowing for a stochastic assignment matrix. The next chapter discusses in detail, several case studies used for evaluating performance of various models. We conclude with guidelines on future research in Chapter 6.

Chapter 2

The Basic Formulation

In this chapter we formalize the ideas discussed in Chapter 1. We first establish the basic equations in the formulation. We then describe an equivalent state-space based formulation of the problem, following which we describe an offline estimation procedure. We conclude the chapter with a discussion of how the inputs to the model may be calibrated from historical data.

2.1 Overview of Methodology

Figure 2-1 illustrates the basic inputs and outputs into an O-D estimation and prediction model system. This model obtains historical O-D flow values from a database. The module also gets a vector of link counts from the surveillance system[1]. For the offline problem, link counts for the entire analysis period (all the departure intervals) would be available. For the real time problem, at the end of each interval h, only the counts corresponding to h would be available. And finally, the module gets (or computes) estimates of "assignment" matrices, discussed in more detail in later sections[2]. By comparison of the link counts that were measured by the surveillance system with the counts obtaining by assigning the estimated or apriori O-D flows

[1]More generally, it gets a set of direct and indirect measurements from the surveillance system (Refer to Chapter 1 for examples of direct and indirect measurements).

[2]As mentioned in Chapter 1, an assignment matrix maps a set of Origin-Destination matrices into a set of link flows.

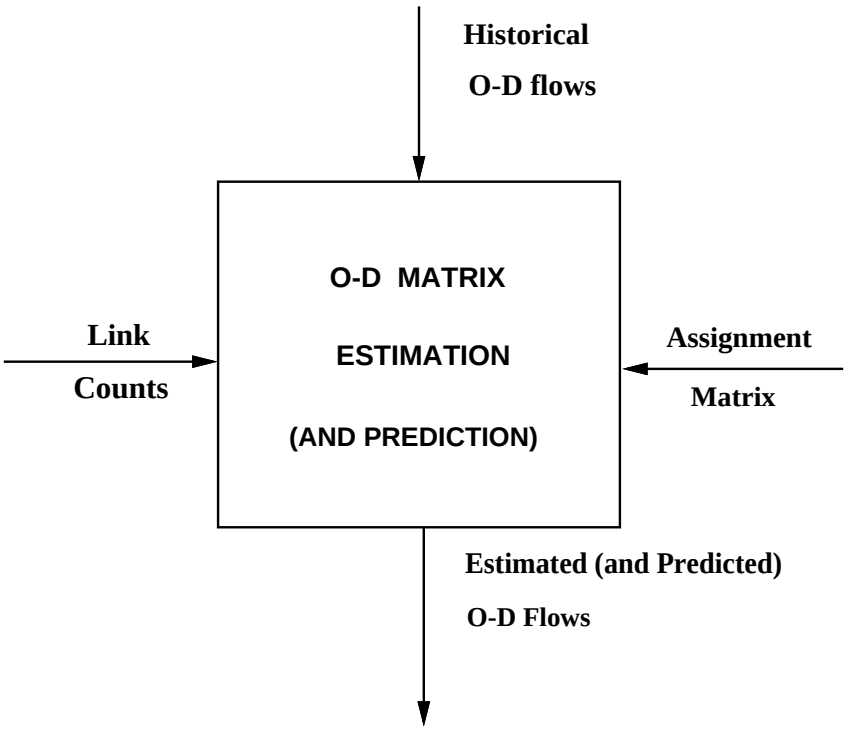

Figure 2-1: Overview of Inputs and Outputs

to the network, the O-D estimation module updates these estimates. For the offline problem, it is possible for this updating to be carried out simultaneously for all the departure intervals. For the real time problem, however, such an updating would have to be performed every estimation interval. Finally, for the real time problem, predictions are generated for intervals $h+1, h+2, ...$ and the process continues.

2.2 Preliminary Definitions

Consider an analysis period of length $\mathcal{T}$ divided into equal intervals $h = 1, 2, ..., N$ of size H. The network is represented as a directed graph that includes a set of consecutively numbered nodes $\mathcal{N}$ and a set of numbered links $\mathcal{L}$. The network is assumed to have n_{LK} links and n_{OD} O-D pairs. It is assumed that n_l of these n_{LK} links are equipped with sensors.

We denote by x_{rh} the number of vehicles between the rth O-D pair that left their origin in interval h and by x_{rh}^{H} the corresponding best historical estimate. A historical estimate typically is the result of estimation conducted during previous days. Further, let the corresponding $(n_{OD} * 1)$ vectors of all O-D flows and the corresponding historical estimates be given by $\mathbf{x}_h$ and $\mathbf{x}_h^H$ respectively. The estimate of $\mathbf{x}_h$ is represented by $\hat{\mathbf{x}}_h$. Finally, denote by y_{lh} the observed traffic counts at detector station l during interval h and by $\mathbf{y}_h$ the corresponding $(n_l * 1)$ vector.

We are now ready to define the set of equations describing the model.

2.3 Direct Measurements

By definition, a direct measurement provides a preliminary estimate of an O-D flow. We therefore express a direct measurement as follows:

$$\mathbf{x}_h^a = \mathbf{x}_h + \mathbf{u}_h \tag{2.1}$$

where $\mathbf{x}_h^a$ denotes the apriori or preliminary estimate of $\mathbf{x}_h$ and $\mathbf{u}_h$ is a vector of random errors.

Equation (2.1) admits a variety of expressions for $\mathbf{x}_h^a$. For example, all of the following are possibilities:

$$\begin{aligned}
\mathbf{x}_h^a &= \mathbf{x}_h^H && (2.2)\\
&= \hat{\mathbf{x}}_{h-1} && (2.3)\\
&= (\mathbf{x}_h^H/\mathbf{x}_{h-1}^H)\hat{\mathbf{x}}_{h-1} && (2.4)\\
&= \mathbf{x}_h^H + \alpha(\hat{\mathbf{x}}_{h-1} - \mathbf{x}_{h-1}^H) && (2.5)\\
&= \Sigma_{p=h-q'}^{h-1}\tau_p(\mathbf{x}_h^H/\mathbf{x}_p^H)\hat{\mathbf{x}}_p \quad , \quad \Sigma_p\tau_p = 1 && (2.6)\\
&= \mathbf{E}_h\mathbf{x}_h^{probe} && (2.7)\\
&= \Sigma_{p=h-q'}^{h-1}\mathbf{f}_h^p\hat{\mathbf{x}}_p && (2.8)\\
&= \mathbf{x}_h^H + \Sigma_{p=h-q'}^{h-1}\mathbf{f}_h^p(\hat{\mathbf{x}}_p - \mathbf{x}_p^H) && (2.9)
\end{aligned}$$

Equations (2.2) and (2.3) are clearly the most straightforward possibilities for incorporating direct measurements – they represent the historical value for interval h and the estimate for the previous interval $h-1$ respectively. We note that Equations (2.2) and (2.3) correspond to the simultaneous and sequential models of Cascetta et al.[13] respectively. Equations (2.4) and (2.5) make use of historical estimates for intervals h and $h-1$ apart from the previous interval estimate. The conjecture underlying Equation (2.4) is that the ratio of interval-over-interval O-D flows is stable on a day-to-day basis. Equation (2.5) attempts to modify the historical estimate $\mathbf{x}_h^H$ based on the *deviation* from historical estimate in the previous interval. Equation (2.6) is a generalization of Equation (2.4) in that it takes into account *many* prior ratios instead of just the ratio $(\mathbf{x}_h^H/\mathbf{x}_{h-1}^H)$.

Equation (2.7) represents the use of information from probe vehicles[3] as direct measurements. $\mathbf{x}_h^{probe}$ represents the number of probe vehicles corresponding to each O-D pair departing during interval h. The matrix $\mathbf{E}_h$ is diagonal and represents an "expansion" factor to account for the fact that probe vehicles constitute only a

[3]A *probe* vehicle is defined as a vehicle equipped with a device that enables a traffic management center to track its progress through the transportation network.

fraction of the total number of vehicles in the network.

Equations (2.8) and (2.9) are most interesting. Equation (2.8) is a more general form of (2.6) and corresponds to the Okutani model. It implies that the O-D flows follow an autoregressive process of order q'. $\mathbf{f}_h^p$ is an $(n_{OD} * n_{OD})$ matrix of effects of $\mathbf{x}_p$ on $\mathbf{x}_h$. On the other hand, Equation (2.9) is a more general form of (2.5) and models the temporal relationship among *deviations* in O-D flows by an autoregressive process of order q'[4]. The matrix $\mathbf{f}_h^p$ is here an $(n_{OD} * n_{OD})$ matrix of effects of $(\mathbf{x}_p - \mathbf{x}_p^H)$ on $(\mathbf{x}_h - \mathbf{x}_h^H)$. This formulation captures correlation over time among deviations which arise from unobserved factors that are correlated over time. Such factors include weather conditions, special events, temporary changes in the transportation network, etc.

A comparison of Equations (2.8) and (2.9) is in order here. An autoregressive process such as (2.8) can only capture temporal interdependencies among O-D flows. Such a model does not represent any structural information about trip patterns. The pattern of O-D trips is a function of spatial and temporal distribution of activities as well as characteristics of the transportation system. It is highly unlikely therefore that a simple autoregressive process (with constant coefficients) would be able to capture the complex structure of activities that result in the spatial and temporal pattern of trip making.

Suppose that O-D matrices have been estimated from historical data for several previous days or months. These already estimated O-D matrices subsume a wealth of information about the relationships that affect trip making and about their variation over space and time. One simple way then of incorporating structural relationships is to include all the prior estimation into the real-time O-D estimation problem. The simplest way to do this is to use *deviations* of O-D flows from best available historical estimates instead of the actual flows themselves as unknown variables. Thus the estimation and prediction process would have indirectly taken into account all the experience gained over many prior estimations and would be richer in its structural content.

[4]This was used in earlier work by Ashok and Ben-Akiva[2].

The idea of deviations overcomes another difficulty that was recognized by Okutani. A normal distribution for traffic variables (such as O-D flows) is a useful property for available statistical tools such as the Kalman Filtering technique used by Okutani. However, the traffic flow variables used by Okutani have skewed distributions whereas the corresponding deviations would have symmetric distributions and hence be more amenable to approximation by a normal distribution.

For these reasons, we adopt Equation (2.9) as the preferred choice for most of the models in this thesis[5]. In view of the fact that we wish to work with deviations instead of the O-D flows themselves, it is convenient to rewrite Equations (2.1) and (2.9) as follows:

$$\partial \mathbf{x}_h^a = \partial \mathbf{x}_h + \mathbf{u}_h \tag{2.10}$$

$$\partial \mathbf{x}_h^a = \Sigma_{p=h-q'}^{h-1} \mathbf{f}_h^p \partial \hat{\mathbf{x}}_p \tag{2.11}$$

where $\partial \mathbf{x}_h$ denotes the deviation in $\mathbf{x}_h$ (i.e., the quantity $(\mathbf{x}_h - \mathbf{x}_h^H)$) and $\partial \mathbf{x}_h^a$ denotes its apriori estimate[6].

Computation of the matrix $\mathbf{f}_h^p$ involves estimation of linear regression models for each O-D pair. The error covariance matrix $\mathbf{Q}_h$ could be approximated from the residuals of these regressions. The exact mechanics will be deferred to Section 2.11.

We close with some important comments. Firstly, it is conceivable that in many situations, one might have multiple sources of information for the *same* direct measurements. In such a situation, we simply write a version of Equation (2.1) (or (2.10)) for each measurement source. The fact that these sources of information could have different degrees of error is taken into account by specifying different variances for the error term $\mathbf{u}_h$ in each of these equations. Thus Equation (2.10) provides an easy way of incorporating the same information from multiple sources.

Secondly, it is conceivable that one might have more than one type of direct

[5]We also use (2.4) for some of our offline models.

[6]Note that in similar fashion, we can rewrite Equation (2.10) for each of the forms (2.2)–(2.8) by subtracting $\mathbf{x}_h^H$ from the respective right hand sides.

measurement. For example, one might wish to use both Equations (2.7) and (2.11). Again, in such a situation we simply use *both* sets of equations and specify error variances separately[7]. Alternatively, we could use a weighted (convex) combination of the two direct measurements where the weights depend on the relative variances of the two[8]. These ideas will be formalized in later sections.

2.4 Indirect Measurements

As mentioned in Chapter 1, link counts constitute the most common type of indirect measurements. The relationship between counts and the unknown O-D flows can be expressed as a linear relationship as follows:

$$y_{lh} = \sum_{p=h-p'}^{h} \sum_{r=1}^{n_{OD}} a_{lh}^{rp} x_{rp} + v_{lh} \tag{2.12}$$

where a_{lh}^{rp} is the fraction of the rth OD flow that departed its origin during interval p and crossed the counting point on link l during interval h. v_{lh} is the measurement error while $(p'+1)$ is the maximum number of time intervals taken to travel between any O-D pair of the network. In matrix form the above equation reduces to:

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p + \mathbf{v}_h \tag{2.13}$$

where the matrix $\mathbf{a}_h^p$ is an $(n_l * n_{OD})$ *assignment* matrix of contributions of $\mathbf{x}_p$ to $\mathbf{y}_h$ and $\mathbf{v}_h$ is the vector of measurement errors[9].

The interpretation of Equation (2.13) is straightforward. The flow across any detector station during interval h is comprised of contributions from O-D flow vectors corresponding to departures during h,h-1,...,h-p'. The assignment matrix consists of the proportions of these O-D flows that constitute the link flow. The error term

[7] If the information from the sources are not independent, covariances have to be specified as well.

[8] Ben-Akiva and Bolduc([5],[6]) use such techniques in a different context of "transferability" of discrete choice model coefficients.

[9] In reality, Equation (2.13) will be highly non-linear because the assignment matrix depends indirectly upon O-D flows. We will consider this issue in Chapter 4.

reflects the possibility of imperfect measurements.

Since we wish to work with deviations, Equation (2.13) can be rewritten as follows:

$$\mathbf{y}_h - \mathbf{y}_h^H = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^p (\mathbf{x}_p - \mathbf{x}_p^H) + \mathbf{v}_h \tag{2.14}$$

where $\mathbf{y}_h^H = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p^H$.

While the error covariance matrix $\mathbf{R}_h$ can be easily computed from historical data (Section 2.11), computation of the matrices $\mathbf{a}_h^p$ is a complicated exercise. The fractions contained in these matrices depend on path choice probabilities as well as the stochastic mapping of time-dependent path flows to link flows[10]. One way of determining the former is by means of discrete choice models. To estimate the latter, one would need in addition, knowledge about time-dependent travel times. Travel times could be obtained from a traffic surveillance system (e.g. probe vehicles or sensors) or from a simulation model. Explicit modeling of the assignment matrices and implications of errors in these on the O-D estimation/prediction process is discussed in detail in Chapter 4.

We make two final observations. Firstly, note that though we have described $\mathbf{y}_h$ as a vector of *link* traffic counts, we could easily handle other types of counts such as screenline counts, cordonline counts, entry/exit counts, etc. by appropriate specification of the assignment matrix $\mathbf{a}_h^p$. Secondly, just as for direct measurements, we could have multiple sensors with different error propensities measure the same link volume. Again, such a situation can be handled by the framework in a natural fashion by appropriate choice of error variances.

2.5 An Equivalent State-Space Model

A classical technique of dealing with dynamic systems is *state-space* modeling (Appendix A). In this section, we demonstrate how the framework presented in previous

[10]Note that in *static* O-D matrix estimation, the mapping of path to link flows is given trivially by a link-path incidence matrix – a matrix of zeros and ones. Because of complicated interaction across time-intervals, this is clearly inappropriate in the dynamic case.

sections can be viewed, under most circumstances, as a state-space based formulation. Since state-space based models have been extensively studied and efficient algorithms have been developed to solve such systems, such an exercise is extremely useful.

To develop a state-space model, we first need to define a *state*. Following the arguments advanced thus far, it is natural to define our state to be the vector of deviations in O-D flows from historical estimates. Once a state is defined, we need to specify *transition* and *measurement* equations.

In a dynamic system, transition equations describe the evolution of the state over time. Measurement equations on the other hand relate the unknown state to their observed indicators. To relate these to the terminology of Sections 2.3 and 2.4, we remark that direct measurements such as (2.6), (2.8), or (2.9) would be expressed by transition equations since they imply a within-day dynamic or evolutionary process that O-D flows or their deviations subscribe to. Direct measurements such as (2.7) would be, in state-space terminology, represented by measurement equations since they do not define a dynamic process that the state evolves according to. For the same reason, indirect measurements in Section 2.4 would *always* be represented by measurement equations in state-space terminology.

Recalling that Equations (2.10) and (2.11) imply an underlying autoregressive process on the O-D flow deviations, we describe the transition equation[11] as follows:

$$\mathbf{x}_{h+1} - \mathbf{x}^{H}_{h+1} = \Sigma^{h}_{p=h+1-q'} \mathbf{f}^{p}_{h+1} (\mathbf{x}_p - \mathbf{x}^{H}_{p}) + \mathbf{w}_{h+1} \tag{2.15}$$

or alternatively,

$$\partial\mathbf{x}_{h+1} = \Sigma^{h}_{p=h+1-q'} \mathbf{f}^{p}_{h+1} \partial\mathbf{x}_p + \mathbf{w}_{h+1} \tag{2.16}$$

where $\mathbf{x}_p$, $\mathbf{x}^{H}_{p}$, $\mathbf{f}^{p}_{h+1}$ are as before and $\mathbf{w}_{h+1}$ is the random error[12]. In scalar form, this

[11] Since we placed no restriction on the number and type of direct measurements, we could in general have multiple direct measurements that could be cast as transition equations. We demonstrate in Section 2.7 how this might be handled.

[12] Notice that the error terms in Equations (2.16) and (2.10) are different (though, as one would expect, they are closely related). We elaborate on this a bit later.

equation can be stated as follows:

$$x_{rh+1} - x^{H}_{rh+1} = \sum_{p=h+1-q'}^{h} \sum_{r'=1}^{n_{OD}} f^{r'p}_{rh+1}(x_{r'p} - x^{H}_{r'p}) + w_{rh+1} \tag{2.17}$$

where the coefficients $f^{r'p}_{rh+1}$ describe the effect of the deviation $(x_{r'p} - x^{H}_{r'p})$ on the deviation $(x_{rh+1} - x^{H}_{rh+1})$ and w_{rh+1} is the random error.

We make the following assumptions:

1. $E[\mathbf{w}_h] = 0$
2. $E[\mathbf{w}_h \mathbf{w}'_l] = \mathbf{Q}_h \delta_{hl}$ where $\delta_{hl} = 1$ if h=l and 0 o.w. $\forall h, l$ and $\mathbf{Q}_h$ is an $(n_{OD} * n_{OD})$ variance-covariance matrix.

The assumption of no serial correlation can be defended because the unobserved factors in the transition equation that could be correlated over time are captured by the historical matrix $\mathbf{x}^{H}_{h}$. In some situations however (e.g. incidents), this assumption might break down. A violation of this assumption can be easily taken care of by using a variant of the estimation algorithm we describe in later sections[13].

The measurement equation which we shall use is identical to (2.13) or in deviation form (2.14), reproduced below for completeness.

$$\mathbf{y}_h - \mathbf{y}^{H}_{h} = \Sigma^{h}_{p=h-p'} \mathbf{a}^{p}_{h} \partial \mathbf{x}_p + \mathbf{v}_h \tag{2.18}$$

We typically assume that

1. $E[\mathbf{v}_h] = 0$
2. $E[\mathbf{v}_h \mathbf{v}'_m] = \mathbf{R}_h \delta_{hm}$ where $\delta_{hm} = 1$ if h=m and 0 o.w. $\forall h, m$ and $\mathbf{R}_h$ is an $(n_l * n_l)$ variance-covariance matrix.

Again, there could be situations in which the assumption of no serial correlation might break down. An example could be, if a specific detector – perhaps because of

[13] An algorithm to handle correlated errors in the transition or measurement equations can be found, for example, in Chui[17].

incorrect calibration – consistently over-estimates or under-estimates a link volume on a particular day. Again, it is easy to relax this assumption and use a variant of the estimation algorithm we describe in later sections.

If additional measurements are available, we simply expand the set of measurement equations to include the additional information. For example, information from probe vehicles would be represented by:

$$\tilde{\mathbf{y}}_h = \partial \mathbf{x}_h + \tilde{\mathbf{v}}_h \tag{2.19}$$

where $\tilde{\mathbf{y}}_h = \mathbf{E}_h \mathbf{x}_h^{probe} - \mathbf{x}_h^H$ (all definitions as before) and $\tilde{\mathbf{v}}_h$ the error[14]. Thus, we would now have two sets of measurements equations[15].

In conclusion, we might mention that the framework in Sections 2.3 and 2.4 does not always lend itself to a state-space interpretation. For example, if we choose Equation (2.2) as a direct measurement instead of Equation (2.9), we cannot write a transition equation. Though choosing (2.2) over (2.9) is not recommended, it could be necessary for example, if existing historical data were insufficient to accurately calibrate $\mathbf{f}_h^p$[16]. In a dynamic traffic management system, however, it is envisaged that data would be available over multiple days, and hence, after a "warm-up" period of a few days, one would be able to calibrate $\mathbf{f}_h^p$ (and other model inputs) leaving no further reason to prefer (2.2). We revisit these issues later.

2.6 State Augmentation

An examination of Equation (2.13) suggests that a link count corresponding to time-interval h provides information not only about $\mathbf{x}_h$ but also about $\mathbf{x}_{h-1}$, $\mathbf{x}_{h-2}$, ...,

[14] $\tilde{\mathbf{v}}_h$ is formally the same as $\mathbf{u}_h$ in Equation (2.1).

[15] An equivalent way of representing (2.19) is as follows:

$$\tilde{\mathbf{y}}_h^* = \mathbf{E}_h^* \partial \mathbf{x}_h + \tilde{\mathbf{v}}_h^* \tag{2.20}$$

where $\mathbf{E}_h^* = \mathbf{E}_h^{-1}$, $\tilde{\mathbf{y}}_h^* = \mathbf{x}_h^{probe} - \mathbf{E}_h^* \mathbf{x}_h^H$ and $\tilde{\mathbf{v}}_h^* = \mathbf{E}_h^* \tilde{\mathbf{v}}_h$. The matrix $\mathbf{E}_h$ is guaranteed to be invertible since it is diagonal with expansion factors ≥ 1.

[16] One way of handling this situation might be to add a random walk in O-D flows or deviations.

$\mathbf{x}_{h-p'}$. Similarly, Equation (2.11) implies a serial correlation in O-D deviations that extends over multiple time periods. This suggests that in order to fully exploit this information, each O-D flow be estimated multiple times. More specifically, it suggests that each O-D flow be estimated $max(p'+1, q')$ times.

The standard technique of achieving this is through State Augmentation[17]. We develop in this section the resulting modifications in specification of the direct and indirect measurements. For the purposes of this section, we assume the existence of an equivalent state-space model, specifically that the O-D flow deviations follow an autoregressive process governed by Equation (2.15). We defer the (less common) situation of no transition equation to a later section.

State Augmentation implies that we re-define the state to include additional variables to be estimated. Since we wish to estimate lagged O-D flows (or rather, O-D flow deviations), we define:

$$\mathcal{X}_h = \begin{bmatrix} \partial\mathbf{x}'_h & \partial\mathbf{x}'_{h-1} & ... & \partial\mathbf{x}'_{h-s} \end{bmatrix}'$$

$$\text{where } s = max(p', q'-1)$$

and correspondingly, the vectors

$$\mathbf{X}_h = \begin{bmatrix} \mathbf{x}'_h & \mathbf{x}'_{h-1} & ... & \mathbf{x}'_{h-s} \end{bmatrix}'$$

$$\mathbf{X}^H_h = \begin{bmatrix} \mathbf{x}^{H'}_h & \mathbf{x}^{H'}_{h-1} & ... & \mathbf{x}^{H'}_{h-s} \end{bmatrix}'$$

We now consider the modifications to the transition and measurement equations.

2.6.1 Transition Equation

Consider the following definitions:

$$\mathbf{F}_h = \begin{bmatrix} \mathbf{f}^h_{h+1} & \mathbf{f}^{h-1}_{h+1} & ... & \mathbf{f}^{h-s}_{h+1} \end{bmatrix}$$

$$\mathbf{\Phi}_h = \begin{bmatrix} & \mathbf{F}_h & \\ \mathbf{I}_{(n_{OD}{}^{s*}n_{OD}{}^{s})} & & \mathbf{0}_{(n_{OD}{}^{s*}n_{OD})} \end{bmatrix}$$

[17] The same technique was employed by Okutani, in his state-space based model.

and

$$\mathbf{W}_{h+1} = \left[\begin{array}{cc} \mathbf{w}'_{h+1} & \mathbf{0}'_{(1*n_{OD}s)} \end{array} \right]'$$

where

$\mathbf{F}_h$ is $(n_{OD} * n_{OD}(s+1))$
$\mathbf{\Phi}_h$ is $(n_{OD}(s+1) * n_{OD}(s+1))$
$\mathbf{W}_{h+1}$is $(n_{OD}(s+1) * 1)$

In the event that $p' > q' - 1$, the additional elements of $\mathbf{\Phi}_h$ are set to zero.
Then, (2.16) can be written as:

$$\mathcal{X}_{h+1} = \mathbf{\Phi}_h \mathcal{X}_h + \mathbf{W}_{h+1} \tag{2.21}$$

From earlier assumptions about $\mathbf{w}_h$, it follows that

1. $E[\mathbf{W}_h] = 0$
2. $E[\mathbf{W}_h \mathbf{W}'_l] = \mathcal{Q}_h \delta_{hl}$ where $\mathcal{Q}_h$ has a top-left block $\mathbf{Q}_h$ and is zero elsewhere.

We also notice that given the definitions $\mathcal{X}_{h+1}, \mathcal{X}_h$ and $\mathbf{\Phi}_h$, Equations (2.10) and (2.11) can be written compactly (for interval $h+1$) as follows:

$$\mathbf{\Phi}_h \hat{\mathcal{X}}_h = \mathcal{X}_{h+1} + \mathbf{U}_{h+1} \tag{2.22}$$

where the augmented error vector $\mathbf{U}_{h+1}$ can be derived by noticing the following:

$$\begin{aligned} \mathbf{\Phi}_h \hat{\mathcal{X}}_h &= \mathbf{\Phi}_h (\mathcal{X}_h + \hat{\mathcal{X}}_h - \mathcal{X}_h) \\ &= \mathbf{\Phi}_h \mathcal{X}_h + \mathbf{\Phi}_h (\hat{\mathcal{X}}_h - \mathcal{X}_h) \\ &= \mathcal{X}_{h+1} - \mathbf{W}_{h+1} + \mathbf{\Phi}_h (\hat{\mathcal{X}}_h - \mathcal{X}_h) \end{aligned}$$

yielding

$$\mathbf{U}_{h+1} = -\mathbf{W}_{h+1} + \mathbf{\Phi}_h \epsilon_h \tag{2.23}$$

where $\epsilon_h = (\hat{\mathcal{X}}_h - \mathcal{X}_h)$ denotes the *estimation error* in $\mathcal{X}_h$[18].

2.6.2 Measurement Equation

Define the $(n_l * n_{OD}(s+1))$ matrix

$$\mathbf{A}_h = \left[\begin{array}{cccc} \mathbf{a}_h^h & \mathbf{a}_h^{h-1} & ... & \mathbf{a}_h^{h-s} \end{array} \right]$$

In the event that $q' - 1 > p'$, the additional elements of $\mathbf{A}_h$ are set to zero.
Then, (2.14) can be written as:

$$\mathbf{y}_h - \mathbf{y}_h^H = \mathbf{A}_h(\mathbf{X}_h - \mathbf{X}_h^H) + \mathbf{v}_h \tag{2.24}$$

or more compactly,

$$\mathcal{Y}_h = \mathbf{A}_h \mathcal{X}_h + \mathbf{v}_h \tag{2.25}$$

where

$\mathcal{Y}_h = \mathbf{y}_h - \mathbf{y}_h^H$ and
$\mathbf{y}_h^H = \mathbf{A}_h \mathbf{X}_h^H$.

We again remark that if additional sets of measurements were available (for example, from probe vehicles), we simply append those to $\mathcal{Y}_h$ and appropriately augment $\mathbf{A}_h$ and $\mathbf{v}_h$.

We are now ready to describe the estimation and prediction methodology.

2.7 Estimation and Prediction

It is convenient to start the presentation of the methodology with reference to the state-space model of Section 2.5. We later provide an equivalent estimator that is based on the idea of fusion of the direct and indirect measurements. In Section 2.11.3, we discuss the case where the framework in Sections 2.3 and 2.4 cannot be represented by an equivalent state-space model.

[18] This explains why the error terms in (2.16) and (2.10) are different.

Equations (2.21) and (2.25) constitute a discrete time linear *Kalman Filter*. The solution of such a system of equations is fairly standard and is summarized below. A detailed derivation may be found in Appendix A.

Assume that the initial state of the system $\mathcal{X}_0$ has known mean $\bar{\mathcal{X}}_0$ and variance $\mathbf{P}_0$. Note that knowledge of $\mathcal{X}_0$ implies knowledge of $\mathbf{x}_0$, $\mathbf{x}_{-1}$, $\mathbf{x}_{-2}$, ... ,$\mathbf{x}_{-s}$ and the corresponding historical estimates. Then, using the assumptions made about the errors in sections 2.3, 2.4 and 2.6.1[19], the following results can be stated[20].

$$\begin{aligned}
\mathbf{\Sigma}_{0|0} &= \mathbf{P}_0 && (2.26)\\
\mathbf{\Sigma}_{h|h-1} &= \mathbf{\Phi}_{h-1}\mathbf{\Sigma}_{h-1|h-1}\mathbf{\Phi}'_{h-1} + \mathcal{Q}_h && (2.27)\\
\mathbf{K}_h &= \mathbf{\Sigma}_{h|h-1}\mathbf{A}'_h(\mathbf{A}_h\mathbf{\Sigma}_{h|h-1}\mathbf{A}'_h + \mathbf{R}_h)^{-1} && (2.28)\\
\mathbf{\Sigma}_{h|h} &= \mathbf{\Sigma}_{h|h-1} - \mathbf{K}_h\mathbf{A}_h\mathbf{\Sigma}_{h|h-1} && (2.29)\\
\hat{\mathcal{X}}_{0|0} &= \bar{\mathcal{X}}_0 && (2.30)\\
\hat{\mathcal{X}}_{h|h-1} &= \mathbf{\Phi}_{h-1}\hat{\mathcal{X}}_{h-1|h-1} && (2.31)\\
\hat{\mathcal{X}}_{h|h} &= \hat{\mathcal{X}}_{h|h-1} + \mathbf{K}_h(\mathcal{Y}_h - \mathbf{A}_h\hat{\mathcal{X}}_{h|h-1}) && (2.32)\\
h &= 1, 2, ...N
\end{aligned}$$

In Kalman Filter terminology, $\hat{\mathcal{X}}_{h|h-1}$ represents a *one-step* prediction of the state $\mathcal{X}_h$. It represents the best knowledge of the deviation $\mathcal{X}_h$ prior to obtaining the link counts for interval h. Equation (2.31) shows how this might be obtained using the autoregressive process on deviations. $\mathbf{\Sigma}_{h|h-1}$ and $\mathbf{\Sigma}_{h|h}$ represent the variances of $\hat{\mathcal{X}}_{h|h-1}$ and $\hat{\mathcal{X}}_{h|h}$. Equation (2.27) shows how $\mathbf{\Sigma}_{h|h-1}$ depends upon both the uncertainty in $\hat{\mathcal{X}}_{h-1|h-1}$ as well as the variances of the error $\mathbf{W}_h$ in the autoregressive process.

The matrix $\mathbf{K}_h$ is called the *gain* matrix. Its interpretation becomes clear from Equation (2.32). Equation (2.32) shows that the *filtered* estimate $\hat{\mathcal{X}}_{h|h}$ can be rep-

[19]We make two additional assumptions. First, the transition and measurement errors are uncorrelated, i.e., $E[\mathbf{w}_h\mathbf{v}'_l] = 0\ \forall h, l$. This is reasonable because they arise from two completely different processes. Second, we assume that the initial state $\mathcal{X}_0$ is independent of the errors $\mathbf{v}_h$ and $\mathbf{W}_h$. Again, this does not seem unreasonable.

[20]The notation $i|j$ indicates an estimate corresponding to interval i based on counts up to and including interval j.

resented as the sum of two terms. The first is simply the one-step prediction (prior estimate) $\hat{\mathcal{X}}_{h|h-1}$. The second term represents the adjustment to be applied to the prior estimate in light of the new measurements $\mathcal{Y}_h$ that have just been obtained. The term $\mathbf{A}_h \hat{\mathcal{X}}_{h|h-1}$ represents, in a sense, a *predicted* deviation in link counts, i.e., the difference between counts obtained by assigning one-step predicted O-D flows and the counts obtained by assigning historical O-D flows. $\mathcal{Y}_h$ represents the deviation in counts *actually* observed. $(\mathcal{Y}_h - \mathbf{A}_h \hat{\mathcal{X}}_{h|h-1})$ therefore represents a "residual". In filter theory, this sequence of residuals is termed the *innovations* sequence. The innovations represent the "new" information in each measurement ($\mathcal{Y}_h$, in our case), i.e., the difference between the actual measurement and the best estimate of the measurement given all past measurements. The gain $\mathbf{K}_h$ can now be interpreted as the weight given to this new information. From Equation (2.28), we can see that as the variance $\mathbf{R}_h$ increases, $\mathbf{K}_h$ decreases and the weight given to this new information decreases as it should.

Equations (2.31) and (2.32) define a *linear* estimator because they involve a linear operation on the measurement data (counts). It is proved in Appendix A that the filter – as given by the above equations – produces the smallest Mean Square Error (MSE) for the state vector among all linear estimators. Under additional conditions of normalcy of the errors, the filter produces the smallest MSE estimate among *all* estimators, whether linear or non-linear. Moreover, the estimate is unbiased and orthogonal to its error.

To extend the model to k-step prediction, all that is required is to multiply the filtered vector by the appropriate $\mathbf{\Phi}$ matrices k times.

What is obtained using the above formulae are the deviations; to obtain the O-D flows themselves, appropriate historical estimates have to be added. A general k-step estimated/predicted value would thus be given by:

$$\hat{\mathbf{x}}_{h+k|h} = \xi(\hat{\mathcal{X}}_{h+k|h}) + \hat{\mathbf{x}}^H_{h+k} \quad \forall k = 0, 1, ... \tag{2.33}$$

where $\xi(.)$ is an operator that extracts the first n_{OD} elements of a vector. Moreover,

one notes from the special structure of the state vector that each flow is filtered[21] $s+1$ times. The first time a flow is filtered (in interval h), the filtered deviations are contained in the first n_{OD} elements of the vector $\hat{\mathcal{X}}_{h|h}$. The next time it is filtered in interval $h+1$, the filtered deviations occupy places $(n_{OD}+1)$ to $(2*n_{OD})$ in the vector $\hat{\mathcal{X}}_{h+1|h+1}$. Thus, all the vectors from $\hat{\mathcal{X}}_{h|h}$ to $\hat{\mathcal{X}}_{h+s|h+s}$ would contain some estimate of the deviation in the filtered flow $\hat{\mathbf{x}}_{h|h}$. By adding to these the vector $\hat{\mathbf{x}}_h^H$, the actual O-D flow values can be retrieved. Since the last estimate makes use of the most information, it has the least variance.

The system of equations presented above can also be interpreted in the context of the direct and indirect measurements of Sections (2.3) and (2.4). Viewed in this fashion, the equations comprising the update step (Equation (2.32)) of the Kalman filter can be viewed as solution of a *mixed* estimation problem during each interval.

We notice that once the direct and indirect measurements are expressed (Equations (2.22) and (2.25)), we have a total of (n_l+n_{OD}) equations and n_{OD} unknowns to be estimated during each time interval[22]. We can now use a Generalized Least Squares (GLS) approach to estimate the O-D flows for each time interval. Specifically, this involves minimization of the following error criterion:

$$\begin{aligned}\hat{\mathcal{X}}_{h|h} &= arg\ min\ (\mathcal{X}_h-\hat{\mathcal{X}}_{h|h-1})'\mathbf{P}_{h|h-1}^{-1}(\mathcal{X}_h-\hat{\mathcal{X}}_{h|h-1}) \\ &\quad +(\mathcal{Y}_h-\mathbf{A}_h\mathcal{X}_h)'\mathbf{R}_h^{-1}(\mathcal{Y}_h-\mathbf{A}_h\mathcal{X}_h) \qquad (2.34)\end{aligned}$$

where $\mathbf{P}_{h|h-1}$ denotes the variance-covariance matrix of the prior estimate $\hat{\mathcal{X}}_{h|h-1}$[23]. If non-negativity constraints on the O-D flows are ignored, the standard GLS estimator can be used to obtain a closed form expression for the O-D flow estimates. In other words, the estimate $\hat{\mathcal{X}}_{h|h}$ would be given by:

$$\hat{\mathcal{X}}_{h|h}=(\mathcal{A}_h'\,\mathcal{P}_h^{-1}\mathcal{A}_h)^{-1}\mathcal{A}_h'\,\mathcal{P}_h^{-1}\mathcal{Z}_h \qquad (2.35)$$

[21] We shall use "filtered" and "estimated" interchangeably.
[22] We would have more than (n_l+n_{OD}) equations if we used additional sets of measurements.
[23] which from equation (2.22) is the variance of the error $\mathbf{U}_h$

where

$$\mathcal{A}_h = \begin{bmatrix} \mathbf{A}_h \\ \mathbf{I}_{(n_{OD} * n_{OD})} \end{bmatrix},$$

$$\mathcal{Z}_h = \begin{bmatrix} \mathcal{Y}_h \\ \hat{\mathcal{X}}_{h|h-1} \end{bmatrix},$$

and

$$\mathcal{P}_h = \begin{bmatrix} \mathbf{R}_h & \mathbf{0}_{(n_l * n_{OD})} \\ \mathbf{0}_{(n_{OD} * n_l)} & \mathbf{P}_{h|h-1} \end{bmatrix}$$

The key result here is as follows. *It can be proved (Appendix B) that the update equation of the Kalman Filter (Equation (2.32)) is identical to Equation (2.35) with the variance* $\mathbf{P}_{h|h-1}$ *comprised within* $\mathcal{P}_h$ *in Equation (2.35) computed recursively using the filter variance propagation equations (2.27),(2.28) and (2.29)*[24].

As we have mentioned earlier, there could arise situations with multiple direct measurements that could be represented by transition equations. Such situations can be easily handled either by Equations (2.35) or an extension of (2.26)–(2.32). We outline the estimation/prediction procedure for the case with two direct measurements; a generalization to an arbitrary number of measurements is obvious. Assume that the two direct measurements are of the form:

$$\mathbf{\Phi}_{h,1} \hat{\mathcal{X}}_h = \mathcal{X}_{h+1} + \mathbf{U}_{h+1,1} \tag{2.36}$$

$$\mathbf{\Phi}_{h,2} \hat{\mathcal{X}}_h = \mathcal{X}_{h+1} + \mathbf{U}_{h+1,2} \tag{2.37}$$

or equivalently,

$$\mathcal{X}_{h+1} = \mathbf{\Phi}_{h,1} \mathcal{X}_h + \mathbf{W}_{h+1,1} \tag{2.38}$$

[24]While a detailed proof is given in the Appendix, we make one quick observation here. From equation (2.23), the variance of the error $\mathbf{U}_h$ is given by:

$$\begin{aligned} var(\mathbf{U}_h) &= \mathcal{Q}_h + \mathbf{\Phi}_{h-1} var(\hat{\mathcal{X}}_{h-1} - \mathcal{X}_{h-1}) \mathbf{\Phi}'_{h-1} \\ &= \mathcal{Q}_h + \mathbf{\Phi}_{h-1} \mathbf{\Sigma}_{h-1|h-1} \mathbf{\Phi}'_{h-1} \end{aligned}$$

which, not-by-accident is identical to (2.27).

$$\mathcal{X}_{h+1} = \mathbf{\Phi}_{h,2}\mathcal{X}_h + \mathbf{W}_{h+1,2} \tag{2.39}$$

We then modify the prediction steps (2.27) and (2.31) as follows:

$$\mathbf{\Sigma}_{h|h-1,1} = \mathbf{\Phi}_{h-1,1}\mathbf{\Sigma}_{h-1|h-1}\mathbf{\Phi}'_{h-1,1} + \mathcal{Q}_{h,1} \tag{2.40}$$

$$\mathbf{\Sigma}_{h|h-1,2} = \mathbf{\Phi}_{h-1,2}\mathbf{\Sigma}_{h-1|h-1}\mathbf{\Phi}'_{h-1,2} + \mathcal{Q}_{h,2} \tag{2.41}$$

$$\mathbf{\Sigma}_{h|h-1} = (\mathbf{\Sigma}^{-1}_{h|h-1,1} + \mathbf{\Sigma}^{-1}_{h|h-1,2})^{-1} \tag{2.42}$$

$$\hat{\mathcal{X}}_{h|h-1,1} = \mathbf{\Phi}_{h-1,1}\hat{\mathcal{X}}_{h-1|h-1} \tag{2.43}$$

$$\hat{\mathcal{X}}_{h|h-1,2} = \mathbf{\Phi}_{h-1,2}\hat{\mathcal{X}}_{h-1|h-1} \tag{2.44}$$

$$\hat{\mathcal{X}}_{h|h-1} = \mathbf{\Sigma}_{h|h-1}(\mathbf{\Sigma}^{-1}_{h|h-1,1}\hat{\mathcal{X}}_{h|h-1,1} + \mathbf{\Sigma}^{-1}_{h|h-1,2}\hat{\mathcal{X}}_{h|h-1,2}) \tag{2.45}$$

where $\hat{\mathcal{X}}_{h|h-1,1}$ and $\hat{\mathcal{X}}_{h|h-1,2}$ denote one-step predictions according to Equations (2.40) and (2.41) respectively while $\mathbf{\Sigma}_{h|h-1,1}$ and $\mathbf{\Sigma}_{h|h-1,2}$ denote their variances. Then a "combined" estimate $\hat{\mathcal{X}}_{h|h-1}$ is constructed by a weighted average, the weights being the inverse covariance matrices. This weighted average represents the minimum variance combination of the two estimates $\hat{\mathcal{X}}_{h|h-1,1}$ and $\hat{\mathcal{X}}_{h|h-1,2}$[25]. This minimum variance is given by Equation (2.42). Once the one-step prediction $\hat{\mathcal{X}}_{h|h-1}$ and its variance are specified, the procedure for computing the gain $\mathbf{K}_h$, $\mathbf{\Sigma}_{h|h}$ and $\hat{\mathcal{X}}_{h|h}$ is exactly the same. For the GLS estimator (2.35), exactly the same equation can be used with $\mathbf{P}_{h|h-1}$ now obtained by (2.42). Alternatively, redefining

$$\mathcal{A}_h = \begin{bmatrix} \mathbf{A}_h \\ \mathbf{I}_{(n_{OD}*n_{OD})} \\ \mathbf{I}_{(n_{OD}*n_{OD})} \end{bmatrix},$$

$$\mathcal{Z}_h = \begin{bmatrix} \mathcal{Y}_h \\ \hat{\mathcal{X}}_{h|h-1,1} \\ \hat{\mathcal{X}}_{h|h-1,2} \end{bmatrix},$$

[25] That this is indeed the minimum variance combination follows from a related proof in Appendix B. Required here is an assumption that the errors $\mathbf{W}_{h+1,1}$ and $\mathbf{W}_{h+1,2}$ are uncorrelated; a more general case can be easily handled if the covariance matrix is specified. Also, if the two errors are unbiased, so is the estimate $\hat{\mathcal{X}}_{h|h-1}$.

and

$$\mathcal{P}_h = \begin{bmatrix} \mathbf{R}_h & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{h|h-1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{h|h-1,2} \end{bmatrix}$$

and applying (2.35) with these new definitions yields the same results.

To summarize, the dynamic O-D estimation problem, just like its static counterpart, can be viewed as one of reconciling information from (at least) two different sources – the link counts and the apriori O-D flows. The difference in the equations we have presented stems from the fact that the apriori information is in the form of an evolutionary process (and an estimate of the O-D flows for the first estimation interval). Additionally, if *observability* (see Section 2.8) conditions are satisfied, the effect of the initial estimates gradually washes away as the estimation proceeds.

2.8 System Observability

A discussion of "observability"[26] which is a desirable property of such dynamic systems is in order here. Essentially, observability defines our ability to determine the initial state vector $\mathcal{X}_0$ uniquely from a set of measurements. Under conditions of non-observability, the effects of the initial estimates do not disappear with time and therefore it is critical to obtain accurate initial values. An analogous situation may be found in conventional static matrix estimation where one typically starts with an apriori matrix and uses traffic counts to modify the apriori estimates. The apriori matrix is necessitated by the fact that the number of observations (traffic counts) is much less than the number of O-D pairs. The apriori matrix in effect increases the number of observations. However, the estimates obtained would always depend on the apriori information provided. In the model proposed here, the initial state vector $\mathcal{X}_0$ and the transition equation provide information similar to that provided by the apriori matrix in conventional estimation by exploiting temporal interdependencies

[26]For a rigorous definition of observability, see for example Gelb[20].

among the O-D flows. The nice feature, however, of dynamic or real-time updating is that under conditions of observability, the influence of the initial value of the state vector would disappear with time.

The most obvious and critical factor affecting observability in our problem is the ratio (n_l/n_{OD}). For a given number of O-D flows, measuring more (independent) counts increases the chances of observability being satisfied. The degree of linkage between O-D flows and counts is another factor. An extreme example of this arises when an entire column of the assignment matrix is zero implying that a particular O-D flow never gets measured[27]. A third consideration is the degree of linkage between O-D flows over time. Again, an extreme example of this arises when the transition matrix is fully populated by zeros.

In the context of observability, results from empirical studies (Chapter 5) conducted on the proposed model are encouraging. It was observed that for different values of initial starting conditions, the model produced identical filtered estimates. Another positive indication while testing the model was that it invariably succeeded in finding a gain matrix that moved the predicted estimates closer to the true values.

2.9 A Smoothing Algorithm

The model and the associated recursive algorithms (equations (2.35) or (2.26)–(2.32)) we have described thus far process measurements sequentially, i.e., interval by interval. More importantly, they address the problem of determining $\mathbf{x}_h$ given link counts up to and including interval h[28]. Such a model is particularly relevant for real-time estimation. While nothing precludes application of the same model for offline estimation (for example, for constructing the historical database that we have so far assumed to exist) a distinguishing feature of the latter problem is that the *entire* time-series of measurements is available for processing. The question that naturally

[27]This might however be overcome if that O-D flow is related to other measurable O-D flows by means of a non-diagonal transition matrix.

[28]As mentioned before, we will throughout this thesis refer to this as the *Estimation* or *Filtering* problem.

arises is whether the extra information represented by measurements during intervals $h+1, h+2, ..., N$ is useful[29] in improving estimates of $\mathbf{x}_h$ obtained using measurements from intervals $1, 2, ..., h$. The answer is affirmative, i.e., the smoothed estimates will have an error covariance that is at most equal to that of the filtered estimates[30]. Intuitively, this is because the transition equation of the state-space model postulates a relationship between states corresponding to multiple time intervals. A measurement corresponding to interval $t > h$ can hence be related to $\mathbf{x}_h$ for all $h < t$ through a single application of the associated measurement equation and repeated application of the transition equation (2.15).

The smoothing algorithm that we describe consists of a set of *backward* recursions that start with the final quantities $\hat{\mathcal{X}}_N$ and $\mathbf{\Sigma}_{N|N}$[31]. The recursion is defined as follows:

$$\begin{aligned} \hat{\mathcal{X}}_{h|N} &= \hat{\mathcal{X}}_{h|h} + \mathbf{G}_h(\hat{\mathcal{X}}_{h+1|N} - \hat{\mathcal{X}}_{h+1|h}) \\ \mathbf{\Sigma}_{h|N} &= \mathbf{\Sigma}_{h|h} + \mathbf{G}_h(\mathbf{\Sigma}_{h+1|N} - \mathbf{\Sigma}_{h+1|h})\mathbf{G}'_h \\ \mathbf{G}_h &= \mathbf{\Sigma}_{h|h}\mathbf{\Phi}'_h\mathbf{\Sigma}^{-1}_{h+1|h} \\ h &= N-1, N-2, ...1 \end{aligned} \tag{2.46}$$

In the above equations, $\hat{\mathcal{X}}_{h|N}$ denotes the estimate of $\mathcal{X}_h$ based on the entire set of link counts (for intervals 1,2,...,N) and $\mathbf{\Sigma}_{h|N}$ denotes its covariance. In order to apply these equations, we need to first apply Equations (2.26)–(2.32) for h=1,2,...,N to obtain $\hat{\mathcal{X}}_{N|N}$ and $\mathbf{\Sigma}_{N|N}$ – the filtered estimate and its covariance for the last interval.

The nice feature of the above equations is that they maintain the principal advantage of the Kalman Filter as a recursive, computationally efficient estimation algorithm. The only extra requirements are that the values of $\hat{\mathcal{X}}_{h|h-1}$ and $\hat{\mathcal{X}}_{h|h}$ $\forall h = 1, 2, ..., N$ along with the covariances have to be stored while running the forward

[29]This problem of determining $\mathbf{x}_h$ given measurements from intervals 1,2,...t^* where $1 \leq h < t^* \leq N$ is referred to in literature as *smoothing*[20].

[30]A proof of this assertion may be found in Gelb[20].

[31]This algorithm is referred to in literature as the *Rauch-Tung-Striebel* fixed-interval optimal smoother. Detailed derivations might be found in Rauch[38], Rauch et al.[39], etc.

filter[32]. Since the algorithm is intended for offline estimation, the additional computational burden imposed by the backward pass may be justifiable.

Just as Equation (2.34) defined a mixed estimator equivalent to (2.26)–(2.32), we next provide an alternate version of (2.46) that lends itself to easier interpretation. This involves minimization of the following error criterion[33]:

$$\begin{aligned}\hat{\mathcal{X}}_{h|N} &= arg\ min\ (\hat{\mathcal{X}}_{h+1|N} - \mathbf{\Phi}_h \mathcal{X}_h)' \mathcal{Q}_{h+1}^{-1} (\hat{\mathcal{X}}_{h+1|N} - \mathbf{\Phi}_h \mathcal{X}_h) \\ &\quad + (\mathcal{X}_h - \hat{\mathcal{X}}_{h|h})' \mathbf{\Sigma}_{h|h}^{-1} (\mathcal{X}_h - \hat{\mathcal{X}}_{h|h}) \end{aligned} \tag{2.47}$$

From Equation (2.47), a GLS estimator similar to (2.35) may easily be constructed. Again, this estimator is applied backwards starting from $h = N-1, N-2, ..., 1$, after a forward pass in which $\hat{\mathcal{X}}_{h|h}$ and $\mathbf{\Sigma}_{h|h}$ for $h = 1, 2, ..., N$ are computed. To accommodate any additional direct measurements describing transition dynamics that may exist, we append additional quadratic terms to (2.47), just as we did to (2.34).

And finally we provide yet another equivalent form of (2.46) that may be viewed as a "simultaneous" estimator. This formulation is stated as follows[34]:

$$\begin{aligned}(\hat{\mathcal{X}}_{0|N}, \hat{\mathcal{X}}_{1|N}, ..., \hat{\mathcal{X}}_{N|N}) &= argmin[(\mathcal{X}_0 - \bar{\mathcal{X}}_0)' \mathbf{P}_0^{-1} (\mathcal{X}_0 - \bar{\mathcal{X}}_0) \\ &+ \sum_{h=1}^{h=N} (\mathcal{Y}_h - \mathbf{A}_h \mathcal{X}_h)' \mathbf{R}_h^{-1} (\mathcal{Y}_h - \mathbf{A}_h \mathcal{X}_h) \\ &+ \sum_{h=1}^{h=N} (\mathcal{X}_h - \mathbf{\Phi}_{h-1} \mathcal{X}_{h-1})' \mathcal{Q}_h^{-1} (\mathcal{X}_h - \mathbf{\Phi}_{h-1} \mathcal{X}_{h-1}) \end{aligned} \tag{2.48}$$

Equations (2.46) and (2.47) are clearly much better forms for computational purposes but (2.48) serves two purposes. First, it has a nice least-squares interpretation[35]. Second, it allows us to evaluate Cascetta et al.'s simultaneous estimator that also attempts to use the entire set of counts (See Section 2.11.3). The key difference is that the latter does not allow for inclusion of information in the form of the autoregressive

[32] There are alternate forms of equations (2.46) that do not require storage of the covariances.

[33] For a proof that this minimization is equivalent to applying (2.46), refer to Rauch et al.[39].

[34] Again, an algebraic proof that this form is equivalent to the two earlier forms is rather tedious and will not be attempted here. The interested reader is again referred to Rauch et al.[39].

[35] If we had additional direct measurements, we append additional quadratic terms to (2.48).

process or indeed, any form of temporal relationship between O-D flows. This could be an important limitation since in a dynamic setting, one would strongly suspect some form of a systematic temporal evolution of O-D flows. In Cascetta et al.'s formulation, any information other than link counts (i.e., any direct measurements) can be specified *only* through an apriori matrix. Another important advantage of (2.48) is that by virtue of its equivalence with (2.46) or (2.47), the solution can be expressed recursively, thus breaking up one big problem into many smaller manageable ones.

2.10 An Approximation

From the nature of the augmentation described in Section 2.6, it can be seen that during each interval, $n_{OD}(s+1)$ flows are estimated[36]. This imposes an enormous computational strain for large and congested networks. For example, the size of the variance covariance matrix $\mathbf{\Sigma}$ after augmentation is $(n_{OD}(s+1) * n_{OD}(s+1))$[37], manipulation of $\mathbf{\Sigma}$ in Equations (2.27), (2.28), and (2.29) therefore becomes cumbersome. As the congestion level in the network increases, the problem becomes worse because the number of lagged states s could increase with increase in travel times. The approximation we propose is based on the conjecture that much of the information about an O-D flow is likely to be provided the first time it is counted. If this were true, O-D flows corresponding to prior departure intervals could be held constant at their prior estimated values and only the flows for the current departure interval need to be estimated. The measurement and transition equations in the state-space model would then be expressed as follows:

$$\mathbf{y}_h = \mathbf{a}_h^h(\mathbf{x}_h - \mathbf{x}_h^H) + \mathbf{b}_h + \mathbf{v}_h \tag{2.49}$$

and

[36] Alternatively each O-D flow is estimated $s+1$ times.
[37] And it is in general, a full matrix.

$$\mathbf{x}_{h+1} - \mathbf{x}^H_{h+1} = \mathbf{f}^h_{h+1}(\mathbf{x}_h - \mathbf{x}^H_h) + \mathbf{c}_{h+1} + \mathbf{w}_{h+1} \tag{2.50}$$

where

$\mathbf{b}_h = \Sigma_{p=h-p'}^{h-1} \mathbf{a}^p_h \hat{\mathbf{x}}_p + \mathbf{a}^h_h \mathbf{x}^H_h$,

$\mathbf{c}_{h+1} = \Sigma_{p=h+1-q'}^{h-1} \mathbf{f}^p_{h+1}(\hat{\mathbf{x}}_p - \mathbf{x}^H_p)$ and $\hat{\mathbf{x}}_p$ is a filtered estimate of $\mathbf{x}_p$[38].

The extent to which the conjecture might be true in a given situation depends on a number of factors. Obviously, it is more likely to hold with low measurement errors – a second measurement might not contribute much over the first. Paradoxically, it might also hold in the presence of very *high* measurement errors – in that case, measurements become so bad that each additional set offers hardly any improvement. Another factor is the error in the transition equation. If the errors $\mathbf{w}_{h+1}$ have a high variance, the importance of the counts as extra sources of information increases and the conjecture is less likely to hold.

Estimation of the O-D deviations from the above system is similar in spirit to those presented earlier except for the presence of constants in the transition and measurement equations. For the sake of completeness, we reproduce them here. Note that the state now comprises O-D flows only of *one* departure interval. The matrix $\mathbf{\Sigma}$ is as before the variance covariance matrix of the state.

$$\begin{aligned}
\mathbf{\Sigma}_{0|0} &= \mathbf{P}_0 \\
\mathbf{\Sigma}_{h|h-1} &= \mathbf{f}^{h-1}_h \mathbf{\Sigma}_{h-1|h-1} \mathbf{f}^{h-1\prime}_h + \mathbf{Q}_h && (2.51) \\
\mathbf{K}_h &= \mathbf{\Sigma}_{h|h-1} \mathbf{a}^{h\prime}_h (\mathbf{a}^h_h \mathbf{\Sigma}_{h|h-1} \mathbf{a}^{h\prime}_h + \mathbf{R}_h)^{-1} && (2.52) \\
\mathbf{\Sigma}_{h|h} &= \mathbf{\Sigma}_{h|h-1} - \mathbf{K}_h \mathbf{a}^h_h \mathbf{\Sigma}_{h|h-1} && (2.53) \\
\hat{\partial \mathbf{x}}_{0|0} &= E(\bar{\mathbf{x}}_0 - \mathbf{x}^H_h) \\
\hat{\partial \mathbf{x}}_{h|h-1} &= \mathbf{f}^{h-1}_h \hat{\partial \mathbf{x}}_{h-1|h-1} + \mathbf{c}_h && (2.54) \\
\hat{\partial \mathbf{x}}_{h|h} &= \hat{\partial \mathbf{x}}_{h|h-1} + \mathbf{K}_h (\mathbf{y}_h - \mathbf{a}^h_h \hat{\partial \mathbf{x}}_{h|h-1} - \mathbf{b}_h) && (2.55) \\
h &= 1, 2, ...
\end{aligned}$$

[38] Though we state the approximation here for a particular transition equation, clearly the approach is generalizable to any type of direct measurement.

where the initial state $\partial \mathbf{x}_0$ has mean $E(\bar{\mathbf{x}}_0 - \mathbf{x}_h^H)$ and variance $\mathbf{P}_0$ respectively.

2.11 Estimation of Model Inputs

The framework and algorithms described thus far require that several matrices be fully known. In this section, we shall discuss how the matrices $\mathbf{f}_h^p$, $\mathbf{x}_h^H$, $\mathbf{Q}_h$ and $\mathbf{R}_h$ may be estimated. Computation of the vectors $\mathbf{a}_h^p$ is the subject of Chapter 4.

2.11.1 Estimating $\mathbf{f}_h^p$

As mentioned earlier, the matrix $\mathbf{f}_h^p$ consists of elements $\{f_{rh}^{r'p}\}$ which are essentially measures of the effects on deviations in the rth O-D flow in period h of lagged O-D flow deviations. This matrix would be estimated offline using historical data on O-D flows[39].

Estimation of the matrix would be done element by element during each interval. For estimation of $f_{rh+1}^{r'p} \quad \forall r' = 1, 2, ..., n_{OD}$, we could have a regression of the form:

$$x_{rh+1} - x_{rh+1}^H = \sum_{p=h+1-q'}^{h} (f_{rh+1}^{1p}(x_{1p} - x_{1p}^H) + ... + f_{rh+1}^{n_{OD}p}(x_{n_{OD}p} - x_{n_{OD}p}^H)) + w_{rh+1} \tag{2.56}$$

where w_{rh+1} is the error. Thus there would be n_{OD} such regressions needed to obtain the entire $\mathbf{f}_h^p$ matrix. Moreover, one would have to obtain such a matrix for each h – i.e., each day of history would yield exactly one observation for calibration. However, if one makes the assumption that the structure of the autocorrelation remains constant with respect to h, one could write equations similar to (2.56) for each interval of *one* day and have enough observations to estimate the elements of the matrix. In that case, the values of the matrix $\mathbf{f}_h^p$ would only depend on the difference $h - p$ and not on individual values of h and p.

[39] Often, such data is not available. In such a situation, one of the simpler offline models described in Section 2.11.3 would be used for the first few days to generate a preliminary database of O-D flows.

To simplify the problem further, it may be reasonable in some situations to assume that deviations in the rth O-D flow would be most affected by those in *the preceding rth O-D flows alone* and that contributions from other O-D pairs would be insignificant in comparison. Under this approximation, we would have n_{OD} regressions of a much simpler form:

$$x_{rh+1} - x^{H}_{rh+1} = \sum_{p=h+1-q'}^{h} f^{rp}_{rh+1}(x_{rp} - x^{H}_{rp}) + w_{rh+1} \tag{2.57}$$

and the $\mathbf{f}^{p}_{h}$ matrix would be diagonal. The value of q' would be obtained from statistical significance tests on regression coefficients for various lags. It is expected that the matrices $\mathbf{f}^{p}_{h}$ would be sparse.

2.11.2 Estimating the error covariances

The error covariances can be obtained from historical data in a fairly straightforward manner. The matrix $\mathbf{Q}_{h}$ would be obtained by an OLS regression on equation (2.56). The (i,j)th element of this matrix could be approximated by

$$Q_{ijh} = \mathbf{e}'_{ih}\mathbf{e}_{jh}/n \tag{2.58}$$

where $\mathbf{e}$ is the OLS residual vector and n is the number of sample observations. The above equation assumes dependence on h. This can be removed by assuming that the structure of the autocorrelation remains constant. In that case, the matrix $\{Q_{ijh}\}$ would reduce to $\{Q_{ij}\}$.

Similarly one can obtain the matrix $\mathbf{R}_{h}$ from the residuals of the measurement equation. Here the residuals $\mathbf{e}_{h}$ would be obtained from computing the differences $(\mathbf{y}_{h} - \Sigma^{h}_{p=h-p'}\mathbf{a}^{p}_{h}\hat{\mathbf{x}}_{p})$ over many days. Each day would yield one value for every residual vector $\mathbf{e}_{h}$. From the values of the residuals $\mathbf{e}_{h}$ over several days, the variance-covariance matrices $\mathbf{R}_{h}$ are calculated. Again, it might reasonable in some situations to let $\mathbf{R}_{h}$ to be invariant (or perhaps to be constant over a peak period), thereby simplifying the process.

2.11.3 Setting up the historical database

So far, we have assumed the existence of a historical database of O-D matrices by departure time. This database would be constructed from results of estimations conducted (using possibly, the smoothing model in Section 2.9) in previous days. The database would be stratified by day-of-week, type of weather, special events, etc. Furthermore, results from the estimation of each day would be used to update the database. There could be different ways of carrying out this updating. The simplest technique is to use the latest available estimate (the estimate obtained during the last day) since this encapsulates all prior history. Another alternative could be to use a moving average of the last few estimates. A third alternative is to use a smoothing formula of the following form:

$$x_{rh}^{H,n} = x_{rh}^{H,n-1} + \alpha(\hat{x}_{rh}^{n} - x_{rh}^{H,n-1}) \tag{2.59}$$

where $x_{rh}^{H,n}$ represents the historical value corresponding to O-D pair r and departure interval h after n days, $\hat{x}_{rh}^{n}$ denotes the estimate on day n and α is a scalar between zero and one.

The final question that remains to be answered pertains to starting the process. For the first few days, the various inputs to the smoothing (or filtering) model such as the error-covariance matrices, the autocorrelation matrix, etc. are likely to be unknown or only approximately known. In such a situation, simpler models such as those proposed by Cascetta et al.[13] can be used. We start by describing these and then suggest some enhancements.

Cascetta et al. propose two estimators. In the first, the O-D flow matrices $\mathbf{x}_h^*$ are obtained sequentially from solving constrained optimization problems of the form:

$$\hat{\mathbf{x}}_h = argmin[f_1(\mathbf{x}_h, \mathbf{x}_h^a) + f_2(\mathbf{y}_h, \hat{\mathbf{y}}_h)] \tag{2.60}$$

over $\mathbf{x}_h > 0$. $\mathbf{x}_h$ is the current value of the demand vector (the variable over which the expression is optimized), $\mathbf{x}_h^a$ is an apriori or starting guess of $\mathbf{x}_h$ (can be obtained

by setting $\mathbf{x}_h^a = \hat{\mathbf{x}}_{h-1}$), $\mathbf{y}_h$ is the measured link counts while $\hat{\mathbf{y}}_h$ is the vector of counts obtained by assigning the decision variable $\mathbf{x}_h$.

f_1 and f_2 depend upon the estimation framework. For example, the GLS formulation of problem (2.60) would be

$$\begin{aligned}\hat{\mathbf{x}}_h &= argmin[(\mathbf{x}_h - \mathbf{x}_h^a)'\mathbf{W}_h^{-1}(\mathbf{x}_h - \mathbf{x}_h^a) + \\ &\quad (\mathbf{y}_h - \sum_{p=h-p'}^{h-1} \mathbf{a}_h^p \hat{\mathbf{x}}_p - \mathbf{a}_h^h \mathbf{x}_h)'\mathbf{R}_h^{-1}(\mathbf{y}_h - \sum_{p=h-p'}^{h-1} \mathbf{a}_h^p \hat{\mathbf{x}}_p - \mathbf{a}_h^h \mathbf{x}_h)] \end{aligned} \quad (2.61)$$

where the optimization is over $\mathbf{x}_h > 0$. $\mathbf{W}_h$ is the variance-covariance matrix of the vector of errors affecting the estimate $\mathbf{x}_h^a$. $\mathbf{R}_h$ is the variance-covariance matrix of the vector of measurement errors and can be obtained as described in Section 2.11.2. In the absence of any prior knowledge, Cascetta et al. suggest the use of identity matrices for the two – reducing the problem to a constrained OLS.

An alternative procedure suggested by them solves for the unknown O-D flows of *several* periods simultaneously. This is computationally expensive since it involves solution of a large optimization problem. The equivalent of equation (2.60) for obtaining the O-D flow vectors for N periods is given by:

$$\begin{aligned}(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_N) &= argmin[f_1(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N; \mathbf{x}_1^a, \mathbf{x}_2^a, ..., \mathbf{x}_N^a) + \\ &\quad f_2(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N; \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_N)] \end{aligned} \quad (2.62)$$

where the minimization is over $\mathbf{x}_i \geq 0 \quad \forall i = 1, 2, ..., N$. All terms in the above expression have the usual meaning. Again f_1 and f_2 depend upon the estimation framework with the GLS formulation given by:

$$\begin{aligned}(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_N) &= argmin \sum_{h=1}^{N} [(\mathbf{x}_h - \mathbf{x}_h^a)'\mathbf{W}_h^{-1}(\mathbf{x}_h - \mathbf{x}_h^a)] + \\ &\quad \sum_{h=1}^{N} [(\mathbf{y}_h - \sum_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p)'\mathbf{R}_h^{-1}(\mathbf{y}_h - \sum_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p)] \end{aligned} \quad (2.63)$$

where all terms have the usual meaning.

A comparison of the two estimators is useful. The sequential estimator has a very "local" outlook. As given by (2.61), it makes use only of the *first* measurement in estimating a particular O-D flow. For all subsequent estimation intervals, the O-D flow is held fixed at the previously estimated value – a procedure that is likely to introduce inaccuracies in the presence of large measurement errors. Indeed, this was the motivation behind the State Augmentation procedure described in Section 2.6. However, it has attractive computational features and in the absence of good apriori O-D information, can be used to provide a starting value during each interval. The simultaneous estimator is obviously more costly computationally, however, it makes use of more information because the entire set of counts is used to estimate each O-D vector.

In light of the above discussion, we define a *rolling-horizon* procedure that attempts to combine the advantages of the sequential and simultaneous estimators as follows:

$$\mathbf{X}_h = \begin{bmatrix} \mathbf{x}'_h & \mathbf{x}'_{h-1} & \ldots & \mathbf{x}'_{h-p'} \end{bmatrix}'$$

$$\mathbf{A}_h = \begin{bmatrix} \mathbf{a}_h^h & \mathbf{a}_h^{h-1} & \ldots & \mathbf{a}_h^{h-p'} \end{bmatrix}$$

$$\mathcal{W}_h = Var(\mathbf{X}_h^a)$$

The GLS formulation corresponding to (2.61) would then entail solving the following optimization problem for $\mathbf{X}_h$ during each interval.

$$\begin{aligned} \hat{\mathbf{X}}_h \;=\; & argmin[(\mathbf{X}_h - \mathbf{X}_h^a)'\mathcal{W}_h^{-1}(\mathbf{X}_h - \mathbf{X}_h^a) + \\ & (\mathbf{y}_h - \mathbf{A}_h\mathbf{X}_h)'\mathbf{R}_h^{-1}(\mathbf{y}_h - \mathbf{A}_h\mathbf{X}_h)] \end{aligned} \tag{2.64}$$

As in the sequential estimator, $\mathbf{X}_h^a$ can be obtained by the relationship $\mathbf{X}_h^a = \hat{\mathbf{X}}_{h-1}$. However, we now compute simultaneously, the O-D vectors over $(p'+1)$ intervals – $\mathbf{X}_h$ now represents the augmented O-D vector. Since the last block of n_{OD} elements of $\hat{\mathbf{X}}_h$ makes use of the most information in estimating $\mathbf{X}_{h-p'}$, it is likely to be statistically

the most efficient estimate of the latter. If additional measurements are available, we add more quadratic terms to (2.64).

Using Equation (2.64) requires calibration of $\mathcal{W}_h^{-1}$ and $\mathbf{R}_h^{-1}$. Wherever possible, these might be obtained from residuals of previous days (as in Section 2.11). In the absence of any prior information, we can use an FGLS procedure[22]. This involves initial OLS estimation (i.e. setting the variances to identity matrices) and computation of the associated residuals (separately for each type of measurement – counts, prior O-D flows, etc.). We next separate the residuals into groups (We hypothesize that variances are constant within each group). One possible criterion for grouping is the size of the fitted values, for example, we could divide the residuals for counts into three groups – low, medium, and high, based on the values of fitted counts. We use the residuals within each group to compute a common variance for that group. These variances are finally used to construct $\mathcal{W}_h^{-1}$ and $\mathbf{R}_h^{-1}$.

We conclude by observing that situations where a state-space model cannot be formulated (due to the absence of a dynamic relationship between O-D flows in the model formulation) can be handled in a straightforward way with the various GLS estimators proposed in this section.

2.12 Conclusion

In this chapter, we have presented the basic structure of the model system. The main conclusion in this chapter is that the dynamic O-D estimation and prediction problem can be viewed as one of integrating multiple sources of information in a consistent and optimal manner. In subsequent chapters, we develop enhancements of this framework and discuss implementation issues.

Chapter 3

Alternate Formulation

In this chapter we discuss an alternate approach to the estimation and prediction problem. This approach is based on the fact that an O-D flow can be decomposed into a departure rate from its origin and a split fraction corresponding to its destination. Each of these two quantities exhibits spatial and temporal variation. In this approach, we attempt to capture the heterogeneity in temporal variation exhibited by these two processes in order to improve the statistical efficiency of the models presented in Chapter 2. This approach is more in line with those proposed by other researchers for closed networks (refer to Section 1.2.1) with one significant difference. In our approach, we explicitly estimate (and predict) the departure rates from each origin while all the previous models (reviewed in Section 1.2.1) assume these to be known inputs.

We begin the chapter with an examination of empirical evidence on the temporal evolution of departure rates and split fractions (shares).

3.1 Stability of "Shares"

The number of trips between O-D pair (i, j) departing i in time interval h can be written as follows:

$$(Number\_of\_trips\_i\_j)_h = (Total\_trips\_from\_i)_h * (Share)_{i\_j,h}$$

Inaudi[25] observed that the shares of each O-D pair in the above expression remained stable over the course of a day relative to the departing trips. We attempted to verify this observation using data on actual time-dependent O-D flows that we obtained for the Massachusetts Turnpike – a 120-mile long expressway that runs east-west from New York state to Boston[1].

The first three blocks of figure 3-1 show the variation of total trips from Framingham and the variation of shares to each of the two destinations connected to Framingham. It can be seen that while the total number of trips changes quite substantially over the four hour period (going from 250 to 500 and back to 100), the share going to Weston does not display any systematic or substantial change – infact it only varies from about 0.97 to a minimum of just under 0.9. The coefficient of variation (standard deviation/mean) for trips is about 0.47 while that for shares to Weston is 0.03. A similar effect can be found in the variation of total trips from Westborough and the shares thereof (shown in the next four blocks of figure 3-1) with the coefficients of variation for trips from Westborough and shares to Weston, 0.4 and 0.09 respectively. These figures support the hypothesis that the shares are more stable with time compared to the total departing trips[2]. In following sections, we present a model that takes advantage of this differential variation.

3.2 Definitions

To formalize the approach, define o_{ih} as the number of trips emanating from origin i during interval h and the corresponding $(n_O * 1)$ vector by $\mathbf{o}_h$ where n_O is the number of origins in the network. Denote the rth share[3] for departures from the origin during interval h by ψ_{rh} and the corresponding $(n_{OD} * 1)$ vector by $\mathbf{\Psi}_h$. The new formulation would then involve estimation of $\mathbf{o}_h$ and $\mathbf{\Psi}_h$ during each interval h.

[1] We describe this data in detail in Chapter 5.

[2] We are interested primarily in the larger shares.

[3] Since there is a one to one correspondence between shares and O-D pairs, we use a single subscript for each share for notational simplicity.

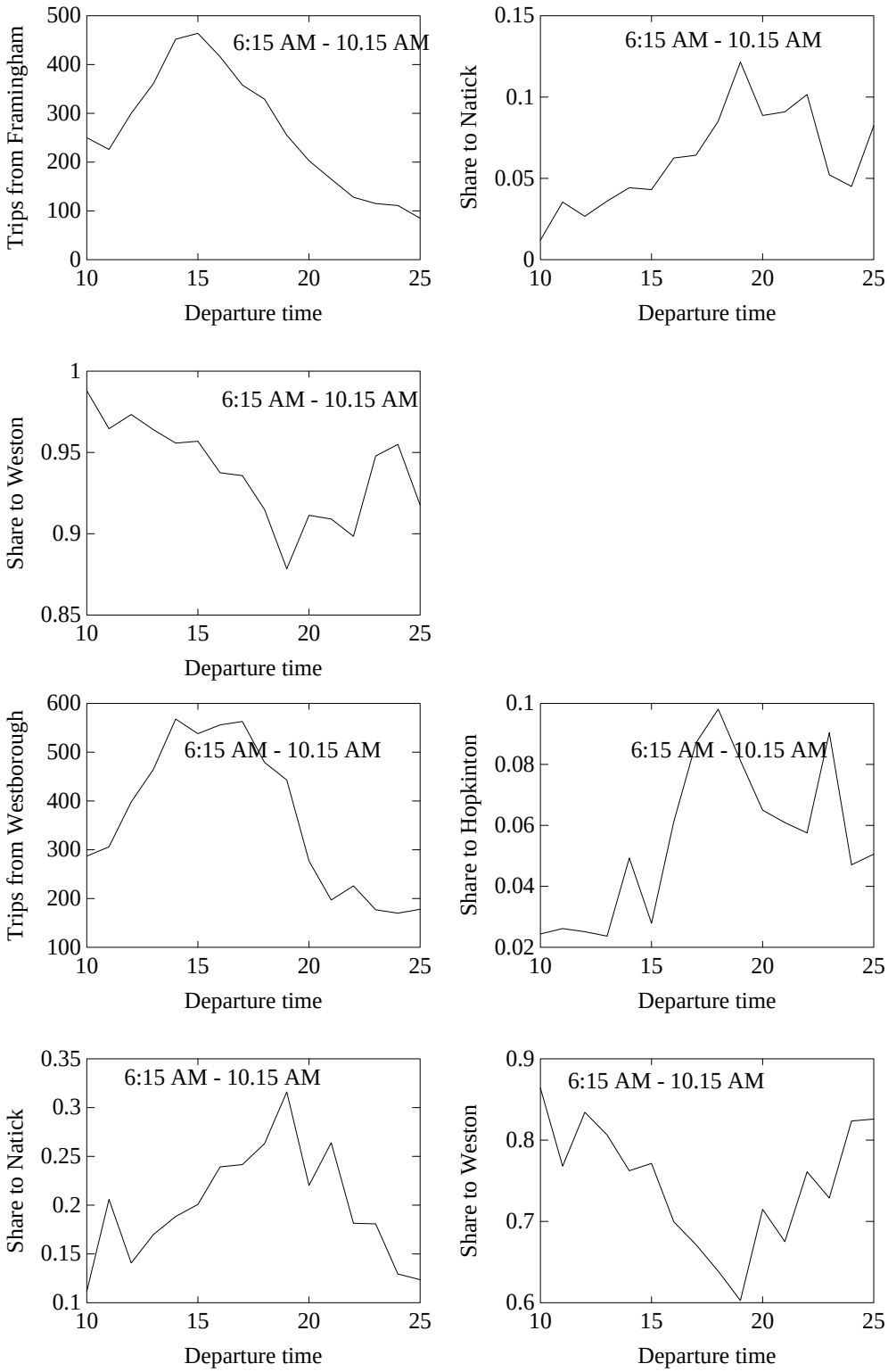

Figure 3-1: Stability of shares with time

3.3 Model Formulation

Direct and indirect measurements for this model would be in terms of trips and shares. Moreover, we wish to work in terms of *deviations* of trips and shares for similar reasons as stated in Chapter 2 for the O-D flows.

We express the direct measurements as follows:

$$\mathbf{o}_h^a = \mathbf{o}_h + \mathbf{u}_h^o \tag{3.1}$$

$$\mathbf{\Psi}_h^a = \mathbf{\Psi}_h + \mathbf{u}_h^\psi \tag{3.2}$$

where $\mathbf{o}_h^a$ and $\mathbf{\Psi}_h^a$ denote the preliminary estimates of $\mathbf{o}_h$ and $\mathbf{\Psi}_h$ respectively and $\mathbf{u}_h^o$ and $\mathbf{u}_h^\psi$ represent $(n_O * 1)$ and $(n_{OD} * 1)$ random errors[4].

Just as there were many ways of specifying $\mathbf{x}_h^a$ in Equation (2.1), there are many ways of specifying $\mathbf{o}_h^a$ and $\mathbf{\Psi}_h^a$. Based on the arguments in Chapter 2, we choose the following representation:

$$\mathbf{o}_h^a = \mathbf{o}_h^H + \Sigma_{p=h-q'_o}^{h-1} \mathbf{\Delta}_h^p (\hat{\mathbf{o}}_p - \mathbf{o}_p^H) \tag{3.3}$$

$$\mathbf{\Psi}_h^a = \mathbf{\Psi}_h^H + \Sigma_{p=h-q'_\psi}^{h-1} \mathbf{\Upsilon}_h^p (\hat{\mathbf{\Psi}}_p - \mathbf{\Psi}_p^H) \tag{3.4}$$

where $\mathbf{o}_h^H$ and $\mathbf{\Psi}_h^H$ are the best historical estimates for trips and shares, $\mathbf{\Delta}_h^p$ and $\mathbf{\Upsilon}_h^p$ are autocorrelation matrices for trips and shares deviations analogous to the $\mathbf{f}_h^p$ matrices in Chapter 2, and q'_o and q'_ψ are the orders of the two autoregressive processes. By specifying a different dynamic process for shares and trips in Equations (3.3) and (3.4), we attempt to capture their differential temporal variation.

Specification of indirect measurements is more complicated. Since an indirect measurement involves a mapping between O-D flows and link counts and the former involves a product of trips and shares, it is non-linear in the state variables. In matrix form, it can be stated as follows:

[4] We could represent the same in deviation form.

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{\Xi}_p \mathbf{o}_p + \mathbf{v}_h \tag{3.5}$$

In the above equation, $\mathbf{\Xi}_p$ is a $(n_{OD} * n_O)$ matrix in which each row has exactly *one* non-zero element corresponding to one O-D pair r[5]. All other terms have the same meaning as before. We note that we could rewrite the above equation in deviation terms on the same lines as in Chapter 2.

The above model can be equivalently represented by a state-space formulation where the state comprises the trip and share deviations. The formulation involves two sets of transition equations as follows:

$$\mathbf{o}_{h+1} - \mathbf{o}_{h+1}^H = \Sigma_{p=h+1-q'_o}^{h} \mathbf{\Delta}_{h+1}^p (\mathbf{o}_p - \mathbf{o}_p^H) + \mathbf{w}_{h+1}^o \tag{3.6}$$

$$\mathbf{\Psi}_{h+1} - \mathbf{\Psi}_{h+1}^H = \Sigma_{p=h+1-q'_\psi}^{h} \mathbf{\Upsilon}_{h+1}^p (\mathbf{\Psi}_p - \mathbf{\Psi}_p^H) + \mathbf{w}_{h+1}^\psi \tag{3.7}$$

where $\mathbf{w}_{h+1}^o$ and $\mathbf{w}_{h+1}^\psi$ are error vectors of dimensions $(n_O * 1)$ and $(n_{OD} * 1)$ respectively and all other definitions are as before. The measurement equation for this formulation is identical to (3.5).

We make the usual assumptions of zero mean and no serial correlation on errors $\mathbf{w}_h^o$, $\mathbf{w}_h^\psi$ and $\mathbf{v}_h$. In addition, we assume that $\mathbf{w}_h^o$ and $\mathbf{w}_h^\psi$ are uncorrelated with $\mathbf{v}_h$.

Just as in Chapter 2, we could represent Equations (3.3) and (3.4) (or equivalently (3.6) and (3.7)) more compactly by defining an augmented state consisting of lagged trip and share deviations.

Estimation of the matrices $\mathbf{\Delta}_{h+1}^p$, $\mathbf{\Upsilon}_{h+1}^p$ as well as the error covariance matrices $\mathbf{Q}_h^o$, $\mathbf{Q}_h^\psi$ and $\mathbf{R}_h$ that represent the variances of errors $\mathbf{w}_h^o$, $\mathbf{w}_h^\psi$ and $\mathbf{v}_h$ respectively, is performed in a manner identical to the procedure described in Section 2.11. Note that we now have separate regressions for the trips and shares.

We next describe the estimation and prediction methodology.

[5]It can be seen that there is a unique mapping between the vector $\mathbf{\Psi}$ and the matrix $\mathbf{\Xi}$. Knowledge of either implies that the other can be constructed.

3.4 Estimation and Prediction

The estimation methodology that we present in this section makes use of the approximation suggested in Section 2.10, i.e., each trip and share deviation is estimated exactly once. One could easily write similar equations for the case when the state is augmented to include lagged trip and share deviations.

One of the most popular ways to tackle the problem of non-linear estimation in dynamic systems has been to use the *Extended* Kalman Filter (EKF) algorithm (See for example Gelb[20]). This involves a first-order Taylor linearization of the measurement equation about the best available estimate of the state vector. The resulting update equations for the filter closely resemble those of the conventional Kalman Filter. Estimates obtained from the EKF could be improved by performing successive iterations of linearization and re-estimation leading to an *Iterated* EKF[6].

We summarize below, the EKF solution for the approximate model[7]:

$$\mathbf{\Sigma}_{0|0} = \begin{bmatrix} \mathbf{P}_0^o & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_0^{\psi} \end{bmatrix} \tag{3.8}$$

$$\mathbf{\Sigma}_{h|h-1} = \begin{bmatrix} \mathbf{\Delta}_h^{h-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Upsilon}_h^{h-1} \end{bmatrix} \mathbf{\Sigma}_{h-1|h-1} \begin{bmatrix} \mathbf{\Delta}_h^{h-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Upsilon}_h^{h-1} \end{bmatrix}' + \begin{bmatrix} \mathbf{Q}_h^o & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_h^{\psi} \end{bmatrix} \tag{3.9}$$

$$\mathbf{K}_h = \mathbf{\Sigma}_{h|h-1}\mathbf{D}_h'(\mathbf{D}_h\mathbf{\Sigma}_{h|h-1}\mathbf{D}_h' + \mathbf{R}_h)^{-1} \tag{3.10}$$

$$\mathbf{\Sigma}_{h|h} = \mathbf{\Sigma}_{h|h-1} - \mathbf{K}_h\mathbf{D}_h'\mathbf{\Sigma}_{h|h-1} \tag{3.11}$$

$$\hat{\mathbf{z}}_{0|0} = \begin{bmatrix} E(\bar{\mathbf{o}}_0 - \mathbf{o}_0^H) \\ E(\bar{\mathbf{\Psi}}_0 - \mathbf{\Psi}_0^H) \end{bmatrix} \tag{3.12}$$

$$\hat{\mathbf{z}}_{h|h-1} = \begin{bmatrix} \mathbf{\Delta}_h^{h-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Upsilon}_h^{h-1} \end{bmatrix} \hat{\mathbf{z}}_{h-1|h-1} + \begin{bmatrix} \Sigma_{p=h-q_o'}^{h-2} \mathbf{\Delta}_h^p \hat{\mathbf{o}}_p \\ \Sigma_{p=h-q_\psi'}^{h-2} \mathbf{\Upsilon}_h^p \hat{\mathbf{\Psi}}_p \end{bmatrix} \tag{3.13}$$

$$\hat{\mathbf{z}}_{h|h} = \hat{\mathbf{z}}_{h|h-1} + \mathbf{K}_h(\mathbf{y}_h - \mathbf{a}_h^h \hat{\mathbf{\Xi}}_{h|h-1}\hat{\mathbf{o}}_{h|h-1} - \Sigma_{p=h-p'}^{h-1}\mathbf{a}_h^p \hat{\mathbf{\Xi}}_{p|p}\hat{\mathbf{o}}_{p|p}) \tag{3.14}$$

$$h = 1, 2, ...$$

[6] An excellent discussion of the EKF as well as other non-linear filtering techniques may be found in Gelb[20].

[7] We assume, in addition, that the initial state is uncorrelated with the transition and measurement errors.

where the modified state vector $\mathbf{z}$ consists of deviations in both trips and shares, the initial state has mean $\begin{bmatrix} E(\bar{\mathbf{o}}_0 - \mathbf{o}_0^H) \\ E(\bar{\mathbf{\Psi}}_0 - \mathbf{\Psi}_0^H) \end{bmatrix}$ and variance $\begin{bmatrix} \mathbf{P}_0^o & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_0^\psi \end{bmatrix}$, $\mathbf{D}_h$ denotes the matrix of first derivatives $\frac{\partial(\mathbf{a}_h^h \mathbf{\Xi}_h \mathbf{o}_h)}{\partial \mathbf{z}}$ evaluated at $\hat{\mathbf{z}}_{h|h-1}$, $\mathbf{\Sigma}_{h|h-1}$ and $\mathbf{\Sigma}_{h|h}$ denote the variances of the estimates $\hat{\mathbf{z}}_{h|h-1}$ and $\hat{\mathbf{z}}_{h|h}$, and $\mathbf{Q}_h^o$ and $\mathbf{Q}_h^\psi$ denote the variance-covariance matrices of the error terms $\mathbf{w}_h^o$ and $\mathbf{w}_h^\psi$ respectively. We observe that Equations (3.8)–(3.14) are identical in spirit (and interpretation) to those for a linear Kalman Filter ((2.51)–(2.55)) except for the presence of the first derivative $\mathbf{D}_h$. We remark, in this regard, that the EKF process is suboptimal since it replaces the non-linearity by a linearization that is only approximate. However, it is a computationally efficient algorithm and hence is widely used when the type of non-linearity is simple (such as in our case, where the non-linearity arises from a product).

A k step prediction of O-D flows involves three steps – a k step prediction of trip and share deviations using (3.13) k times, adding to these deviations the corresponding historical estimates for interval $h + k$ and finally, constructing the predicted O-D flows by multiplying the predicted trips and shares.

3.5 Comments

A comparison of the trip/share approach with that of the previous chapter reveals that each has its advantages. The trip/share approach clearly provides a better way of modeling the dynamic evolution of O-D flows. This could result in a model with better predictive capabilities. On the other hand, it suffers from several disadvantages. First, the model is non-linear and the EKF procedure involves approximations because of the linearization. There is also the related issue of extra computational requirements – especially for the iterated version of the EKF. The final issue with the trip/share model comes from realizing that it makes no attempt to satisfy the natural constraints that for each origin, shares headed toward each destination should be between zero

and one and that the shares should sum up to one[8]. For these reasons, it is difficult to say categorically, which is a "better" approach.

3.6 Conclusion

We have presented a different formulation in this chapter that models departure rates and destination shares separately. Arguably, such a formulation could result in a model with better predictive capabilities. In Chapter 5, we compare the performance of this model with that presented in Chapter 2.

[8]Empirical evidence (Chapter 5) however indicated that it was rare for an estimated share to be negative or greater than one.

Chapter 4

The Assignment Matrix

In previous chapters, we have talked briefly about the assignment matrix and its key role in the O-D estimation and prediction problem. While the role of this matrix in static estimation is well understood and modeled, the same cannot be stated in the dynamic context. We devote this chapter to a detailed examination of this matrix and of how errors in this matrix can be explicitly captured in the model formulation.

We start with a description of the assignment matrix as a function of link travel times and route choice fractions. We then focus our attention on situations where the assignment matrix is imperfectly known or endogenous. These could occur for example, when travel times are unobserved, or are subject to large measurement errors. We finally discuss modeling strategies that incorporate the effect of an imperfect or a stochastic assignment matrix into the O-D estimation and prediction process.

4.1 Parameterizing the Assignment Matrix

Let each O-D pair r be connected by a set of paths $\mathcal{K}_r$. Assume that there exist in total K paths between the n_{OD} O-D pairs in the network i.e. $K = \|\mathcal{K}_1 \cup \mathcal{K}_2 \cup ... \cup \mathcal{K}_{n_{OD}}\|$. Each path $k = 1, 2, .., K$ corresponds to a unique O-D pair. Denote by L_k the set of links comprised in path k. Finally, denote by $\mathcal{F}_h^k$ the flow along path k departing the

origin in interval h. Thus during any interval h, the following relationship holds:

$$x_{rh} = \Sigma_{k \in \mathcal{K}_r} \mathcal{F}_h^k \tag{4.1}$$

Let q_{kh} denote the fraction of travelers corresponding to O-D pair r and departure interval h that choose path k, with $\Sigma_{k \in \mathcal{K}_r} q_{kh} = 1 \ \ \forall \ \ r, h$. We then get the following relationship between O-D and path flows:

$$\mathcal{F}_h^k = x_{rh} q_{kh} \tag{4.2}$$

Recognizing that the link flows are comprised of contributions from many different path flows, Equation (2.12) can be restated in terms of path flows as follows:

$$y_{lh} = \Sigma_{p=h-p'}^{h} \Sigma_{k=1}^{K} \alpha_{lh}^{kp} \mathcal{F}_p^k + v_{lh} \tag{4.3}$$

where α_{lh}^{kp} defines a mapping between path and link flows and is defined as the contribution of the kth path flow departing the origin during interval p towards the flow across detector l during interval h[1].

Finally, as shown by Cascetta et al.[13], a simple manipulation of equations (2.12), (4.2) and (4.3) yields the following expression for the assignment matrix:

$$a_{lh}^{rp} = \Sigma_{k:k \in \mathcal{K}_r} \alpha_{lh}^{kp} q_{kp} \tag{4.4}$$

Analytical expressions for the link-path incidence fractions can be obtained using information about link travel times. In addition to these travel times however, an assumption about movement of vehicles through the network is required. For example, Cascetta et al.[13] derive expressions based on the assumption that vehicles within a group (k, p) (henceforth referred to as a *packet*) are uniformly comprised within the departure duration H and stay within this interval as they move across the network. In other words, vehicles within a packet are uniformly distributed between the leader

[1] In conventional static estimation, this would be either one or zero. A matrix of these fractions is the familiar link-path incidence matrix.

and the last follower over a span of time H. This assumption can be easily relaxed to permit the effects of "stretching" and "squeezing" of packets as they traverse the network. Such effects could be significant for example, if trip durations are relatively large or travel time variations across successive time-intervals are significant. Under this situation, the link-path incidence fractions would be given by the following expression:

$$\begin{aligned}
\alpha_{lh}^{kp} &= 1 && \text{if } (h-1)H < \eta_{1l}^{kp} < \eta_{2l}^{kp} < hH \\
&= (hH - \eta_{1l}^{kp})/(\eta_{2l}^{kp} - \eta_{1l}^{kp}) && \text{if } (h-1)H < \eta_{1l}^{kp} < hH < \eta_{2l}^{kp} \\
&= H/(\eta_{2l}^{kp} - \eta_{1l}^{kp}) && \text{if } \eta_{1l}^{kp} < (h-1)H < hH < \eta_{2l}^{kp} \\
&= (\eta_{2l}^{kp} - (h-1)H)/(\eta_{2l}^{kp} - \eta_{1l}^{kp}) && \text{if } \eta_{1l}^{kp} < (h-1)H < \eta_{2l}^{kp} < hH \\
&= 0 && \text{otherwise}
\end{aligned} \tag{4.5}$$

where η_{1l}^{kp} and η_{2l}^{kp} represent the crossing times of the first and last vehicle in the packet (k,p) at detector l. To use the above relationship, one would in addition have to know the departure times of the first and last vehicles. A convenient assumption might be to have the first depart at the beginning of a departure interval and the last at the end.

Travel times (or more typically, speeds) can be obtained either from a traffic surveillance system (e.g. sensors on the roadway, video cameras, probe vehicles) or a DTA model. In addition to link and path travel times, information about the path choice fractions q_{kh} is required in order to apply equation (4.4). One way of obtaining these is by using discrete choice models that utilize information about generalized costs along different paths during each interval. An example of such a model is provided by Cascetta et al.[14].

To summarize, computation of the assignment matrix is highly complicated. Moreover, the estimates obtained from application of equations (4.4) and (4.5) may suffer from errors on several fronts

- Travel times obtained from the surveillance system are subject to measurement

error due to, for example, sensor malfunction.

- The assumption of uniform distribution of vehicles within a packet might be invalid under certain situations e.g. an incident.
- Choice fractions obtained from route choice models might be erroneous because of inaccuracies either in the model coefficients or in the data.
- True departure times of first and last vehicles are unknown.

Finally, there could be scenarios in which some (or all) travel times might be entirely unobserved (are endogenous). These are explored in greater detail in following sections.

4.2 Endogeneity in the Assignment Matrix

We turn our attention now to the case of erroneous travel times resulting in an imperfect assignment matrix. In such a case, we can write:

$$\mathbf{a}_h^{p\bullet} = \mathbf{a}_h^p + \nu_h^p \tag{4.6}$$

where $\mathbf{a}_h^{p\bullet}$ denotes the erroneous assignment matrix and ν_h^p a random error. Equation (2.13) becomes

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^{p\bullet} \mathbf{x}_p - \nu_h^p \mathbf{x}_h + \mathbf{v}_h \tag{4.7}$$

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^{p\bullet} \mathbf{x}_p + \tilde{\mathbf{v}}_h \tag{4.8}$$

where the new error term $\tilde{\mathbf{v}}_h = \mathbf{v}_h - \nu_h^p \mathbf{x}_h$. Clearly, $\mathbf{a}_h^{p\bullet}$ and $\tilde{\mathbf{v}}_h$ are correlated because of Equation (4.6). Thus, applying a GLS estimator of the form (2.35) to (4.8) yields biased and inconsistent estimates for $\mathbf{x}_h$[2]. Depending on the level of uncertainty in the assignment matrix, the extent of bias might be significant.

[2]This constitutes a standard error-in-variables problem in econometrics (Greene[22]).

The impact of this inconsistency could be even more severe if the O-D flows and assignment matrix were to be obtained by an iterative scheme[3]. This could be the case, for example, if the travel times were entirely unobserved and the assignment fractions were obtained by applying the following series of steps: (a) Load a preliminary set of O-D flows to a traffic simulator. (b) Use the observed assignment matrix to recompute the O-D flows. (c) Repeat (a) using updated O-D flows. In such a situation, even if convergence in O-D flows and assignment fractions were to be attained[4], the final estimates could be highly biased since errors in the assignment matrix are not explicitly accounted for, during each iteration. It is likely, of course, that these iterations may reduce the error in the assignment matrix.

This brings us to the central theme of this chapter. In the following sections, we describe two approaches that explicitly take into account the stochasticity of the assignment matrix and accordingly modify the formulations developed in previous chapters.

4.3 Modeling a Stochastic Assignment Matrix

In the first approach, we envisage adding randomness to the assignment matrix by means of additional measurement equations of the following form:

$$\mu(\mathbf{a}_h^{p\bullet}) = \mu(\mathbf{a}_h^{p}) + \mu(\nu_h^{p}) \tag{4.10}$$

In the above equations, $\mathbf{a}_h^{p\bullet}$ is the (erroneous) value of the assignment matrix computed from equations (4.4) and (4.5) using measured or estimated travel times

[3]In such a situation, the assignment fractions are endogenous and indirectly depend on the O-D flows. To see this dependence, we first notice that link travel times depend directly upon link flows. The latter is related to path flows (since a link flow is essentially a weighted combination of several path flows using that link). Finally, the path flows are related to the O-D flows through equation (4.2). Thus, equation (2.13) should be expressed as:

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h} \mathbf{a}_h^{p}(\mathbf{x}_h, \mathbf{x}_{h-1}, ..., \mathbf{x}_{h-p'})\mathbf{x}_p + \mathbf{v}_h \tag{4.9}$$

[4]which in itself is far from guaranteed

and route choice fractions. $\mu(.)$ is an operator that re-arranges the elements of a $(m*n)$ matrix into a $(mn*1)$ vector. The error term ν_h^p reflects the fact that the estimate $\mathbf{a}_h^{p\bullet}$ is subject to error. As mentioned earlier, this error could arise either because of imperfect measurements/estimates of travel times and route-choice fractions or because of incorrectness of assumptions involved in equations (4.5).

Equation (4.10) represents a straightforward approach of incorporating stochasticity in the assignment matrix. The disadvantage of this method is the additional computational load imposed by adding a large number of decision variables $\{\mathbf{a}_h^p\}$. In practical applications therefore, it may be necessary to prune the number of assignment fractions that are considered random in order to make the technique computationally tractable. For example, one might wish to add additional equations (4.10) only for assignment fractions corresponding to O-D pairs with high flows.

Deeper examination of the assignment matrix and its dependence on travel times and route-choice fractions as given by equations (4.4) and (4.5) suggests an alternative approach. Define the assignment matrix $\mathbf{a}_h^p$ by the following relationship:

$$\mathbf{a}_h^p = a(\mathbf{t}_h, \mathbf{t}_{h-1}, ..., \mathbf{t}_{h-p'}, \mathbf{q}_p) \tag{4.11}$$

where $\mathbf{t}_h$ denotes a $(n_{LK}*1)$ vector of travel times for interval h, $\mathbf{q}_p$ a $(K*1)$ vector of route choice fractions for departure time-interval p and the function $a(.)$ defined by equations (4.4) and (4.5)[5]. The above equation can be compactly represented as:

$$\mathbf{a}_h^p = a(\mathbf{T}_h, \mathbf{q}_p) \tag{4.12}$$

where $\mathbf{T}_h$ denotes the augmented vector $[\mathbf{t}_h \mathbf{t}_{h-1} ... \mathbf{t}_{h-p'}]'$.

If we assume that the above relationship is exact, the only remaining sources of errors in the assignment matrix are those in $\mathbf{T}_h$ and $\mathbf{q}_p$. This provides the motivation for the second approach. Instead of directly dealing with the assignment matrix

[5]Note that $a(.)$ could also in general be a function of *future* travel times. This dependence comes from equation (4.5) as well as the fact that the $\mathbf{q}$'s could depend on future travel times. The formulation to be presented captures errors in future travel times through stochasticity of the $\mathbf{q}$'s.

fractions as in the earlier approach, we work with the underlying travel times and route-choice fractions in the second. We thus specify measurement equations of the following form:

$$\mathbf{T}_h^{\bullet} = \mathbf{T}_h + \mathbf{\Lambda}_h^T \tag{4.13}$$

$$\mathbf{q}_p^{\bullet} = \mathbf{q}_p + \mathbf{\Lambda}_p^q \tag{4.14}$$

where $\mathbf{T}_h^{\bullet}$ and $\mathbf{q}_p^{\bullet}$ denote the measured/estimated values of $\mathbf{T}_h$ and $\mathbf{q}_p$ and $\mathbf{\Lambda}_h^T$ and $\mathbf{\Lambda}_p^q$ represent vectors of random error terms.

Again for computational reasons, one might wish to treat some or all of the route choice fractions as fixed. Note that if the route choice fractions $\mathbf{q}_p$ can be explicitly represented as a function of travel times $\mathbf{t}_p$ (for example using a logit formula[6]), this relationship could be substituted into equation (4.12), and equations (4.14) are rendered unnecessary – travel times are the only additional variables to be estimated.

Both the approaches presented above have advantages and shortcomings. While the first approach is likely to be more computation intensive (due to the size of the assignment matrix), it could be useful in modeling situations where the assumptions behind equations (4.5) break down. For example, the assumption that vehicles within a packet stay uniformly distributed between a leader and follower might be violated on urban networks with traffic control signals. On the other hand, the second approach is more efficient because it benefits more directly from the information contained in equation (4.12). It could be used in freeways or on moderately congested arterials. It could also be more useful when specific information can be obtained on sensor errors, for example, when it is known that a particular sensor systematically overestimates or underestimates, say travel speeds.

[6]Such a model would use *path* travel times. If we assume additivity, these could be obtained from link travel times.

4.4 The Enhanced Model

4.4.1 Direct and Indirect Measurements

To extend the framework developed in Chapters 2 to incorporate the ideas described thus far in this chapter, we need to specify additional measurement equations. For the second approach, we augment the state by the travel times (or speeds) and route choice fractions and add new direct measurements as follows:

$$\mathbf{t}_h^\bullet = \mathbf{t}_h + \mathbf{\Lambda}_h^t \tag{4.15}$$

$$\mathbf{q}_h^\bullet = \mathbf{q}_h + \mathbf{\Lambda}_h^q \tag{4.16}$$

where $\mathbf{t}_h^\bullet$ and $\mathbf{q}_h^\bullet$ denote the measured/estimated values of $\mathbf{t}_h$ and $\mathbf{q}_h$ and $\mathbf{\Lambda}_h^t$ and $\mathbf{\Lambda}_p^q$ represent vectors of random error terms.

In addition, we hypothesize that the ratio of travel times (or speeds) and route-choice fractions over two successive intervals is stable on a day-to-day basis. This information can be represented in the form of additional direct measurements:

$$t_{ih}^a = t_{ih} + \gamma_{ih}^t \tag{4.17}$$

$$q_{ih}^a = q_{ih} + \gamma_{ih}^q \tag{4.18}$$

where

$$t_{ih}^a = \frac{t_{ih}^H}{t_{ih-1}^H}\hat{t}_{ih-1} \tag{4.19}$$

$$q_{ih}^a = \frac{q_{ih}^H}{q_{ih-1}^H}\hat{q}_{ih-1} \tag{4.20}$$

In the above equations, t_{ih} and q_{ih} denote the ith travel time and route-choice fraction, t_{ih}^a and q_{ih}^a denote preliminary estimates of these, and the superscript H, as always, indicates historical values. In matrix form, Equations (4.17) and (4.18) can

be represented as follows:

$$\mathbf{t}_h^a = \mathbf{t}_h + \mathbf{\Gamma}_h^t \tag{4.21}$$

$$\mathbf{q}_h = \mathbf{q}_h + \mathbf{\Gamma}_h^q \tag{4.22}$$

with

$$\mathbf{t}_h^a = \mathbf{M}_h^t \hat{\mathbf{t}}_{h-1} \tag{4.23}$$

$$\mathbf{q}_h^a = \mathbf{M}_h^q \hat{\mathbf{q}}_{h-1} \tag{4.24}$$

where $\mathbf{M}_h^t$ and $\mathbf{M}_h^q$ are diagonal matrices with the (i,i)th element given by (t_{ih+1}^H / t_{ih}^H) and (q_{ih+1}^H / q_{ih}^H) respectively.

Similarly for the first approach, we have two additional sets of direct measurements. The first is simply (4.10), i.e.,

$$\mu(\mathbf{a}_h^{p\bullet}) = \mu(\mathbf{a}_h^p) + \mu(\nu_h^p) \tag{4.25}$$

In addition, we hypothesize that the ratios $(a_{lh+1}^{rp+1} / a_{lh}^{rp})$ remain stable on a day-to-day basis yielding the familiar form:

$$\mu(\mathbf{a}_h^p)^* = \mu(\mathbf{a}_h^p) + \mathbf{\Gamma}_h^a \tag{4.26}$$

$$\mu(\mathbf{a}_h^p)^* = \mathbf{M}_h^a \mu(\hat{\mathbf{a}}_{h-1}^{p-1}) \tag{4.27}$$

$$\{\mathbf{M}_h^a\}_{ii} = [\mu(\hat{\mathbf{a}}_h^{pH})]_i / [\mu(\hat{\mathbf{a}}_{h-1}^{p-1H})]_i \tag{4.28}$$

where $\mu(\mathbf{a}_h^p)^*$ denotes the preliminary estimate and $\mathbf{M}_h^a$ is a diagonal matrix whose ith element is given by (4.28). $\mathbf{\Gamma}_h^a$ represents the random error while as before, the superscript H indicates a historical value.

4.4.2 State-Space formulation

Both of the above models can be expressed easily in state-space form. For the second approach, the state is now comprised of O-D deviations augmented by travel times and route-choice fractions. We represent (4.15) and (4.16) as additional measurement equations. The information in (4.17) and (4.18) can be expressed as transition equations as follows:

$$\mathbf{t}_{h+1} = \mathbf{M}^t_{h+1}\mathbf{t}_h + \mathbf{\Theta}^t_{h+1} \tag{4.29}$$

$$\mathbf{q}_{h+1} = \mathbf{M}^q_{h+1}\mathbf{q}_h + \mathbf{\Theta}^q_{h+1} \tag{4.30}$$

where $\mathbf{\Theta}^t_{h+1}$ and $\mathbf{\Theta}^q_{h+1}$ are error terms and all other definitions are as before. A complete specification of this approach, therefore, requires (2.50), (4.29) and (4.30) as transition equations and Equations (2.49), (4.15) and (4.16) as measurement equations.

Similarly for the first approach, the state is comprised of O-D deviations augmented by assignment fractions. The additional measurement equation is given by (4.10). The information in (4.26) can be expressed as a transition equation as follows:

$$\mu(\mathbf{a}^{p+1}_{h+1}) = \mathbf{M}^a_{h+1}\mu(\mathbf{a}^p_h) + \mathbf{\Theta}^a_h \tag{4.31}$$

where $\mathbf{\Theta}^a_h$ is the random error and other definitions are as before. A complete specification of this approach therefore involves Equations (2.50) and (4.31) as transition equations and Equations (2.49) and (4.10) as measurement equations.

4.4.3 Estimation and Prediction

Estimation and prediction of O-D flows and additional travel time or assignment parameters may be carried out as in Chapter 3 using the EKF, with the usual assumptions on the error terms. Additional input parameters $\mathbf{M}^t_h$, $\mathbf{M}^q_h$ and $\mathbf{M}^a_h$ may be calibrated from historical data in a similar fashion. The variances of the extra error terms can be calibrated from residuals corresponding to previous days in a

straightforward fashion using similar formulae as in Section 2.11.

4.4.4 Comments

We make an important observation. In this entire section, we have described "approximate" versions of stochastic assignment matrix models in the following sense. We can conceive of link counts during interval h providing indirect information not only about O-D flows corresponding to $h-1, h-2, ...h-p'$ but also about travel times and route-choice fractions $\mathbf{t}_h, \mathbf{t}_{h-1}, ..., \mathbf{t}_{h-p'}, \mathbf{q}_h, \mathbf{q}_{h-1}, ..., \mathbf{q}_{h-p'}$ (and hence about assignment matrices $\mathbf{a}_h^{h-1}, \mathbf{a}_h^{h-2}, ..., \mathbf{a}_h^{h-p'}$). Nothing precludes us from constructing an augmented model (just as in Section 2.6) to estimate during each interval, lagged travel times and route-choice fractions or assignment matrices. No doubt this would lead to more efficient models; however, the associated computational overheads could make the enhanced models intractable. For notational simplicity, we have avoided detailed exposition of such a model. For small networks, this could be useful.

4.5 Modified Offline Models

The models described in the previous section require historical data in order to calibrate the additional input parameters. Just as we described the use of the sequential estimator proposed by Cascetta et al.[13] in Section 2.11.3 in the context of calibrating input parameters for the models in Chapter 2, we present in this section a modification of the sequential estimator that incorporates a stochastic assignment matrix.

Consider the second of the two approaches suggested in Section 4.3. Since the sequential model only estimates O-D flows corresponding to one period and holds O-D flows corresponding to prior periods constant, equations (2.13), (4.13) and (4.14) are modified as follows:

$$\mathbf{y}_h = \Sigma_{p=h-p'}^{h-1} a(\hat{\mathbf{T}}_{h-1}, \mathbf{t}_h, \hat{\mathbf{q}}_p)\hat{\mathbf{x}}_p + a(\mathbf{t}_h, \mathbf{q}_h)\mathbf{x}_h + \mathbf{v}_h \quad (4.32)$$

$$\mathbf{t}_h^{\bullet} = \mathbf{t}_h + \mathbf{\Lambda}_h^t \quad (4.33)$$

$$\mathbf{q}_h^{\bullet} = \mathbf{q}_h + \mathbf{\Lambda}_h^q \tag{4.34}$$

where $\hat{\mathbf{x}}_p$, $\hat{\mathbf{T}}_p$, and $\hat{\mathbf{q}}_p$ denote estimates from prior intervals that are fixed during interval h.

A GLS based solution would then involve minimization of the following error criterion for each interval:

$$[\hat{\mathbf{x}}_h, \hat{\mathbf{t}}_h, \hat{\mathbf{q}}_h] = arg\ min\ [(\mathbf{x}_h - \hat{\mathbf{x}}_{h-1})' \mathbf{W}_h^{-1} (\mathbf{x}_h - \hat{\mathbf{x}}_{h-1}) + (\mathbf{y}_h - \mathbf{y}_h^{\bullet})' \mathbf{V}_h^{-1} (\mathbf{y}_h - \mathbf{y}_h^{\bullet}) + (\mathbf{t}_h - \mathbf{t}_h^{\bullet})' \mathbf{P}_h^{t^{-1}} (\mathbf{t}_h - \mathbf{t}_h^{\bullet}) + (\mathbf{q}_h - \mathbf{q}_h^{\bullet})' \mathbf{P}_h^{q^{-1}} (\mathbf{q}_h - \mathbf{q}_h^{\bullet})] \tag{4.35}$$

where $\mathbf{y}_h^{\bullet} = \Sigma_{p=h-p'}^{h-1} a(\hat{\mathbf{T}}_{h-1}, \mathbf{t}_h, \hat{\mathbf{q}}_p) \hat{\mathbf{x}}_p + a(\mathbf{t}_h, \mathbf{q}_h) \mathbf{x}_h$. $\mathbf{V}_h$, $\mathbf{P}_h^{t^{-1}}$ and $\mathbf{P}_h^{q^{-1}}$ represent the covariances of the error terms $\mathbf{v}_h$, $\mathbf{\Lambda}_h^t$ and $\mathbf{\Lambda}_h^q$ in equations (4.32), (4.33) and (4.34) respectively; their inverses as before reflect the degree of confidence placed on the various sources of information. $\mathbf{W}_h$ represents the covariance of the error in the prior estimate $\hat{\mathbf{x}}_{h-1}$ (we use the estimate for the previous interval $h-1$ as apriori value for interval h). The optimization would be subject to non-negativity constraints on $\mathbf{x}_h$, $\mathbf{t}_h$ and on the fact that $\Sigma_{k \in \mathcal{K}_r} q_{kp} = 1 \ \forall \ (r,p)$ and $0 \leq q_{kp} \leq 1 \ \forall \ (k,p)$.

Following a similar vein, the first approach suggested in Section 4.3 would involve minimization of the following error criterion:

$$\begin{aligned}[\hat{\mathbf{x}}_h, \hat{\mathbf{a}}_p^h] = arg\ min\ [&(\mathbf{x}_h - \hat{\mathbf{x}}_{h-1})' \mathbf{W}_h^{-1} (\mathbf{x}_h - \hat{\mathbf{x}}_{h-1}) \\ &+ (\mathbf{y}_h - \mathbf{y}_h^{\bullet})' \mathbf{V}_h^{-1} (\mathbf{y}_h - \mathbf{y}_h^{\bullet}) \\ &+ \Sigma_{p=h-p'}^{h} (\mu(\mathbf{a}_p^h) - \mu(\mathbf{a}_h^{p^{\bullet}}))' (\mathbf{P}_h^a)^{-1} (\mu(\mathbf{a}_p^h) - \mu(\mathbf{a}_h^{p^{\bullet}}))]\end{aligned} \tag{4.36}$$

where $\mathbf{P}_h^a$ is the covariance matrix for the error term ν_h^p in equation (4.10). $\mathbf{a}_h^{p^{\bullet}}$ represents the value of the assignment matrix computed from the equations (4.4) and (4.5) using measured or estimated travel times and route choice fractions[7].

[7] In Section 2.11.3, we described a modification of the sequential estimator to estimate, in addition,

4.6 Conclusion

In this chapter, we have described in detail, the role played by the assignment matrix in the O-D Estimation and Prediction problem. Recognizing the fact that in most practical situations, these matrices are likely to be, at best, imperfectly known, we have proposed two models that explicitly capture their stochasticity. Both the models fall naturally into the overall framework described in Chapters 1 and 2. In the next chapter, we evaluate the performance of each of the models developed in this thesis thus far.

several lagged O-D flows. In similar fashion, it is possible to modify error criteria (4.35) and (4.36) to construct a model with increased decision variables – travel times, route choice fractions or assignment fractions corresponding to prior intervals (See also Section 4.4.4). Again, to keep the notation simple, we choose not to provide a detailed treatment.

Chapter 5

Case Studies

In previous chapters, we have developed a suite of models for Dynamic O-D Estimation and Prediction. The objective of this chapter is to use actual traffic data to demonstrate the performance of the various models. While tests using real data are extremely useful, in most cases, the true O-D flows are unknown. In an attempt to derive further insights into performance characteristics of different models, we therefore generate synthetic traffic data to supplement that actually observed.

5.1 Data Description

For this research, we had available to us three different data sources. We describe briefly the main features of each.

5.1.1 The Massachusetts Turnpike

This stretch of I-90 from New York State to Weston (I-95) comprises a distance of about 120 miles with 15 entry/exit ramps. Any combination of an entry and exit ramp constitutes an acceptable O-D pair. Hence, the network has 210 possible O-D pairs. However, in the analysis, only East Bound traffic was considered; hence the number of O-D pairs was reduced by half. Data on traffic movements was available for three days.

Each line of the data file contained information specific to a particular vehicle. Information was available on entry and exit ramps of the vehicle, entry and exit times, vehicle type and transaction type (whether toll was paid in cash/non-cash). The information about the entry times of the vehicles, however, was not considered sufficiently accurate[1]. Hence, the entry times were back-calculated from the exit-times assuming a uniform average speed[2].

The advantage of this dataset is that true O-D flows can be computed – a luxury not available in a vast majority of applications.

5.1.2 I-880 near Hayward, California

The second dataset covered a 5.2 mile (NorthBound) stretch of I-880 near Hayward, California. This section had 4 on-ramps and 5 off ramps with 20 O-D pairs. Ten minute detector data on traffic volumes and average speeds was available at 10 detector locations for a 2.5 hour morning peak period. Data over seven days was available. The advantage of using this dataset was that unlike the Turnpike data, congestion level was heavy during certain time intervals with speeds reaching 15-20 mph at some locations. A schematic layout of this network is shown in Figure 5-1.

5.1.3 Amsterdam Beltway

This is a 32 km freeway encircling the city of Amsterdam with 20 entrance and exit ramps. The layout of this network is depicted in Figure 5-2. Both the datasets mentioned earlier, suffer from the limitation that they do not capture route choice. This dataset is intended to address this limitation by allowing for two routes between each O-D pair – clockwise and anticlockwise. Moreover, in this dataset, we use a combination of actual and synthetic data to gain additional insight into model

[1]The data was generated from tickets used during toll collection. As each vehicle enters the freeway, the driver is issued a ticket (at the entry ramp). Toll collectors are expected to punch the ticket just before handing it to the driver. This time is then recorded as the entry time of the vehicle. In reality however, during periods of heavy demand, toll collectors often "gang-punch" tickets in advance to help expedite the process of issuing. The entry times hence are generally unreliable.

[2]More on the speed assumption later.

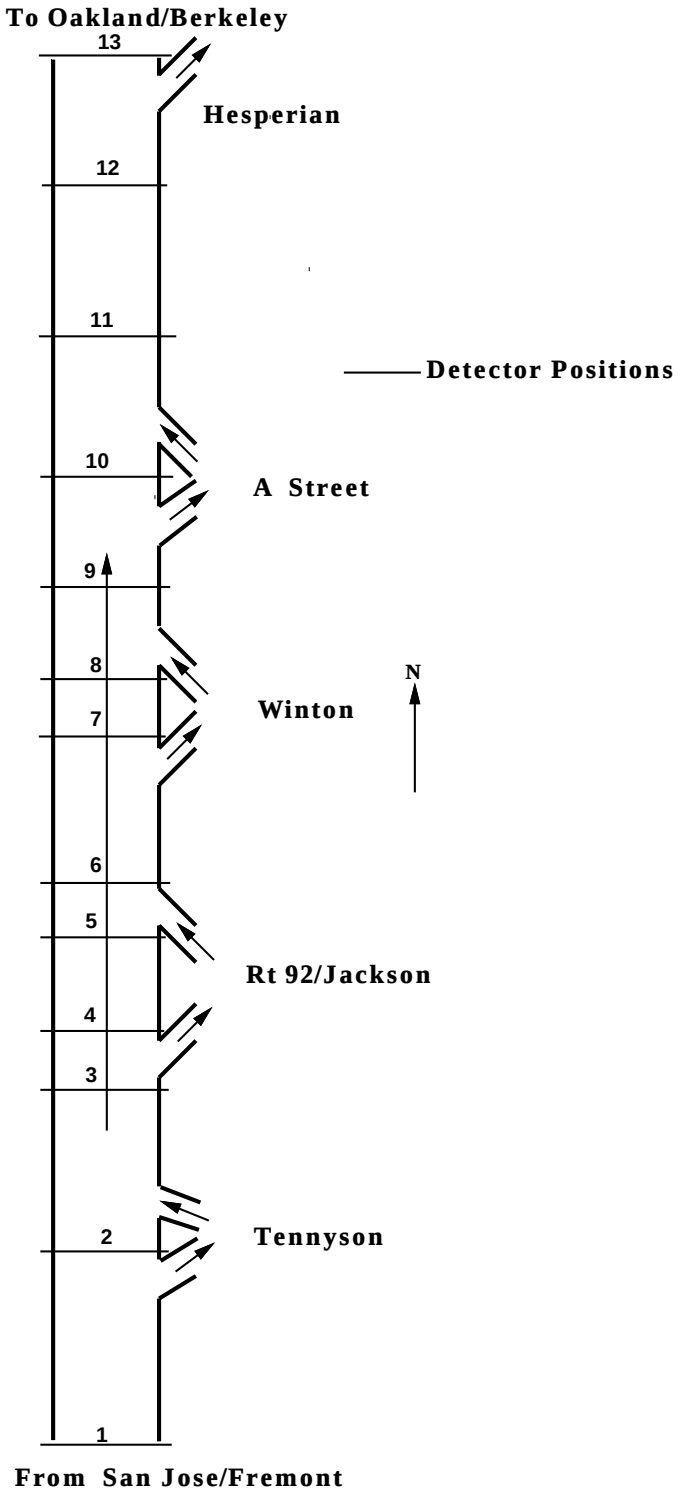

Figure 5-1: Section of I-880 North

performance. After removing sensors with large errors from the dataset, information on average speeds and link counts was available at 65 locations over one minute intervals for one day. Synthetic data was generated for an additional day using a process described in Section 5.2.3.

5.2 Implementing the models

5.2.1 The Massachusetts Turnpike

As mentioned earlier, the data-file consisted of information about every vehicle that used the turnpike on the three days. The period of analysis chosen was from 6:15 A.M. to 9:45 A.M. with the length of each departure interval chosen to be 15 minutes. Since the data was in disaggregate form, it had to be aggregated to obtain time-dependent traffic counts for each of the 14 links in the network. This aggregation was carried out by assuming an average vehicle speed of 55 mph and using this speed to calculate the entry time of each vehicle on each link[3]. Under the assumption that a counting station is located near the entrance to each link, the traffic counts for an interval represent the number of vehicles that enter each link during that interval. Hence, time dependent counts at these hypothetical counting stations could be easily abstracted from information about network-entering times and average speeds. Since to employ the Kalman Filter, the initial state of the system is required and since the initial state in our case encompasses several prior time-intervals (because of the transformation of variables), all vehicles that had entering times after 4.00 A.M. were processed. The values of the assignment matrices and counts were computed such that the relationship

$$y_{lh} = \sum_{p=h-p'}^{h} \sum_{r=1}^{n_{OD}} a_{lh}^{rp} x_{rp} \tag{5.1}$$

held exactly.

[3]This assumption of a uniform average speed for all vehicles at all times is unrealistic. However, for the purposes of implementing and evaluating the proposed model, all that is needed is a set of "reasonable" O-D matrices, counts and assignment matrices consistent with each other. The issue of whether the speeds assumed for generating counts is realistic is not directly relevant.

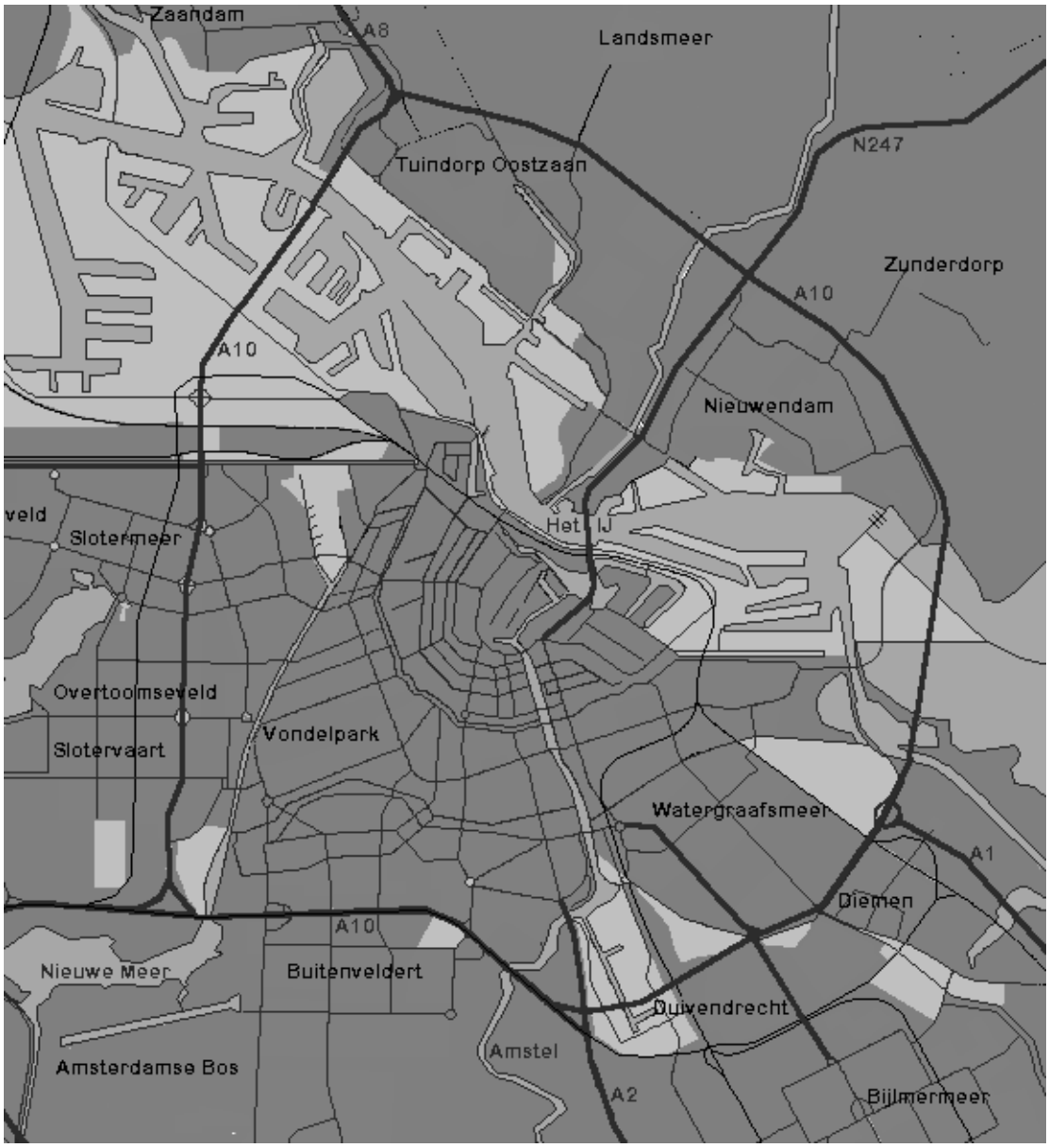

Figure 5-2: The Amsterdam Beltway (A10)

The details are as follows. All the variables were initialized to zero. The data file was processed line by line. Assuming the origin and destination of the vehicle currently being processed corresponded to the O-D pair r and that its entrance time was comprised in the interval p, the variable x_{rp} was incremented by one. Then, the movement of the vehicle was tracked using the constant speed assumption. Since the network was linear, no path choice model was required. For each link l in the path of the vehicle the entering time was computed. Assuming that this time fell in the interval h, the variables y_{lh} and a_{lh}^{rp} were incremented by one[4]. When all the vehicles entering the network during the interval p were treated, the variables a_{lh}^{rp} were divided by x_{rp} to obtain the actual fractions. This procedure ensured that Equation (5.1) was satisfied exactly.

The data for the third day (of the three days for which data was available) was chosen for implementing various models. The historical database of O-D flows and counts was created from the data of the first day. The transition matrices in Sections 2.5 and 3.3 were computed by simple Ordinary Least Squares regressions using the deviation of O-D flow values of the second day from those of the first. Two assumptions were made in the process of obtaining these matrices. Firstly, it was assumed that the structure of the autocorrelation remained constant over the whole day so that one could have enough observations in one day to estimate all the parameters. Secondly, it was assumed that a flow between O-D pair r for a period was related only to rth O-D flows of prior intervals. The covariance of the transition equation error was retrieved from these regressions as explained in Section 2.11.2. By virtue of the first assumption, this matrix was time-invariant. It was found that an autoregressive process of order 4 fit the data best. Because of the exactness of equation (5.1), there was no measurement error in the problem. The obvious choices for the initial estimates for the state-vector and its initial covariance matrix were the corresponding historical values.

[4]Note that this procedure ignores vehicles that had entering times before 4.00 A.M. but remain on the network after 4.00 A.M. There is therefore an implicit assumption that all vehicles entering before 4.00 A.M. have left the network before 6.15 A.M.

This case study was primarily used to compare the O-D flow deviation based models with the trip/share deviation based models. For the former, there were 105 unknown flows to be estimated in each 15-minute departure interval, while for the latter, there were 119 state variables. However, for the model with state augmentation (Section 2.6), there were also flows corresponding to s prior intervals to be estimated. In this case study, p' was set equal to 8 since the maximum time taken to traverse the network was about 120 minutes ($= 8 * 15$). Since q' was only 4, the value of s was 8. Hence the number of unknown O-D flows to be estimated during each interval equaled 945.

5.2.2 The I-880 dataset

The first six days of data were used in constructing the historical database and calibrating model parameters. The historical database was constructed using GLS based models as in Cascetta et al.[13][5]. Data from the last day was used for testing the models. The transition matrices and the error variances were computed exactly as in the earlier dataset. Again, an autoregressive process of order four was used. Since the maximum travel time between any O-D pair was about 9 minutes, the value of $s = \max(p', q' - 1) = \max(1,3) = 3$. Thus there were (3+1)*20=80 O-D flows to be estimated in each interval for the model with state augmentation. Measurement error covariance matrices for each interval h were computed from the residuals obtained for h from the GLS procedure[6] for the first six days.

In addition, this dataset was also used to evaluate the performance of models with stochastic assignment matrices. For the offline models (i.e. those in Section 4.5), two apriori O-D matrices were used[7]. The first was obtained from the estimation results

[5]In Cascetta's sequential model, the apriori estimate for the O-D flow $\mathbf{x}_h$ is taken to be the estimate $\mathbf{x}^*_{h-1}$ (See section 2.11.3). In our offline model, we multiply this quantity by the factor $\mathbf{x}^H_h / \mathbf{x}^H_{h-1}$ where the superscript H refers to historical values – in this case, those computed for previous days. This factor helps in accounting for the interval-over-interval variation in O-D flows more effectively. Note that this factor cannot be computed for the first day since there is no prior history.

[6]For the first two days of offline estimation, we used an FGLS procedure as in Section 2.11.3.

[7]Thus there would be two quadratic terms in the error criterion with different weighting factors.

of previous days i.e. the apriori matrix for interval h corresponded to the estimated matrix for interval h for a previous day. The second matrix corresponded to the estimate obtained on the same day in interval $h-1$[8]. The covariance matrices for both error terms were calibrated from residuals for the previous days exactly as in Section 2.11.

For stochastic assignment matrix modeling using the second approach, two sets of travel speeds[9] were used. The first corresponded to the estimation result for the previous period $(h-1)$. The second corresponded to the values measured by the sensors during the current period h. Again, error variances for both error terms were calibrated from residuals for previous days.

Similarly for the first approach, two sets of assignment matrix fractions were used. The first corresponded to the estimation result for the previous period $(h-1)$ while the second, to the values computed using equations (4.5) and the measured speeds during the current period h. As before, error variances were calibrated from residuals for previous days. For this case study, 39 of the 106 assignment fractions were estimated along with the O-D flows. The remaining fractions corresponded to O-D pairs with very low flows and hence their randomness was not considered.

Since $p' = 1$, during any interval p, only $\mathbf{a}_p^{p-1}$ and $\mathbf{a}_p^p$ needed to be estimated. Because no route choice existed, the elements of $\mathbf{a}_p^{p-1}$ had to obey the following relationship.

$$a_{lp}^{rp-1} = 1 - a_{lp-1}^{rp-1} \tag{5.2}$$

The above relationship had to be satisfied for all (l, r) pairs for which $l \in L_k$ for any $k \in K_r$. We enforced this constraint in the following manner. For any interval h, we estimated the assignment fractions a_{lh}^{rh}. We then computed a_{lh+1}^{rh} using (5.2) and moved to interval $h+1$.

In the implementation of the real-time models with stochastic assignment matri-

[8]In other words, we used Equations (2.2) and (2.3) as direct measurements.

[9]Speeds were used as the fundamental entities instead of travel times in all the models i.e. all the equations described in Sections 4.3, 4.4 and 4.5 were written for speeds rather than travel times. The structure of the equations, however, was exactly the same.

ces, i.e. the models in Section 4.4, matrices $\mathbf{M}_h^t$, $\mathbf{M}_h^q$ and $\mathbf{M}_h^a$ were directly computed from historical estimates. The error variances for the measurement equation were the same as for the modified offline methods described earlier. Constraints (5.2) were enforced exactly as before. We also note that the EKF procedure does not guarantee that the estimated assignment fractions lie between zero and unity as required. We therefore truncate each negatively estimated fraction to zero and each fraction greater than unity to one[10].

5.2.3 The Amsterdam Beltway

The main objective of this dataset was to provide a combination of synthetic and actual data to better evaluate different models. Synthetic data was generated by the following series of steps:

1. *Generate "True" O-D flows and speeds for Day 1:* This process is shown schematically in Figure 5-3. The sequential model of Cascetta et al. (Equation (2.61)) was implemented for one day of actual (observed) volume and speed data. This involved estimating 365 O-D pairs for each interval using 65 measurements on link counts and speeds. A constrained FGLS procedure was used. For route-choice, a simple logit model with path travel time as the only attribute for each route was used. The coefficient for this variable was fixed such that the resulting model provided best fit to observed link counts (See Section 5.3.3). The estimated O-D flows and observed speeds were then taken to be the "true" ones for subsequent analysis.

2. *Generate "True" O-D flows and speeds for Day 2:* This process is shown schematically in Figure 5-4. This consists of the following steps.

   - Generate O-D flow deviations and speeds for interval 1: True O-D flows for the first interval of the first day are perturbed to yield O-D flows for

[10] Our empirical results indicated that these were rare occurrences.

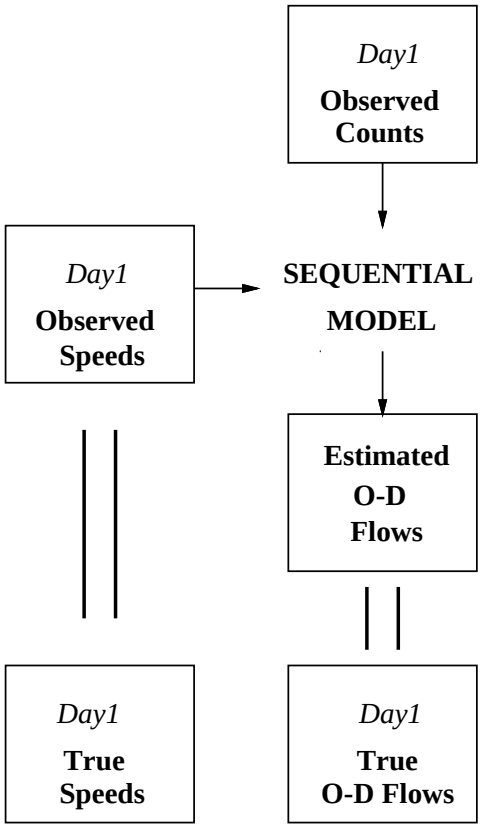

Figure 5-3: Generating True O-D Flows and Speeds for Day1

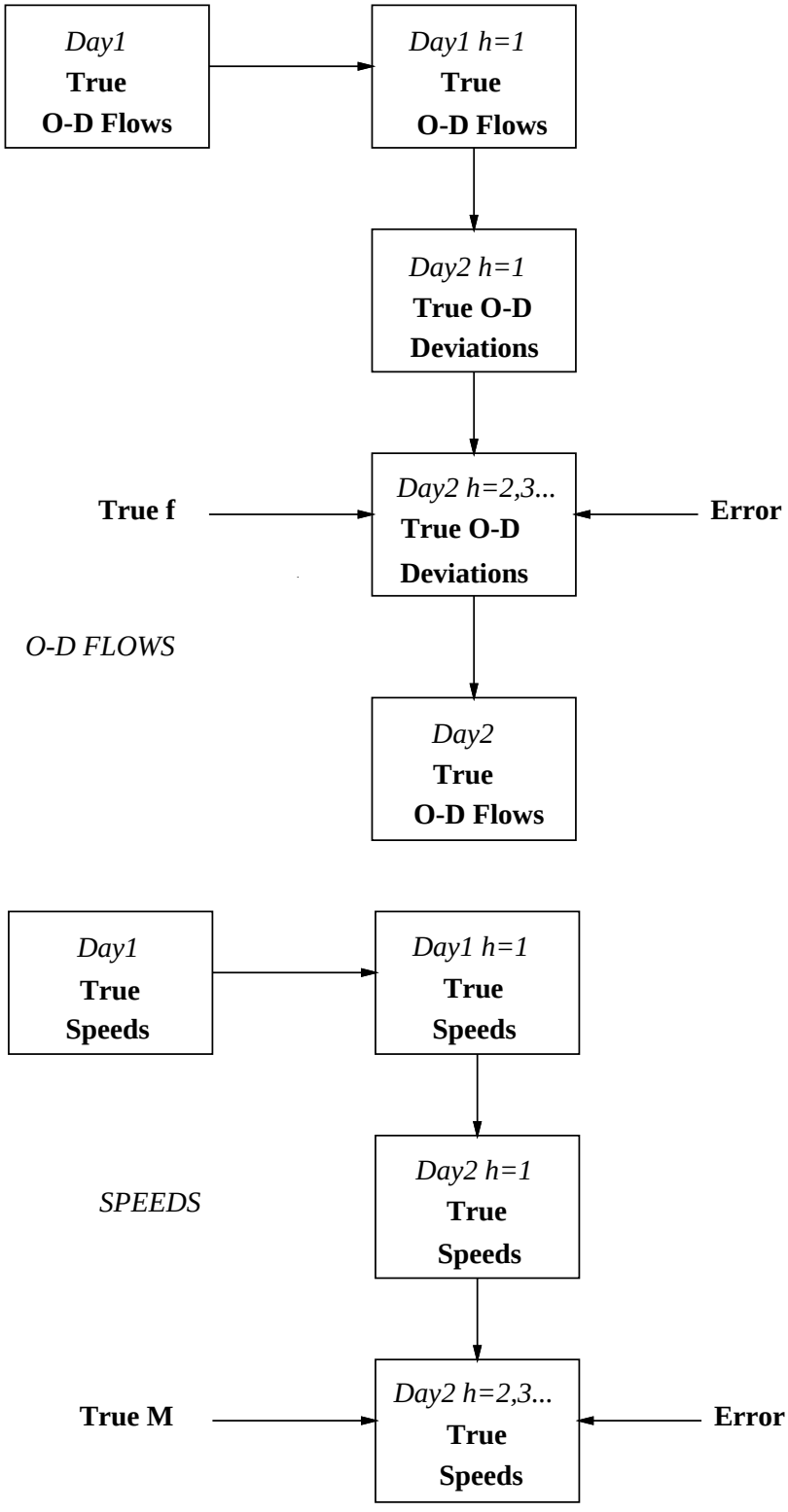

Figure 5-4: Generating True O-D Flows and Speeds for Day2

the first interval of the second day according to the following formula:

$$od\_interval\_1\_day2 = od\_interval\_1\_day1 * (1.0 + \delta_1 * U) \qquad (5.3)$$

where U is a random number uniformly distributed between zero and one and δ_1 is a scalar between zero and one[11]. Before applying the equation, O-D pairs were divided into different groups based on size. Different δ_1 values were used for each group[12]. Once the O-D flows for interval one for Day 2 are computed, deviations for interval one readily follow.

Similarly, the first interval speeds for Day 2 are computed from those for Day 1 by using the following equation:

$$spd\_interval\_1\_day2 = spd\_interval\_1\_day1 * (1.0 - \delta_2 * U) \qquad (5.4)$$

where δ_2 is positive and between zero and one[13].

- Generate O-D flows and speeds for other intervals: Once O-D flow deviations for the first interval were obtained, the transition equation (2.17) was then applied recursively to generate deviations for subsequent intervals. The errors w_{rh} were generated from a uniform distribution between -E and +E[14]. This process required knowledge of the transition matrices $\mathbf{f}_h^p$; accordingly, these were assigned fixed values between 0 and 1[15].

  Speeds for subsequent intervals were computed by a similar procedure. This involved using equation (4.29)[16]. The matrix $\mathbf{M}_h^t$ was computed from speeds of the first day. Again, the errors $\mathbf{\Theta}_h^t$ were generated from a uniform distribution between fixed thresholds -V and +V.

[11] Day 2 can therefore, be construed as a "high volume" day.

[12] The Turnpike data, for which true O-D flows were known for multiple days, was taken as a guideline for choosing δ_1.

[13] reflecting the fact that higher O-D flows for day two would lead to lower speeds

[14] The thresholds E were stratified by size of O-D flow.

[15] Again, the Turnpike data provided useful guidelines on choosing the elements of the transition matrix. The transition matrices were assumed to be diagonal and q' was set equal to 4, as in the other case studies.

[16] for speeds instead of travel times

Once O-D flows and speeds were generated for the second day, the testing procedure was straightforward (Figure 5-5). Essentially, the true speeds and O-D flows were used to compute true link counts. These were then perturbed using the equations

$$Measured\_counts = True\_counts * (1 - \delta_{cts} + 2 * \delta_{cts} * U) \quad (5.5)$$

$$Measured\_speeds = True\_speeds * (1 - \delta_{spd} + 2 * \delta_{spd} * U) \quad (5.6)$$

to generate measured counts and speeds respectively which were finally used by various models in an effort to replicate the original true set of O-D flows[17]. For each scenario, multiple runs were conducted because data was generated stochastically and average error estimates were computed.

We conclude this section with a brief discussion on computation of variance-covariance matrices required as input for various models. Variance of the initial state for O-D flow deviations and speeds may be obtained *exactly* since the initial state for Day 2 is obtained by using equations (5.3) and (5.4) for O-D flows and speeds respectively[18]. For the model with state defined by O-D deviations and assignment fractions, variances were obtained by drawing a large sample of speeds from the uniform distribution, computing assignment fractions corresponding to each, and finally computing the sample variance using these assignment fractions.

Similarly, since measurement errors for counts and speeds were generated by Equations (5.6) and (5.5), and transition errors $\mathbf{w}_h$ and $\mathbf{\Theta}_h^t$ were generated from the uniform distribution mentioned earlier, measurement and transition error variances could be computed *exactly* for most of the models. Again, the only situation where error variances could not be obtained exactly was for the model with stochastic assignment fractions. For this case, as with the initial state variance, transition error variances were approximated by generating a large sample of errors in speeds, $\mathbf{\Theta}_h^t$ (according to the governing uniform distribution), computing the residuals $\mathbf{\Theta}_h^a$ corresponding to

[17] The offline analysis for Day 1 showed that 69 of the 365 O-D pairs had zero flows almost all day. For computational convenience, these were fixed at zero and only 291 O-D pairs were considered for subsequent estimation and prediction.

[18] The variance of a random variable uniformly distributed between a and b is $(b-a)^2/12$.

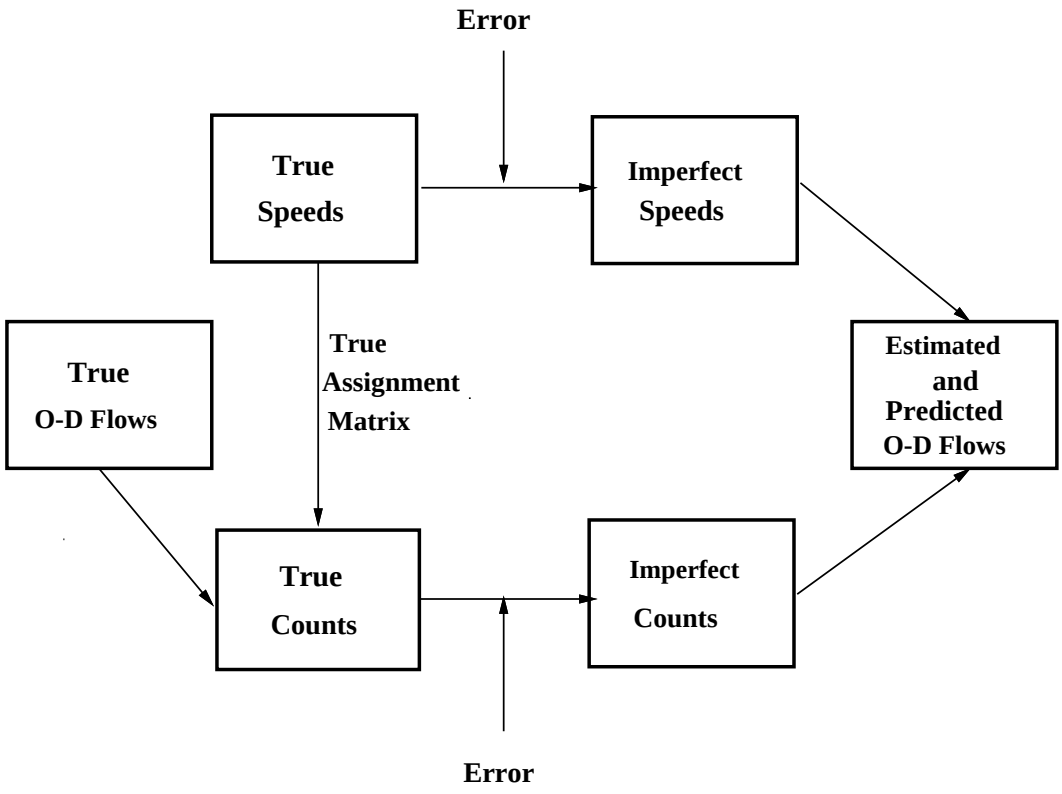

Figure 5-5: Testing procedure with synthetic data

these, and finally, computing the sample variance from the residuals. Exactly the same procedure was used for computing measurement error covariance, except that now, a large sample of errors $\mathbf{\Lambda}_h$ was generated.

5.3 Results

In what follows, we use the following shorthand for describing various models.

- Model with state given by O-D flow deviations of current and prior departure intervals (Section 2.7): *Base*
- Approximate model with state given by O-D flow deviations only of current departure interval (Section 2.10): *Base-Appx*
- Same as Model *Base-Appx* but O-D flow deviations smoothed (Section 2.9): *Sm-Base-Appx*
- Trip/share based model with the approximation in Section 2.10: *T/s-Appx*
- Basic Offline model (as in Cascetta et al.[13]): *Off-Base*
- Modified (rolling horizon) Offline Model (Section 2.11.3, Equation (2.64)): *Off-Mod-Base*
- Offline model with stochastic speeds (Section 4.5): *Off-Stoc-Spd*
- Offline model with stochastic assignment fractions (Section 4.5): *Off-Stoc-Assg*
- *Base-Appx* with stochastic speeds (Section 4.4): *Stoc-Spd*
- *Base-Appx* with stochastic assignment fractions (Section 4.4): *Stoc-Assg*

5.3.1 The Turnpike Data

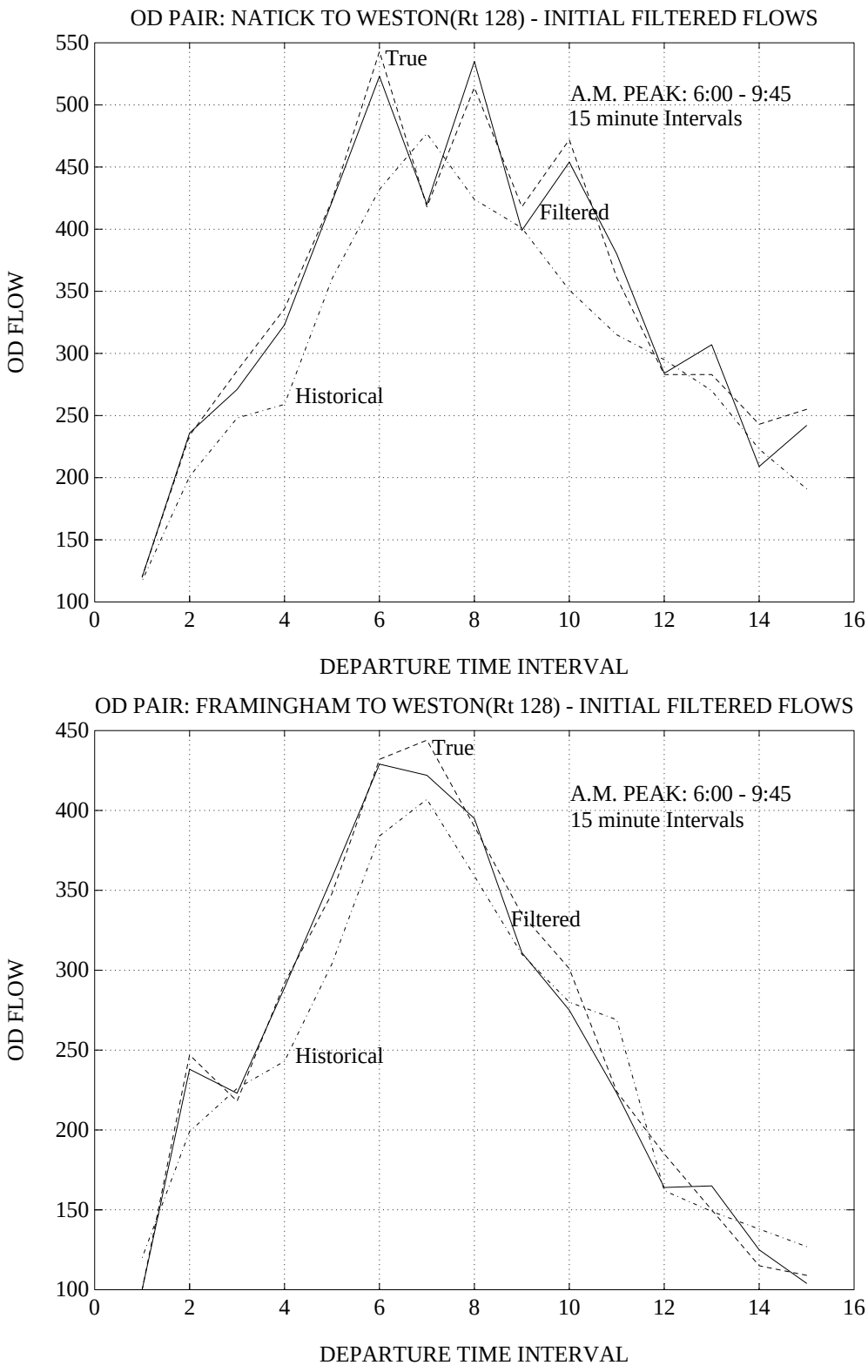

Figure 5-6: Typical Filter Estimates For *Base*

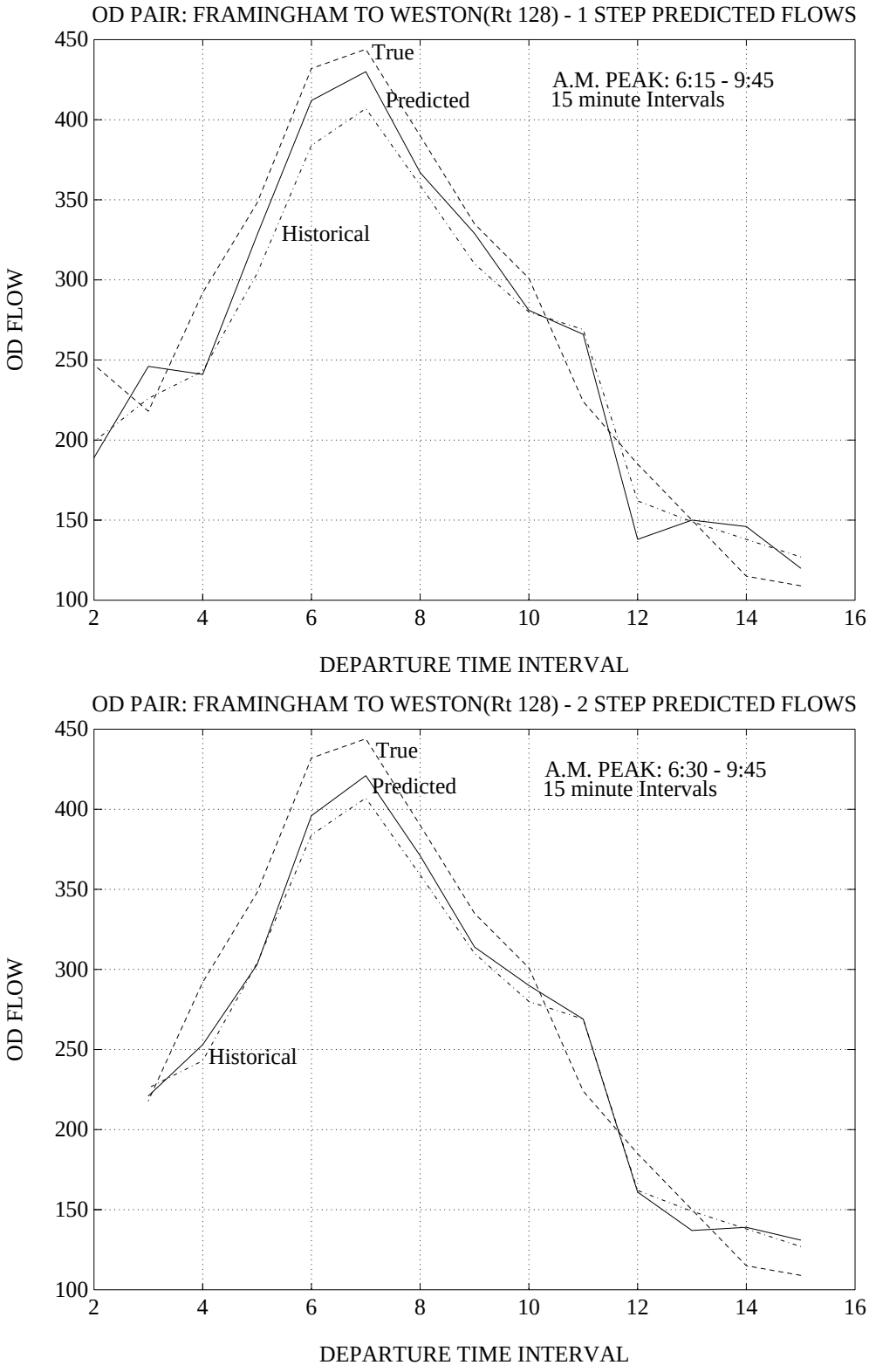

Figure 5-7: One Step and Two Step Predictions for *Base*

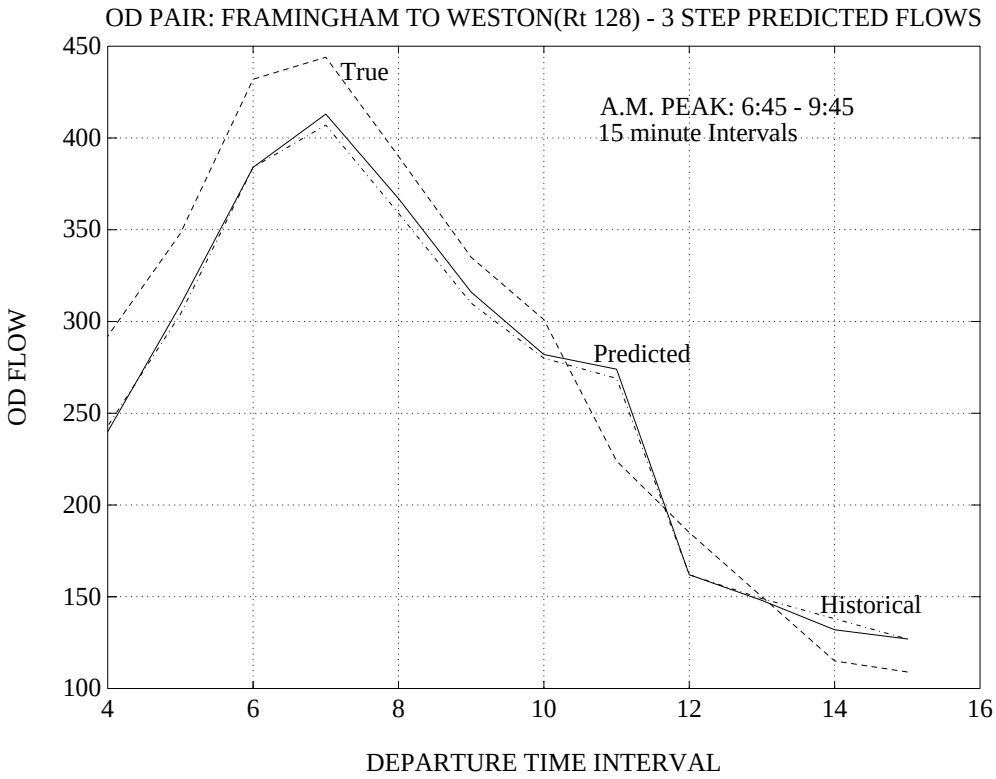

Figure 5-8: Three Step Prediction for *Base*

Figures 5-6 through 5-8 show results for Model *Base* presented for some O-D pairs[19]. Note that Figure 5-6 shows the *initial* filtered flow – because of transformation of variables, each flow is filtered 9 times. It is apparent from figure 5-6 that the filtered estimates are significantly closer to the true values than the corresponding historical values for both O-D pairs Natick-Weston and Framingham-Weston. Figures 5-7 and 5-8 show one-step (15 minute), two-step (30 minute) and three-step (45 minute) predictions. It can be seen that the quality of the predictions deteriorates progressively as the prediction time-step is increased and that the predicted estimates tend to converge to the historical values. This is to be expected given the autoregressive formulation because for every one-step-ahead prediction, we are effectively multiplying the deviation (in the prior interval) by a fraction. The deviations are small given the limited variability in traffic flow over the three days. Multiplying them repeatedly by a fraction reduces them still further. Adding them to the historical values – which are much larger in magnitude in comparison – hence yields estimates that do not differ by much from the historical values.

In the above implementation, there is little variability in the traffic demand over the three days of analysis and hence, the historical flows themselves provide a good approximation to the true values. To test the performance of the filter in replicating the true values in the presence of poor historical information, another implementation was carried out using very poor historical values[20]. The choice of initial conditions is also therefore very poor. Figures 5-9 and 5-10 show results for O-D pair Framingham-Weston. It is seen that the filtering procedure is fairly robust and the quality of the filtered estimates does not seem to be very sensitive to historical information. However, the predicted estimates – though significantly better than the historical values – leave room for improvement.

The results presented thus far were for specific O-D pairs. The following two statistics were computed to get a better overall comparison of the historical estimates and those estimated by the model.

[19]The results in this section for model *Base* come from earlier work by Ashok and Ben-Akiva[1].
[20]These actually corresponded to historical values for the evening peak.

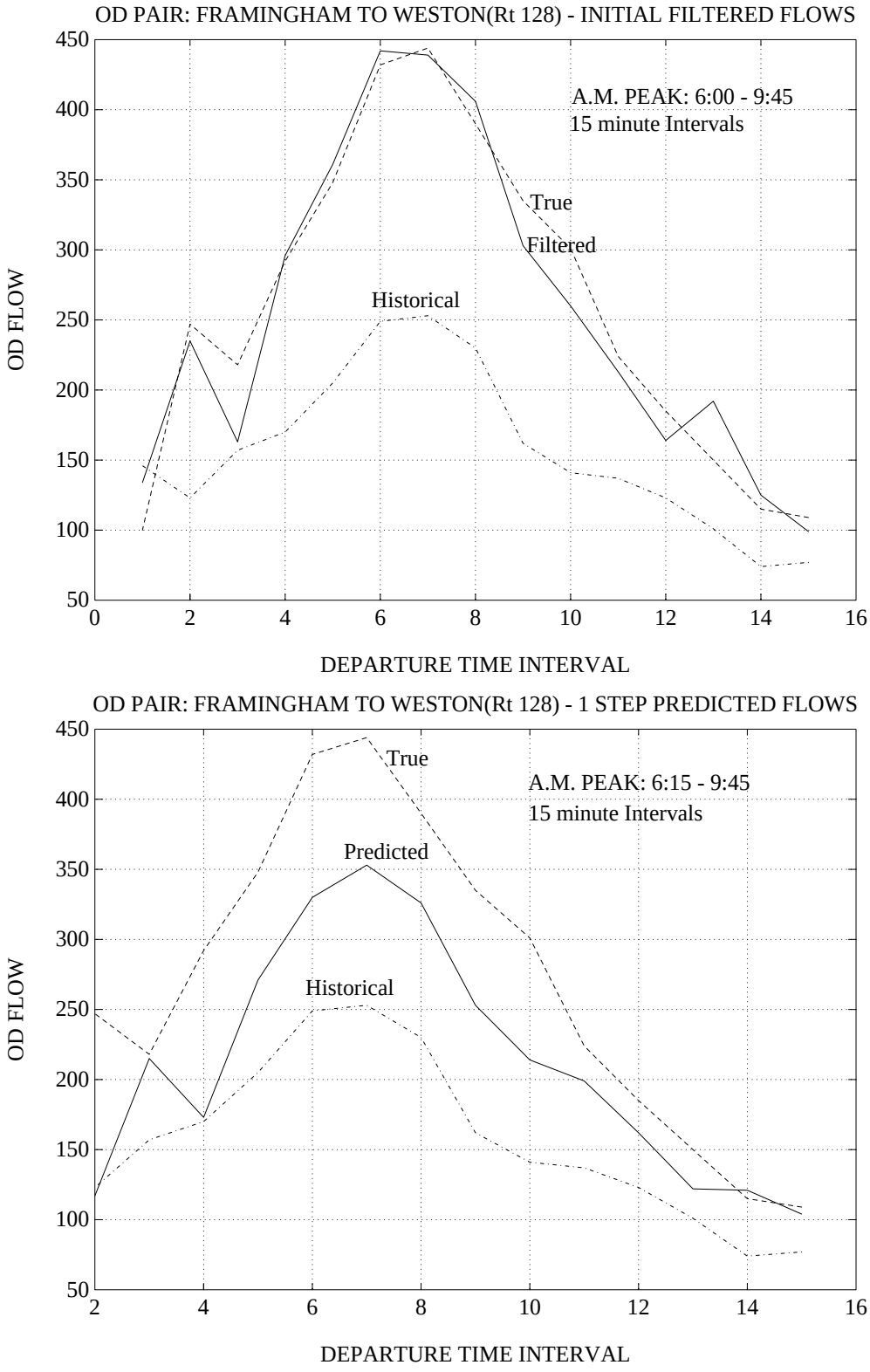

Figure 5-9: Estimates and One-step Predictions with poor history: Model *Base*

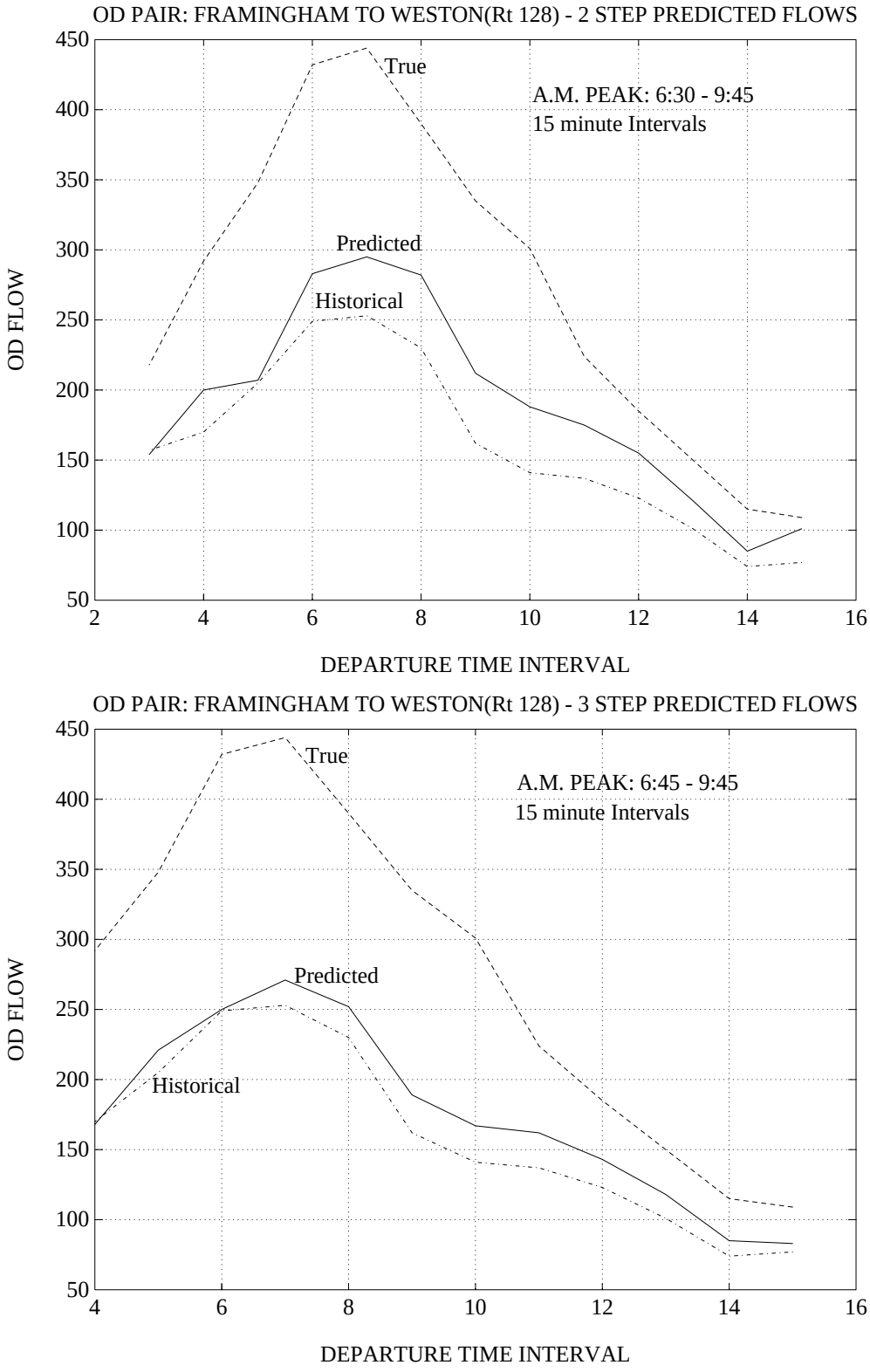

Figure 5-10: Two-step and Three-step Predictions with poor history: Model *Base*

		Base	*Base A*	*Base B*	*Base-Appx*	*Historical*
RMS Error	Filtered	5.3449	*	*	5.7798	8.7015
	1-Step Predicted	9.1061	13.1631	10.9520	9.1916	8.9501
	2-Step Predicted	8.7558	18.2373	9.4515	8.7290	9.1146
	3-Step Predicted	9.3090	25.3705	11.6322	9.2937	9.2490
RMSN Error	Filtered	0.2905	*	*	0.3141	0.4729
	1-Step Predicted	0.4783	0.6913	0.5752	0.4878	0.4701
	2-Step Predicted	0.4502	0.9378	0.4860	0.4525	0.4687
	3-Step Predicted	0.4721	1.2866	0.5899	0.4701	0.4690

Table 5.1: RMS and Normalized RMS Error Values (I-90)

1. Root Mean Square (RMS) Error = $\sqrt{\frac{\sum_i (x_i - \hat{x}_i)^2}{N}}$
2. Root Mean Square Error Normalized (RMSN) = $\frac{\sqrt{N \sum_i (x_i - \hat{x}_i)^2}}{\sum_i x_i}$

   where the summation is over all O-D pairs and all intervals for which analysis was carried out.

Shown in Table 5.1 are results for *Base* and *Base-Appx*. The second and third columns (*Base A* and *Base B*) correspond to *Base* models that use alternate schemes for predictions. Model *Base A* corresponds to a "no prediction" case – the O-D flows estimated by *Base* during a given interval are the "predictions" for all future intervals. *Base B* corresponds to a prediction method that uses constant deviations i.e., the deviations in O-D flows estimated by *Base* during a given interval are assumed to be identical for all future intervals. Next in the table are errors corresponding to Model *Base-Appx*. The final column displays the errors when the historical O-D flows for each interval were used in lieu of estimates/predictions from the two models.

We make the following observations. First, the RMS errors are relatively low and the filtered estimates from both *Base* and *Base-Appx* are much more accurate than the historical values. We next see that there is some loss in accuracy in moving from *Base* to *Base-Appx* especially in the filtered estimates; however these have to be traded off against the vast computational gains – instead of estimating 945 parameters in *Base*, only 105 parameters need to be estimated in *Base-Appx*. While predictions

from *Base* are significantly better than those from Models *Base A* and *Base B*, they are not much different from the historical values. As mentioned earlier, this could be because there was not much variability in O-D flows during the three day period and the historical flows provided good approximations to the O-D flow on the day of interest.

Following the observation that most of the O-D flows in the case study were extremely small, we next computed errors for *Base* and *Base-Appx* only for the high flows. Table 5.2 shows that RMSN errors are drastically reduced showing that both models perform significantly better while estimating and predicting high O-D flows.

Of course one reason for the small gap in performance between *Base* and *Base-Appx* could be the fact that there was no measurement error in the problem. This could explain why most of the information about an O-D flow could be obtained from just one measurement. To perform a more "fair" comparison[21] between the two models, we first perturbed the assignment matrix using the following formula:

$$a_{new} = a_{correct}[(1-\delta) + U * 2\delta]$$

for different values of δ. U is a random number drawn from a uniform distribution between zero and one. The performance of filtered estimates for a value of $\delta = 0.2$[22] is shown in Table 5.3. Expectedly, the errors are higher for both models with the gap between *Base* and *Base-Appx* widening.

Table 5.4 (only for high O-D flows) confirms our conclusion from Figures 5-9 and 5-10 that both models continue to be significantly superior to historical values when an extremely poor historical database is used.

Overall, results seem to indicate that the augmentation of the state-vector with variables corresponding to prior departure intervals does not offer significant improvement to warrant the extra computations that are required.

Tables 5.5 and 5.6 show RMS and RMSN errors for Models *T/s-Appx* based on

[21] In general, interaction across multiple intervals came from *two* sources – the presence of O-D pairs on the network over many time intervals *and* the multi-order autoregressive formulation. Hence, the comparisons in Tables 5.1 and 5.2 have some value even in the absence of measurement error.

[22] This can be loosely interpreted as a $\pm$20% error in the assignment matrix.

			Base	*Base-Appx*	*Historical*
RMS Error	Flows ≥ 100	Filtered	15.0397	15.8635	38.6223
		1-Step Predicted	39.2869	39.8425	39.5175
		2-Step Predicted	36.7074	36.6537	39.8209
		3-Step Predicted	40.1668	40.1617	40.6043
	Flows ≥ 150	Filtered	15.1953	16.3323	44.8590
		1-Step Predicted	44.8964	45.5796	45.3617
		2-Step Predicted	42.0653	42.0238	46.2092
		3-Step Predicted	47.1325	47.1934	47.7631
RMSN Error	Flows ≥ 100	Filtered	0.0584	0.0616	0.1499
		1-Step Predicted	0.1484	0.1505	0.1493
		2-Step Predicted	0.1380	0.1378	0.1497
		3-Step Predicted	0.1503	0.1503	0.1519
	Flows ≥ 150	Filtered	0.0472	0.0507	0.1393
		1-Step Predicted	0.1378	0.1399	0.1392
		2-Step Predicted	0.1263	0.1262	0.1388
		3-Step Predicted	0.1382	0.1384	0.1400

Table 5.2: RMS and RMSN Error Values for high O-D flow pairs (I-90)

		Base	*Base-Appx*	*Historical*
RMS Error	Filtered	7.4698	8.4265	8.7015
	1-Step Predicted	8.7998	8.4310	8.9501
	2-Step Predicted	9.0479	8.5518	9.1146
	3-Step Predicted	9.3242	9.7057	9.2490
RMSN Error	Filtered	0.4060	0.4579	0.4729
	1-Step Predicted	0.4670	0.4474	0.4701
	2-Step Predicted	0.4691	0.4434	0.4687
	3-Step Predicted	0.4716	0.4909	0.4690

Table 5.3: RMS and RMSN errors with erroneous assignment matrix (I-90)

			Base	*Base-Appx*	*Historical*
RMS Error	Flows ≥ 100	Filtered	47.7839	57.3556	141.6710
		1-Step Predicted	111.2135	109.1876	144.6273
		2-Step Predicted	132.6936	131.5689	146.6274
		3-Step Predicted	144.6440	144.3848	149.9019
	Flows ≥ 150	Filtered	46.8046	58.4643	162.3158
		1-Step Predicted	121.3750	119.6109	164.3451
		2-Step Predicted	149.8922	149.2868	168.4647
		3-Step Predicted	167.0358	167.4073	174.5907
RMSN Error	Flows ≥ 100	Filtered	0.1854	0.2226	0.5498
		1-Step Predicted	0.4201	0.4124	0.5463
		2-Step Predicted	0.4988	0.4946	0.5512
		3-Step Predicted	0.5412	0.5403	0.5609
	Flows ≥ 150	Filtered	0.1454	0.1816	0.5042
		1-Step Predicted	0.3725	0.3671	0.5043
		2-Step Predicted	0.4502	0.4483	0.5059
		3-Step Predicted	0.4897	0.4908	0.5119

Table 5.4: RMS and RMSN Error Values with poor historical information (I-90)

the departure-rate/share based formulation. The table shows two different versions of these. The trip-share based models require that the estimated shares lie between zero and unity and that the shares corresponding to each origin add to unity. Since the EKF procedure does not guarantee this, we truncate each negatively estimated share[23] to zero and normalize all shares during each interval. The "A" model incorporates this truncation/normalization feature while the "B" model ignores these constraints.

We observe that both *T/s-Appx A* and *T/s-Appx B* out-perform the linear models in predictive power. At an intuitive level, this could be explained by the fact that the transition equation now allows for differential variability of departing trips and shares. On the other hand, they exhibit worse performance in filtering which could be because of the non-linearity in the measurement equation and the resulting approximation. Also, the "B" model is marginally inferior to the "A" model in its filtered estimates.

The number of iterations in implementation of the Iterated EKF algorithm de-

[23]Our empirical results indicated that this was an infrequent occurrence and and happened only for O-D pairs with extremely low flow.

		T/s-Appx A	*T/s-Appx B*	*Historical*
RMS Error	Filtered	6.8346	7.0359	8.7015
	1-Step Predicted	8.9679	8.9654	8.9501
	2-Step Predicted	8.3042	8.3049	9.1146
	3-Step Predicted	9.2028	9.1839	9.2490
RMSN Error	Filtered	0.3714	0.3824	0.4729
	1-Step Predicted	0.4710	0.4709	0.4701
	2-Step Predicted	0.4270	0.4271	0.4687
	3-Step Predicted	0.4667	0.4657	0.4690

Table 5.5: RMS and RMSN Errors for alternate formulation (I-90)

			T/s-Appx A	*T/s-Appx B*	*Historical*
RMS Error	Flows ≥ 100	Filtered	23.2466	23.8482	38.6223
		1-Step Predicted	39.1845	38.9744	39.5175
		2-Step Predicted	34.7828	34.7210	39.8209
		3-Step Predicted	40.3939	40.2259	40.6043
	Flows ≥ 150	Filtered	25.3686	26.3215	44.8590
		1-Step Predicted	44.9034	44.6657	45.3617
		2-Step Predicted	39.9146	39.8508	46.2092
		3-Step Predicted	47.2156	46.9441	47.7631
RMSN Error	Flows ≥ 100	Filtered	0.0902	0.0925	0.1499
		1-Step Predicted	0.1480	0.1472	0.1493
		2-Step Predicted	0.1308	0.1305	0.1497
		3-Step Predicted	0.1511	0.1505	0.1519
	Flows ≥ 150	Filtered	0.0788	0.0818	0.1393
		1-Step Predicted	0.1378	0.1371	0.1392
		2-Step Predicted	0.1199	0.1197	0.1388
		3-Step Predicted	0.1384	0.1376	0.1400

Table 5.6: RMS/RMSN Error Values for high flows using alternate formulation (I-90)

Iterations	*T/s-Appx A*	*T/s-Appx B*
1	7.5903	10.9918
2	6.8239	7.0034
3	6.8302	7.0341
6	6.8346	7.0353
7	6.8346	7.0359

Table 5.7: RMS Errors in filtered estimates vs number of iterations (I-90)

pends upon both the sensitivity of the results to number of iterations and the computational effort associated with each additional iteration. Table 5.7 shows the filtering errors associated with different number of iterations using the "A" and "B" models. It is seen that the error values quickly stabilize.

5.3.2 The I-880 Data

Results from the I-880 dataset showed the same overall trends for the various models as can be seen in Table 5.8. Note that the errors in Table 5.8 are with respect to *link counts* and not the true O-D flows. This error measure should be interpreted with caution since (a) the measured counts are themselves erroneous and (b) it is possible that even though the estimated link counts closely match the measured link counts, the estimated O-D flows differ considerably from the true O-D flows. Thus, while these results are useful for identifying general trends, to reach definitive conclusions based on these would be premature.

Again, *Base-Appx* displays higher errors relative to *Base*. Another interesting observation relates to the variances of the estimated O-D flows in *Base*. Since each O-D flow is estimated multiple (in this case four) times, one would expect the variance of the estimates to decrease with each successive estimate. This was borne out in the results. Figure 5-11 shows the relationship between the variance of the filtered estimates and the number of estimates for several O-D pairs for a specific departure interval. It is seen that most of the reduction in variance takes place within two estimations. This could be because in this case-study, vehicles can remain on the

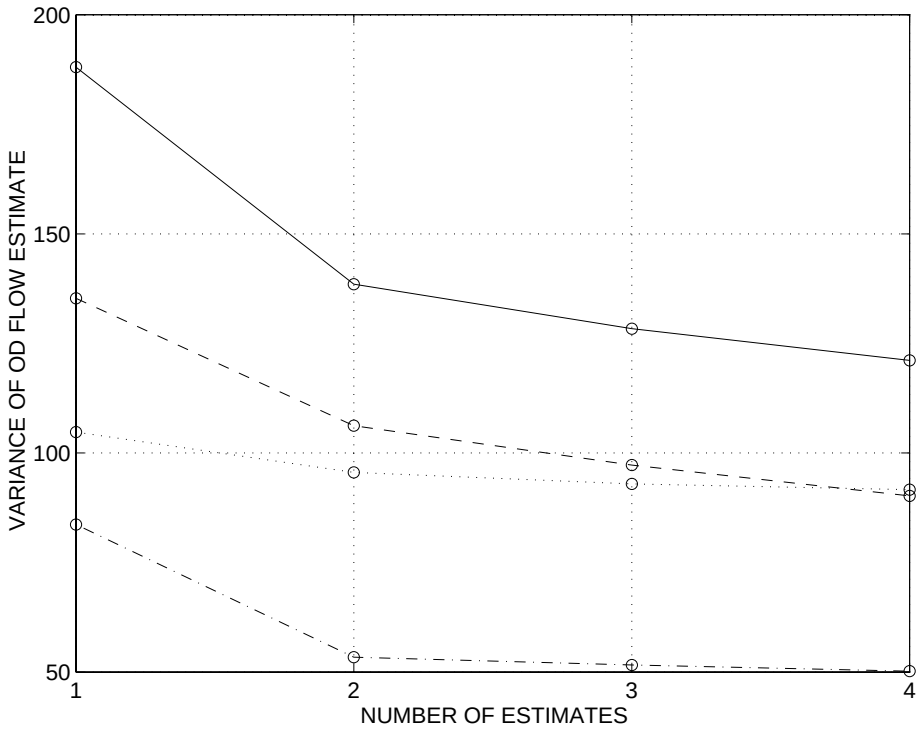

Figure 5-11: Variance of Filtered OD flows vs number of estimates (I-880)

network for at most two successive time-intervals[24]. Finally, *T/s-Appx A* and *T/s-Appx B* out-perform their linear counterpart (*Base-Appx*) in predictive power just as in the I-90 dataset[25].

We turn our attention next to the models with a stochastic assignment matrix. Table 5.9 shows the relative performance of the four offline models. The last column shows the effect of assigning historical O-D flows to the network (where the assignment matrix is computed using measured speeds). We observe that though all the models significantly outperform the historical values in terms of fit to link counts, there is almost no difference between the performance of the three models. An examination of the estimated O-D flows for the three indicated very little difference as well. One

[24]That there is still a reduction in variance beyond two estimations is because the autoregressive process is of order 4.

[25]Predicted values for link counts depend on predicted O-D flows and predicted assignment fractions. Those in Table 5.8 have been obtained by using predicted (or historical, in the case of the last column) O-D flows and assignment matrices *held constant*, based on the speeds observed for the *current* interval observed speeds.

		Base	*Base-Appx*	*T/s-Appx A*	*T/s-Appx B*	*Historical*
RMS Error	Filtered	20.8819	27.6405	29.8567	23.2492	85.1615
	1-Step	57.9496	77.6307	67.3107	70.6092	82.2823
	2-Step	52.9301	70.0270	55.4017	58.3487	76.7824
	3-Step	47.7663	92.8057	47.3866	54.3810	63.8978
RMSN Error	Filtered	0.0197	0.0261	0.0282	0.0219	0.0808
	1-Step	0.0546	0.0732	0.0635	0.0666	0.0776
	2-Step	0.0500	0.0662	0.0524	0.0552	0.0726
	3-Step	0.0457	0.0887	0.0453	0.0520	0.0611

Table 5.8: RMS and RMSN Errors in Link Volumes (I-880)

Errors	*Off-Base*	*Off-Mod-Base*	*Off-Stoc-Spd*	*Off-Stoc-Assg*	*Historical*
RMS	26.9748	26.5299	26.8931	26.7943	85.6206
RMSN	0.0254	0.0247	0.0254	0.0253	0.0812

Table 5.9: RMS and RMSN Errors in Link Volumes Using Offline Models

explanation for these results could be that for a linear network (with no route-choice) where speeds do not change drastically interval-over-interval, Model *Off-Base* is fairly insensitive to errors in the assignment matrix[26].

[26]Indeed, for Model *Off-Stoc-Spd*, an RMSN measure that compared estimated speeds with the measured ones indicated a difference of 13.7% between the two. It is interesting that this large difference in speeds does not seem to have translated into comparable differences in O-D and link flow estimates. Again, this might have to do with the linear structure of the network.

		Base-Appx	*Stoc-Spd*	*Stoc-Assg*	*Historical*
RMS Error	Filtered	27.6405	21.7885	21.3744	85.1615
	1-Step Predicted	79.4600	111.1101	72.4859	88.1974
	2-Step Predicted	77.4488	116.1716	67.2875	86.2065
	3-Step Predicted	102.3850	102.1264	74.9402	68.7522
RMSN Error	Filtered	0.0261	0.0205	0.0202	0.0808
	1-Step Predicted	0.0749	0.1062	0.0683	0.0832
	2-Step Predicted	0.0732	0.1126	0.0636	0.0815
	3-Step Predicted	0.0979	0.1012	0.0717	0.0657

Table 5.10: RMS and RMSN Errors in Link Volumes

And finally, we conclude this section with Table 5.10 that displays results from the state-space models based on stochastic assignment matrix[27]. Again, results using the historical O-D flows have been shown for comparison.

We make several observations. Firstly, as in the offline case, all the three models show significantly better performance in estimation, compared to historical values. Except for *Stoc-Spd*, one-step and two-step predictions are better as well. The most interesting find however is that unlike in the offline case, Models *Stoc-Spd* and *Stoc-Assg* significantly outperform *Base-Appx*. One reason could be that insofar as the state-space model works with deviations in O-D flows and uses an autoregressive process, it represents a different statistical model with different properties compared to the GLS based offline model[28]. Also, unlike the offline methods, the transition equations in the real-time case provide an explicit modeling of the relationship between speeds and assignment fractions across time-intervals[29]. We finally notice that *Stoc-Spd* performs worse than the others in prediction indicating perhaps a need for examining other alternatives to equation (4.29).

5.3.3 The Amsterdam Data

We start with results for the first day. Figure 5-12 shows the fit to counts when the offline sequential model *Off-Base* is applied to Day 1. Results are shown for different values of the travel time coefficient in the logit model. Figure 5-12 indicates that best results are provided for an all or nothing assignment to the shortest path. We therefore used a very low value of $\beta = -6$ for all subsequent analysis. This has the effect of magnifying even extremely small travel time differences between alternate paths. Once O-D flows were estimated for the first day using *Off-Base*, data was

[27] In this case, a k step ahead prediction for *Base-Appx* uses the assignment fractions based on observed speeds k steps ahead. Predictions for *Stoc-Spd* and *Stoc-Assg* use the transition equations (4.29) and (4.31). Historical predictions use the same assignment fractions as *Base-Appx*. This explains the difference in entries between Tables 5.10 and 5.8 for *Base-Appx* and *Historical*.

[28] For example, the error terms in equation (2.50) are more amenable to a normal approximation than those in Cascetta et al.

[29] These equations also used historical speeds and assignment fractions – something the offline models did not.

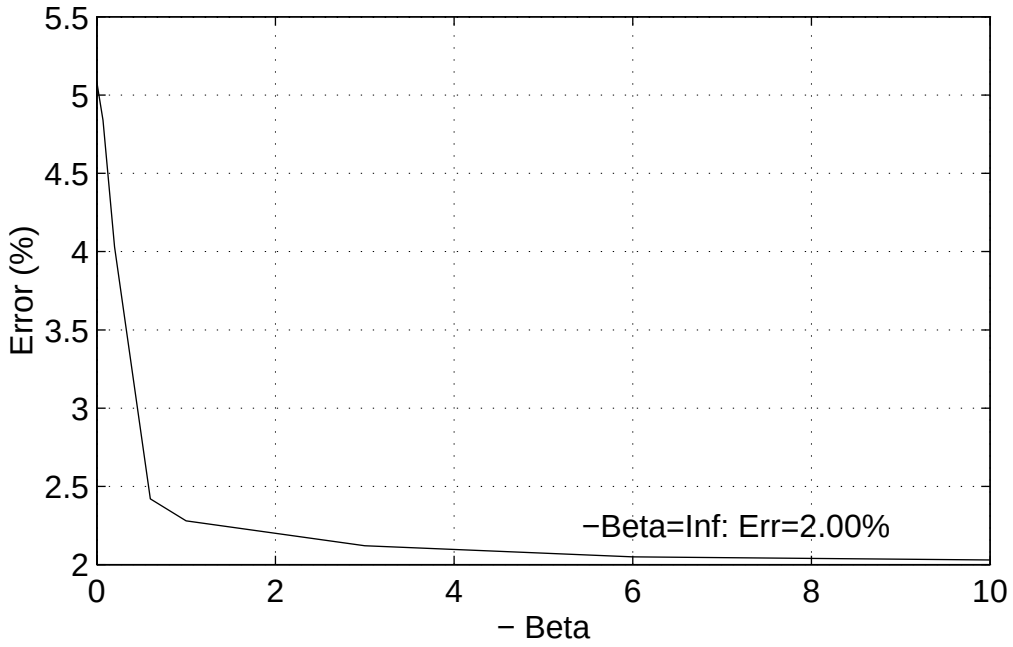

Figure 5-12: Error in Fit to counts for Model *Off-Base* for Day 1

generated for the second day using the procedure explained in the previous section. The remainder of this section deals with application of various models on synthetic data for the second day.

Figure 5-13 shows the degradation in model performance[30] as the measurement error parameter δ_{cts} (equation (5.5)) is varied[31]. Errors are stratified by size of O-D flow. Each bar shows the RMS/RMSN errors for a specific value of δ_{cts}. For comparison, errors in employing historical O-D flows (in other words, the difference between O-D flows on the first and second days) are shown in the last bar. It can be seen that the model is fairly robust with respect to quality of link counts. A similar conclusion is reached in Figure 5-14 which investigates the extent of bias as the error parameter in speeds (equation (5.5)) is varied. The bias becomes significant, however, for very large values of δ_{spd}.

[30]The notation (xx, yy) in figures in this section indicate that δ_{cts}, the error parameter for counts is xx and δ_{spd}, the error parameter for speeds is yy.

[31]Recall that $\delta_{cts} = e$ implies a measurement error within $\pm e\%$ in link counts.

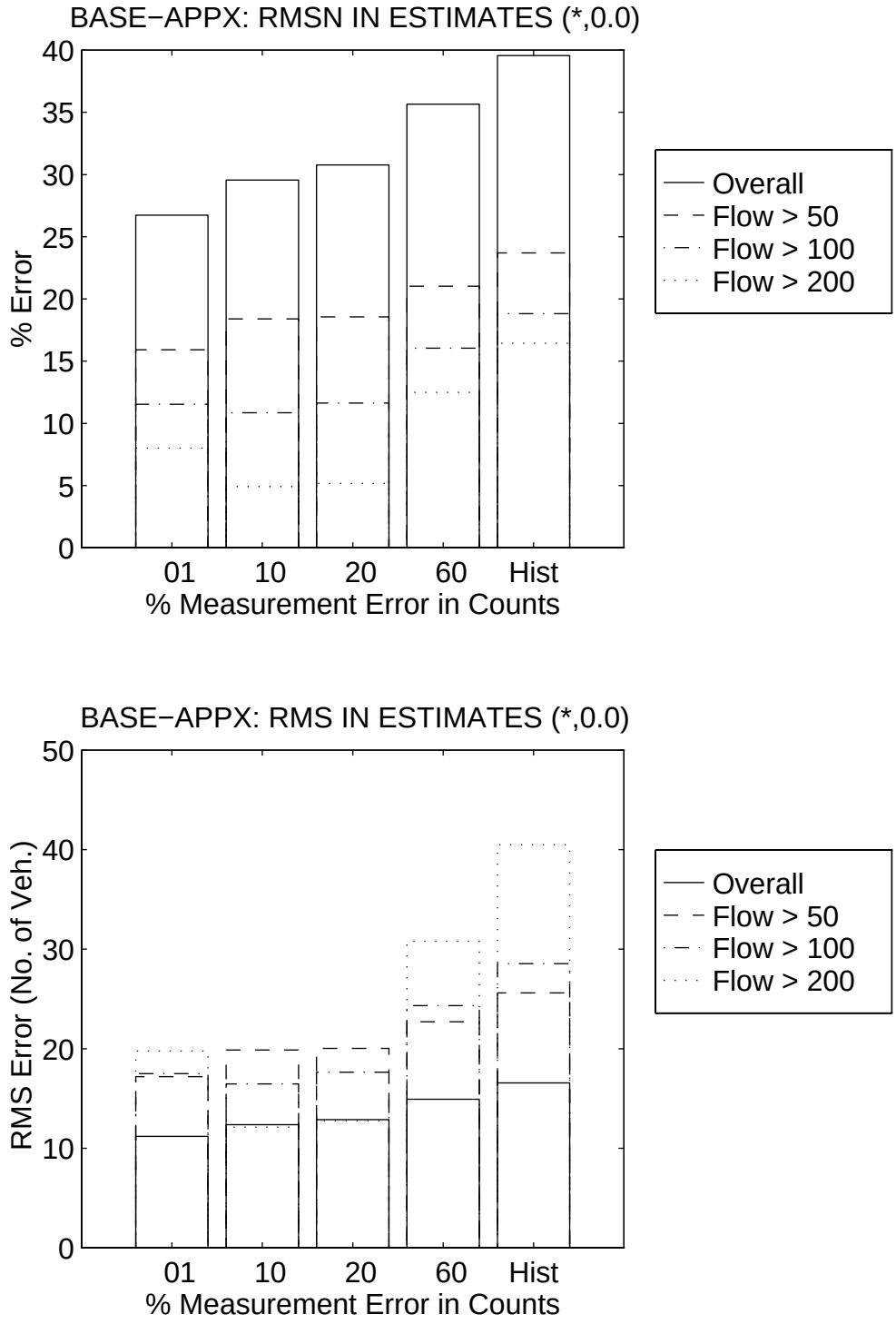

Figure 5-13: Model performance as a function of accuracy of counts : *Base-Appx*

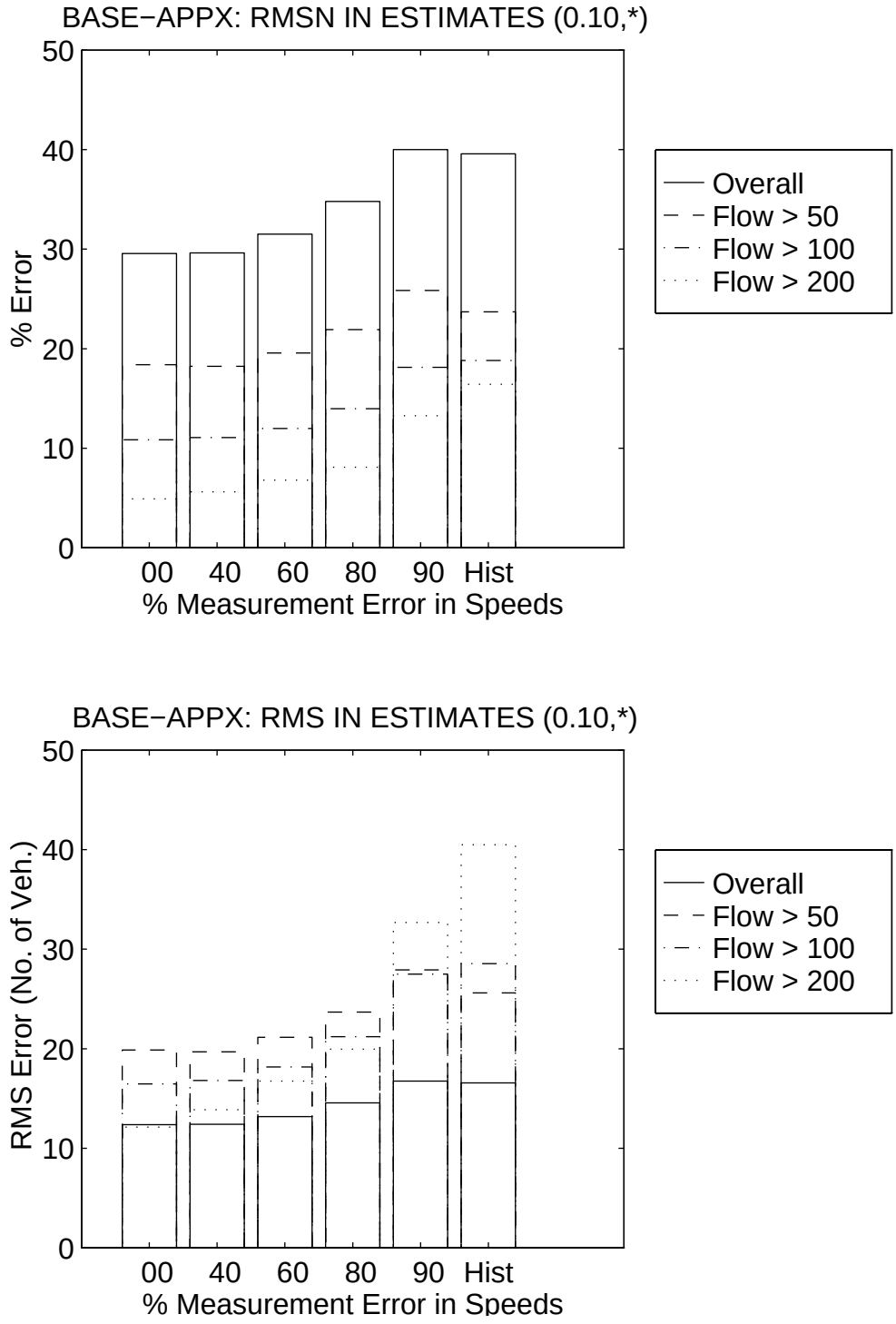

Figure 5-14: Model performance as a function of accuracy of speeds : *Base-Appx*

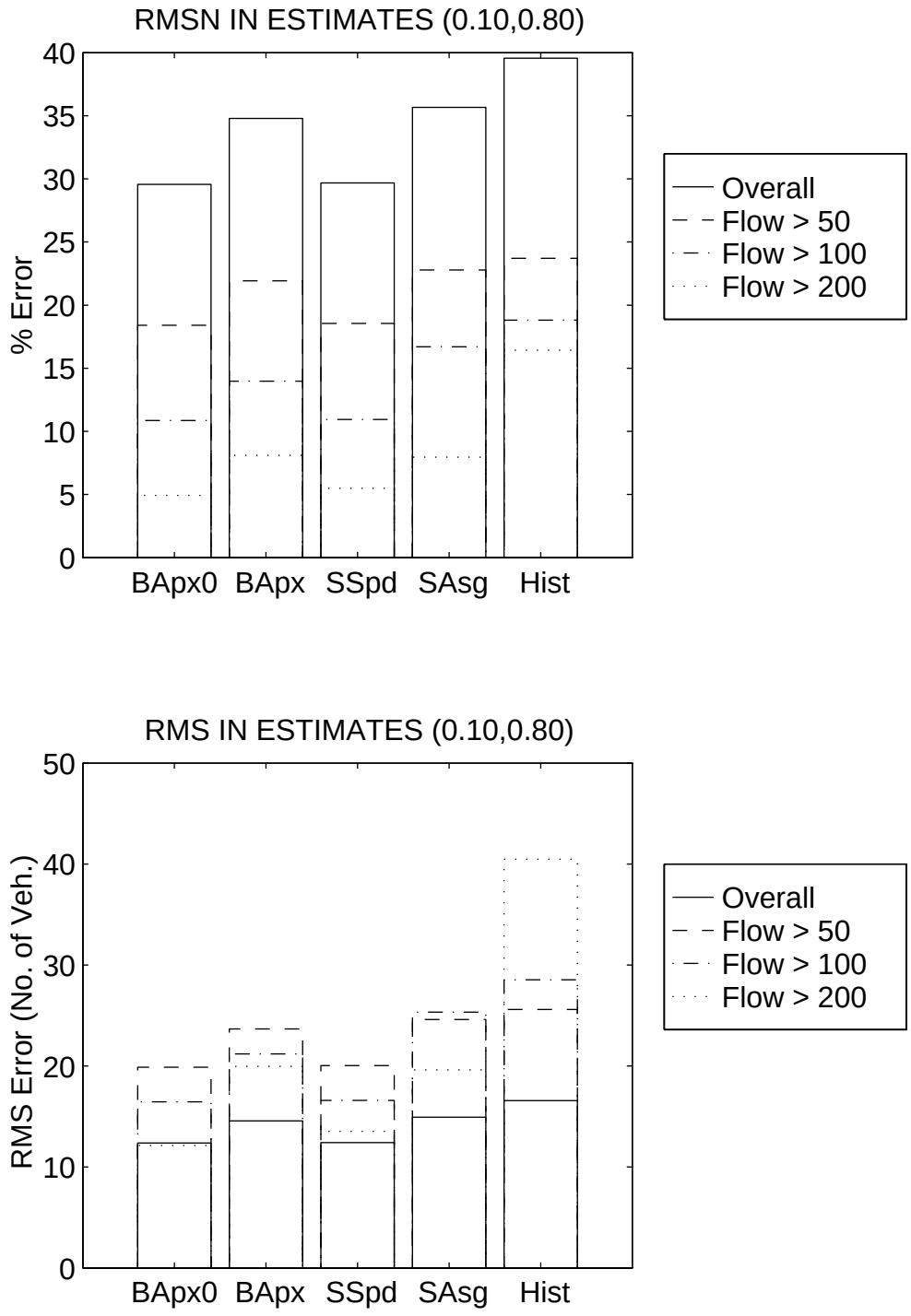

Figure 5-15: Fixed and Stochastic Assignment Matrix Models

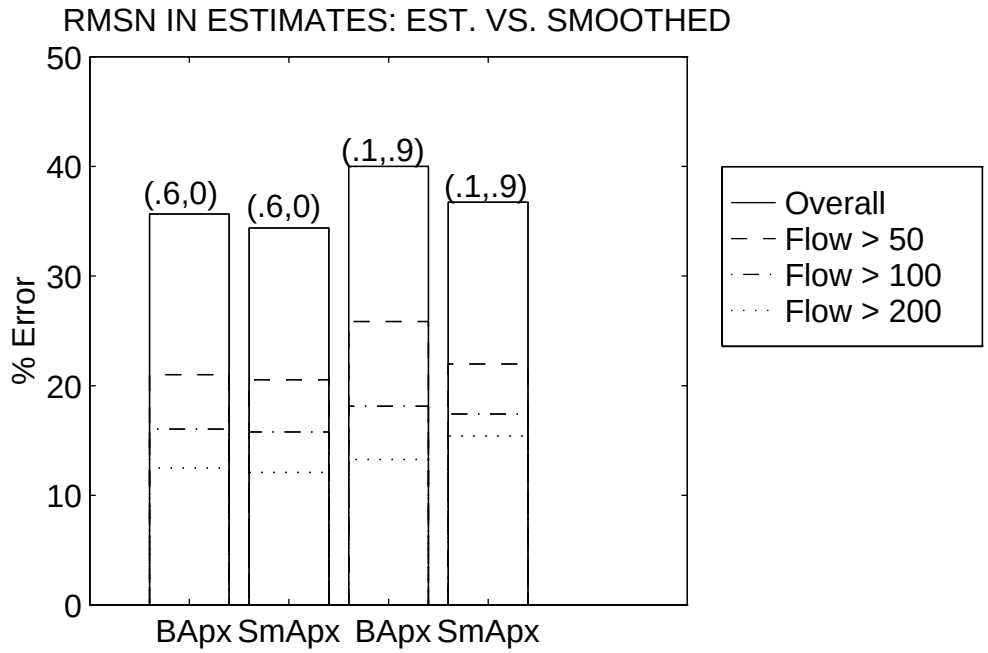

Figure 5-16: Estimation vs Smoothing

We move next to Figure 5-15 that compares the fixed and stochastic assignment matrix models for $\delta_{cts} = 0.1$ and $\delta_{spd} = 0.8$. The first bar pertains to the model *Base-Appx* with true speeds, i.e., $\delta_{spd} = 0$. The second bar pertains to the application of *Base-Appx* with $\delta_{spd} = 0.8$. The third and fourth bars pertain to the *Stoc-Spd* and *Stoc-Asg* models respectively. Again, errors in historical estimates are shown for comparison. We see that the model *Stoc-Spd* performs extremely well – in fact, its performance almost rivals that of *Base-Appx* with true speeds. The same cannot be said, however, for the *Stoc-Asg* model which does no better than *Base-Appx*. This poor performance of *Stoc-Asg* relative to *Stoc-Spd* could be because of the following reasons:

- Speeds for the second day were generated by application of equation (4.29). Thus, the extra information provided by (4.29) was extremely valuable for model *Stoc-Spd* (since this was *true* information). Relatively, the information provided by (4.29) for *Stoc-Asg* was not as useful since the transition equation (4.31) used by *Stoc-Asg* was not completely consistent with the data generation procedure.

- Because of computational reasons, all the assignment fractions could not be estimated. Only those corresponding to high O-D flows (234 in number) were considered stochastic.

- Round-off errors could have contributed to inaccuracies in the final results.

The conclusion to be drawn from Figure 5-15 is that a good knowledge of the underlying process (both in terms of the structure of the transition equations, as well as in the error covariances) that describes temporal evolution of speeds or assignment fractions is needed in order for these more complicated models to perform better. In this context, an advantage of the stochastic speeds approach over the stochastic assignment fractions approach is that since speeds are directly observed, it could be easier to calibrate more complicated transition equations for this approach. A reassuring result – particularly for the stochastic speeds approach – in applying these models is that the impact of nonlinearities in the measurement equation do not appear to be the source of biases.

We finally compare results from the smoothed and estimated models in Figure 5-16. The first pair of bars are for the case $\delta_{cts} = 0.6$, $\delta_{spd} = 0$ while the last two are for $\delta_{cts} = 0.1$, $\delta_{spd} = 0.9$. We see that some gain is realized from using smoothed estimates. We also observe a reduction (albeit small) in the variance of the estimated O-D flows in Figure 5-17.

5.4 Major Findings

We conclude the chapter with a summary of the major findings from the three case studies.

1. Estimated O-D flow values for virtually all the models tested are substantially better (in the sense of being closer to the true values) compared to the corresponding historical values. Also, the estimation procedure seems to be fairly robust with respect to quality of historical information.

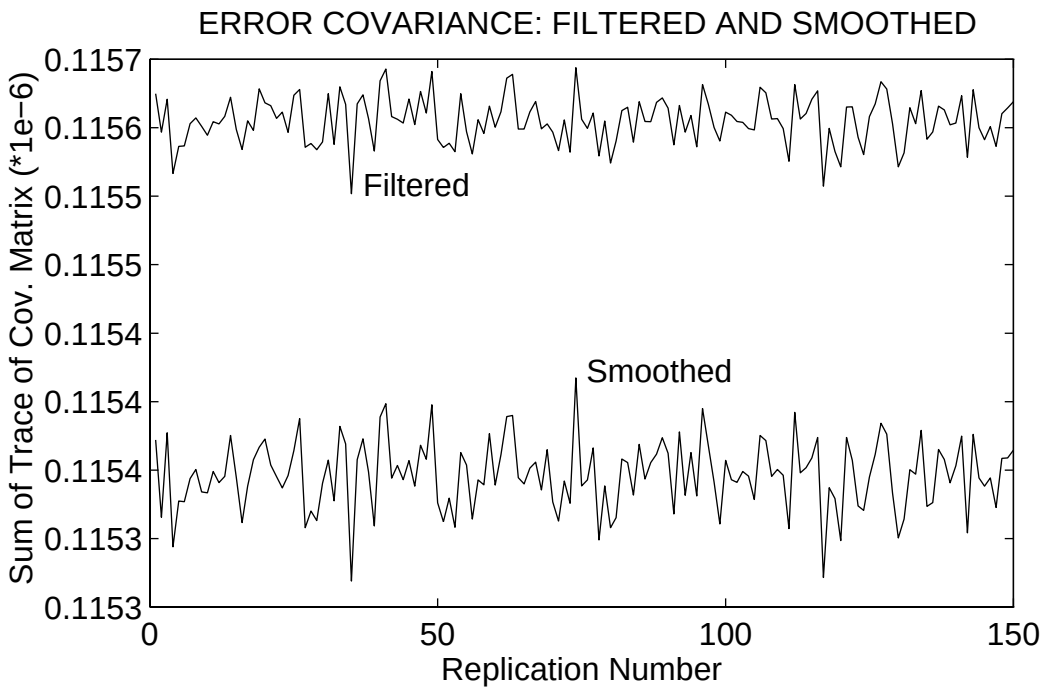

Figure 5-17: Decrease in Variance due to Smoothing

2. The one-step and two-step predictions perform better than the corresponding historical values for O-D pairs with high O-D flows. If there is not much variability between the historical O-D flows and the actual flows during the period of analysis, the predictions are not much better than the historical values. There is not much to be gained from three step predictions over the historical values. The predicted values tend to converge to the historical values with increasing prediction horizon.

3. The approximation introduced in the measurement equation by Model *Base-Appx* has only a slight impact on quality of estimated O-D flows relative to *Base*. Because of the computational savings, it seems the preferred method for real-time implementation.

4. Model *Base-Appx* is fairly robust with respect to measurement error in link counts and speeds. However, the bias due to inaccuracies in measuring speeds

can be significant at high levels of error.

5. Formulations that are based on modeling departure rates and shares separately perform better than their linear counterparts in predictions. The non-linear models tend to converge fairly quickly to reasonable values.

6. Smoothed O-D Flows are in general better than estimates, both in terms of reducing the RMS error as well as in reducing the variance of the estimates. These are therefore the preferred models for offline estimation.

7. In offline estimation, generalized models that incorporate a stochastic assignment matrix do not show much improvement over conventional models. As mentioned earlier, this could be because of the linear network structure in the I-880 case study. The modified offline models *Off-Mod-Base* that estimate O-D flows corresponding to multiple departure intervals simultaneously offer improvement over *Off-base*.

8. In real-time estimation and prediction, the two models with stochastic assignment matrix show mixed performance. A critical factor in the success of such models seems to be good understanding of the transition dynamics and the associated variance-covariance matrices.

9. The models with stochastic assignment matrices are computationally more intensive than the conventional models. In practical applications therefore, a judicious choice of additional decision variables (travel times, route-choice fractions and assignment parameters) should be made.

In the next chapter, we describe other implementation issues related to the model system.

Chapter 6

Conclusion

We begin this chapter with an assessment of the contribution of this research to the state of the art of O-D estimation and prediction models. We then provide a brief discussion of some practical issues that arise in implementation of such models. We conclude with further suggestions for future research.

6.1 Contribution to state of the art

This research represents an advancement of the state of the art in estimation and prediction of time-dependent Origin-Destination flows. All the models developed in this thesis can be applied to open networks and represent significant modeling improvements over their predecessors. Specifically,

- The framework developed here allows for incorporation of multiple sources of information with different degrees of reliability in a natural fashion. Many of the existing models that use limited sources of information can be expressed as special cases lying within the framework. Estimation and Prediction are conducted simultaneously within the same framework. The framework is used to address both the real-time and offline problems.

- An alternate way of formulating the real-time estimation and prediction problem in terms of originating trips and destination shares has been developed. This

formulation is based on the empirical observation that destination shares show greater within-day stability compared to the originating trips. The resulting models have shown improved predictive ability over the conventional models.

- It is shown that models based on a deterministic assignment matrix lead to biased and inconsistent estimates of the O-D flows since the assignment matrix is usually computed with error. As a remedy, two models based on a stochastic mapping between counts and O-D flows have been developed. This represents a fundamental extension of existing work.

- To improve computational aspects of these models, an approximation based on estimating each O-D flow only once – the first time it is measured – has been developed and tested.

- The models have been subjected to rigorous empirical testing based on a combination of actual and synthetic traffic data with encouraging results. With this elaborate process of validation, the model system is ready for implementation within a prototype traffic management system and for offline planning applications.

6.2 Application Issues

In this section, we focus on issues related to implementation of the models within a dynamic traffic management system.

6.2.1 Estimation Interval

An important issue in dynamic estimation is the choice of estimation interval. Of primary relevance here is the time granularity required for the application (for example, a DTA) that makes use of these matrices. Also, if the time intervals involved are very short, the predictability of the autoregressive process would be reduced since over very small intervals of time, fluctuations in traffic movements are essentially random.

Very large estimation intervals on the other hand hold out the danger of masking the information contained in time-varying link counts. And finally, computational considerations are also important in choice of estimation interval because for small intervals, the number of lags to be considered (and hence the dimensionality of the augmented state vector) would be high. Our empirical work indicates that as a rule of thumb, an interval of 10-15 minutes seems to be reasonable.

6.2.2 Computing the Assignment Matrix

At various points in this thesis, we have dealt with the assignment matrix – a crucial input into the O-D estimation and prediction process. When travel times in the network are observable (and route-choice fractions can be estimated), the analytical expressions given in Chapter 4 may be used to compute the matrix. In addition, to accommodate considerations of erroneous travel times or route-choice fractions, a stochastic assignment matrix based model might be preferred. There would be situations, however, where the available surveillance system only allows for measurement of link counts. Under such a scenario, the assignment matrix would be obtained through an iterative application of the O-D estimation and prediction module and a DTA.

To see how the iterative scheme would proceed, consider a time instant t which corresponds to the end of the departure time-interval h and the beginning of departure time-interval $h+1$. At this point in time, the O-D module gets a set of link counts for the departure interval h from the surveillance system. Further it has available, a set of trial assignment matrices $\mathbf{a}_h^h$, $\mathbf{a}_h^{h-1}$, $\mathbf{a}_h^{h-2}$, ... , $\mathbf{a}_h^{h-p'}$. Using these counts and assignment matrices, the O-D estimation module computes filtered O-D flows for interval h (apart from updated estimates for prior intervals $h-1$,$h-2$, ... , $h-p'$). These filtered O-D flows may be quite different from the (one-step predicted) flows that were used to determine the assignment matrices in the first place. Hence, the newly filtered flow of interval h (and possibly also the newly updated flows of prior intervals) is reloaded again on to the DTA to get revised estimates of the assignment matrices. This cycle of filtering/updating–computation of assignment

matrices–filtering/updating continues till convergence is reached. Once convergence is attained, a one-step prediction is performed to generate estimates of O-D flows for interval $h+1$. This is now used along with the filtered/updated flows of intervals h, $h-1$, ... , $h+1-p'$ to get preliminary estimates of assignment matrices for interval $h+1$ and the process continues.

We note that in the event of an iterative solution technique, though convergence cannot be guaranteed, empirical study (Chapter 5) indicates that the filtering procedure is fairly robust with respect to the quality of assignment matrices and hence one would expect the quality of the O-D flow estimates to get better with each iteration. Also, wherever computational resources permit, one of the stochastic assignment matrix based methods (Chapter 4) should be used. In that case, one need not have to iterate until convergence – consistent estimates of O-D flows are obtained at each step.

6.2.3 Missing Measurements

Sensor failures are not uncommon occurrences. It is important, therefore, that the models are robust with respect to such failures. The framework presented in this thesis can easily accommodate missing observations arising as a result of sensor failures. The easiest way to do this would be to assign a default value (perhaps a historical average if one exists) and a very high measurement error variance to the missing observation. The solution procedure would automatically attach a low weight to the problematic observation and the estimation and prediction methodology would otherwise remain the same.

6.2.4 Computational Issues

Given the large size of most real-life traffic networks, computational considerations assume an important role in any practical implementation. In general, the computational costs of implementing various models proposed in this thesis seem to be a function primarily of the following four parameters:

1. Number of O-D pairs and measurements.

2. Spatial distribution of the network.

3. Congestion level in the network.

4. Degree of autoregressive process.

The number of O-D pairs is an obvious parameter since it is directly related to the dimensionality of the unknown vector to be estimated. The number of measurements dictates the size of the matrix to be inverted in the estimation process. The spatial distribution of the network is important because the number of lagged intervals in the measurement equation depends upon the maximum travel time between any two points in the network. This in turn depends upon whether the network is "clustered" or "dispersed". For the same number of O-D pairs therefore, one would expect the computational costs associated with a linear network (a freeway or arterial) to be higher compared to an urban network of a few intersections[1]. The congestion level in the network is also related to this fact. High congestion levels in the network would increase the maximum travel times and hence increase the number of lags. Similarly, a high order autoregressive process implies an augmented state of higher dimension.

We can conceive of either *modeling* or *numerical* devices to improve the computational performance of the O-D estimation/prediction models proposed in this thesis. We discuss various possibilities below.

Modeling Approximations

Approximate Model

In Section 2.10, we presented an approximate model that did not require re-estimation of already computed O-D flows. For networks with large number of O-D pairs, this seems to be the preferred approach for real-time estimation. For offline estimation. a full blown augmented state model could be used. Infact, the number of lags could vary anywhere from zero to p' depending on the computational resources available.

[1]Unless the latter experiences high congestion.

State Reduction

This might be accomplished in two ways. The first and most obvious is to define origins and destinations in an aggregate fashion. Another technique is based upon the fact that in any practical situation, an overwhelmingly large proportion of O-D flows are either zero or extremely small. It might be desirable in some situations therefore, to fix deviations in these flows to zero. The measurement equation of course has to be adjusted suitably by adding a term that would reflect the contribution of all the "constant" O-D flows to the observed link volumes. The additional term would be time-varying with estimates obtained from historical data. It is clear that there would be dramatic savings in computational requirements by employing this approach; whether this is achievable without significant loss in accuracy is a matter of empirical testing.

Spatial Decomposition

Another idea to reducing the computational load in the O-D estimation process is to split the network into subnetworks. This would split the big problem into several problems of significantly smaller size. The disadvantages in this approach of course are that (a) each smaller network has less information to work with and (b) O-D pairs would now be "locally" defined within each subnetwork. Regarding the former, a mechanism for interaction and information transfer between these subnetworks would appear to be the ideal solution. One way of doing this might be to incorporate spatial correlation factors in the transition equation (for each subnetwork) that relate O-D flows passing through a subnetwork in a particular time interval to O-D flows passing through adjacent "upstream" subnetworks in prior time intervals. Separate measurement equations would be specified for each subnetwork. All of this would introduce a significant complication in the modeling process but is computationally attractive since each subnetwork could be processed in parallel by different processors with some real-time information exchange between them during each interval.

Numerical Improvements

At question here are two issues – numerical robustness of the filtering equations and relative speeds of various algorithms.

The conventional approach to recursive estimation involves propagation of a state estimate and an error covariance matrix from stage to stage. For the system described by

$$\mathcal{X}_{h+1} = \mathbf{\Phi}_h \mathcal{X}_h + \mathbf{W}_h \tag{6.1}$$

$$\mathcal{Y}_h = \mathbf{A}_h \mathcal{X}_h + \mathbf{v}_h \tag{6.2}$$

with definitions as in Section 2.6, this involves the following equations:

$$\mathbf{\Sigma}_{h|h-1} = \mathbf{\Phi}_{h-1} \mathbf{\Sigma}_{h-1|h-1} \mathbf{\Phi}'_{h-1} + \mathcal{Q}_{h-1} \tag{6.3}$$

$$\mathbf{K}_h = \mathbf{\Sigma}_{h|h-1} \mathbf{A}'_h (\mathbf{A}_h \mathbf{\Sigma}_{h|h-1} \mathbf{A}'_h + \mathbf{R}_h)^{-1} \tag{6.4}$$

$$\mathbf{\Sigma}_{h|h} = \mathbf{\Sigma}_{h|h-1} - \mathbf{K}_h \mathbf{A}_h \mathbf{\Sigma}_{h|h-1} \tag{6.5}$$

$$\hat{\mathcal{X}}_{h|h-1} = \mathbf{\Phi}_{h-1} \hat{\mathcal{X}}_{h-1|h-1} \tag{6.6}$$

$$\hat{\mathcal{X}}_{h|h} = \hat{\mathcal{X}}_{h|h-1} + \mathbf{K}_h (\mathcal{Y}_h - \mathbf{A}_h \hat{\mathcal{X}}_{h|h-1} - \mathcal{B}_{h-1}) \tag{6.7}$$

In several practical problems, propagation of the error covariance matrix by means of Equations (6.3) and (6.5) results in a matrix which is not positive semidefinite. This may occur, for example, when a particular linear combination of state vector components is known with great precision[2] while other combinations are less observable. This has given rise to development of alternate recursive relationships that propagate a state estimate and a *square root* error covariance instead[3]. In other words, the *square root* filter involves replacement of the covariance matrix $\mathbf{\Sigma}$ by the square root

[2]For example, the combination of O-D flows constituting an observable entry ramp flow

[3]The square root covariance $\mathbf{S}$ is defined by the following relationship:

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{S}' \tag{6.8}$$

$\mathbf{S}$ is not uniquely determined by this relationship. This lack of uniqueness is not generally a problem as a unique square root may be defined, for example, by Cholesky decomposition.

covariance $\mathbf{S}$, then replacing Equations (6.3) and (6.5) by equivalent relationships for propagating the square root. This approach is motivated by two considerations: (a) The product $\mathbf{SS}'$ can never be indefinite even in the presence of roundoff errors, while roundoff errors sometimes cause the computed value of $\mathbf{\Sigma}$ to be indefinite; (b) the numerical conditioning of $\mathbf{S}$ is generally much better than that of $\mathbf{\Sigma}$. In this spirit, several recursive square root solutions have been proposed for the filtering and smoothing problems[4].

Square root algorithms impose some additional computational burden over the conventional algorithm. While there has been significant work is designing faster versions of these (see for example Bierman[11]), the problem still remains computationally challenging. For O-D Estimation and Prediction, it might be possible to realize significant gains by recognizing that most of the large matrices involved in the model are likely to be very sparse. This highlights the need for design of efficient data structures for handling and storage of sparse matrices in computation.

6.2.5 An Ongoing Application

We conclude this section with a brief description of an ongoing application of models presented in this thesis. A Dynamic Traffic Assignment (DTA) system to support real-time applications such as dynamic route guidance and adaptive traffic control is currently being developed[37]. The main components of this DTA are as follows:

- Real-time O-D estimation and prediction
- Network tracking simulator to continuously assess the current state of the network
- Network performance simulator to anticipate future traffic conditions
- Route guidance generator based on predicted traffic conditions.

The DTA is designed to reside in Traffic Management Centers.

[4] A review and references might be found in Sorenson[41].

For real-time O-D estimation and prediction, the model being implemented in the initial prototype is the approximate model described by equations (2.49) and (2.50), i.e.,

$$\mathbf{y}_h = \mathbf{a}_h^h(\mathbf{x}_h - \mathbf{x}_h^H) + \mathbf{b}_h + \mathbf{v}_h \tag{6.9}$$

and

$$\mathbf{x}_{h+1} - \mathbf{x}_{h+1}^H = \mathbf{f}_{h+1}^h(\mathbf{x}_h - \mathbf{x}_h^H) + \mathbf{c}_{h+1} + \mathbf{w}_{h+1} \tag{6.10}$$

where all terms have the same meaning as before. The assignment matrix is obtained from the network tracking simulator, possibly by successive iterations as explained in Section 6.2.2.

For numerical robustness, a square root algorithm is proposed. The candidate algorithm is given by the following series of steps[17]:

1. Compute $\mathbf{J}_0^0 = (Var(\partial\mathbf{x}_0))^c$, where the superscript c denotes the Cholesky square root.

2. For h = 1,2,..., compute

$$\mathbf{J}_h^{h-1} = [\mathbf{f}_h^{h-1}\mathbf{J}_{h-1}^{h-1} \quad \mathbf{Q}_{h-1}^c] \tag{6.11}$$

$$\mathbf{G}_h = (\mathbf{a}_h^h\mathbf{J}_h^{h-1}\mathbf{J}_h^{h-1\prime}\mathbf{a}_h^{h\prime} + \mathbf{R}_h)^c \tag{6.12}$$

$$\mathbf{J}_h^h = \mathbf{J}_h^{h-1}[\mathbf{I} - \mathbf{J}_h^{h-1\prime}\mathbf{a}_h^{h\prime}(\mathbf{G}_h')^{-1}(\mathbf{G}_h + \mathbf{R}_h^c)^{-1}\mathbf{a}_h^h\mathbf{J}_h^{h-1}] \tag{6.13}$$

3. Compute $\hat{\partial\mathbf{x}}_{0|0} = E(\bar{\mathbf{x}}_0 - \mathbf{x}_h^H)$ and for h = 1,2,..., using the information from step (2), compute the matrix:

$$\mathbf{K}_h = \mathbf{J}_h^{h-1}\mathbf{J}_h^{h-1\prime}\mathbf{a}_h^{h\prime}(\mathbf{G}_h')^{-1}\mathbf{G}_h^{-1} \tag{6.14}$$

and

$$\hat{\partial\mathbf{x}}_{h|h-1} = \mathbf{f}_h^{h-1}\hat{\partial\mathbf{x}}_{h-1|h-1} + \mathbf{c}_h \tag{6.15}$$

$$\hat{\partial\mathbf{x}}_{h|h} = \hat{\partial\mathbf{x}}_{h|h-1} + \mathbf{K}_h(\mathbf{y}_h - \mathbf{a}_h^h\hat{\partial\mathbf{x}}_{h|h-1} - \mathbf{b}_h) \tag{6.16}$$

In the above equations, the matrices $\mathbf{J}$ represent the *square root* covariance of the state, i.e., $\mathbf{J}_h^{h-1}\mathbf{J}_h^{h-1'} = \mathbf{\Sigma}_{h|h-1}$ and $\mathbf{J}_h^h\mathbf{J}_h^{h'} = \mathbf{\Sigma}_{h|h}$. These definitions imply that the matrix $\mathbf{K}_h$ in Equation (6.14) is identical to the gain matrix of earlier chapters. We notice that we only have to invert triangular matrices and in addition, these matrices are square roots of the ones which might have very large or very small entries.

Since the O-D module in this application continuously interacts with the tracking and performance simulators, efficient data structures for storing and manipulating common data items such as O-D flows and assignment matrices are needed. In designing these data structures, there are two important considerations. First, these matrices are highly sparse. Second, the data structures should be designed taking into account the needs of the application that uses the data. For example, the data structure holding the O-D matrix should enable efficient iteration over origins and destinations. The assignment matrix should allow for efficient iteration over O-D pair r and departure interval p given a sensor l and an estimation interval h. Candidate data structures that satisfy these considerations are shown schematically in Figure 6-1. For the O-D matrix, only non-zero flows are stored. For each origin, a list of destinations with non-zero flows can be accessed sequentially. Likewise, for each destination, all the origins with non-zero flows can be accessed efficiently. For the assignment matrix, each link l (for a given estimation interval h) is connected to a list of prior departure intervals of size $p'+1$. Each element p of this list is further connected to another list containing non-zero contributions from O-D pairs r that departed during p. This structure permits efficient access to, and summation of, contributions of the vehicle groups (r,p) to the link flows (l,h).

The system is being implemented using object oriented techniques using the programming language C++.

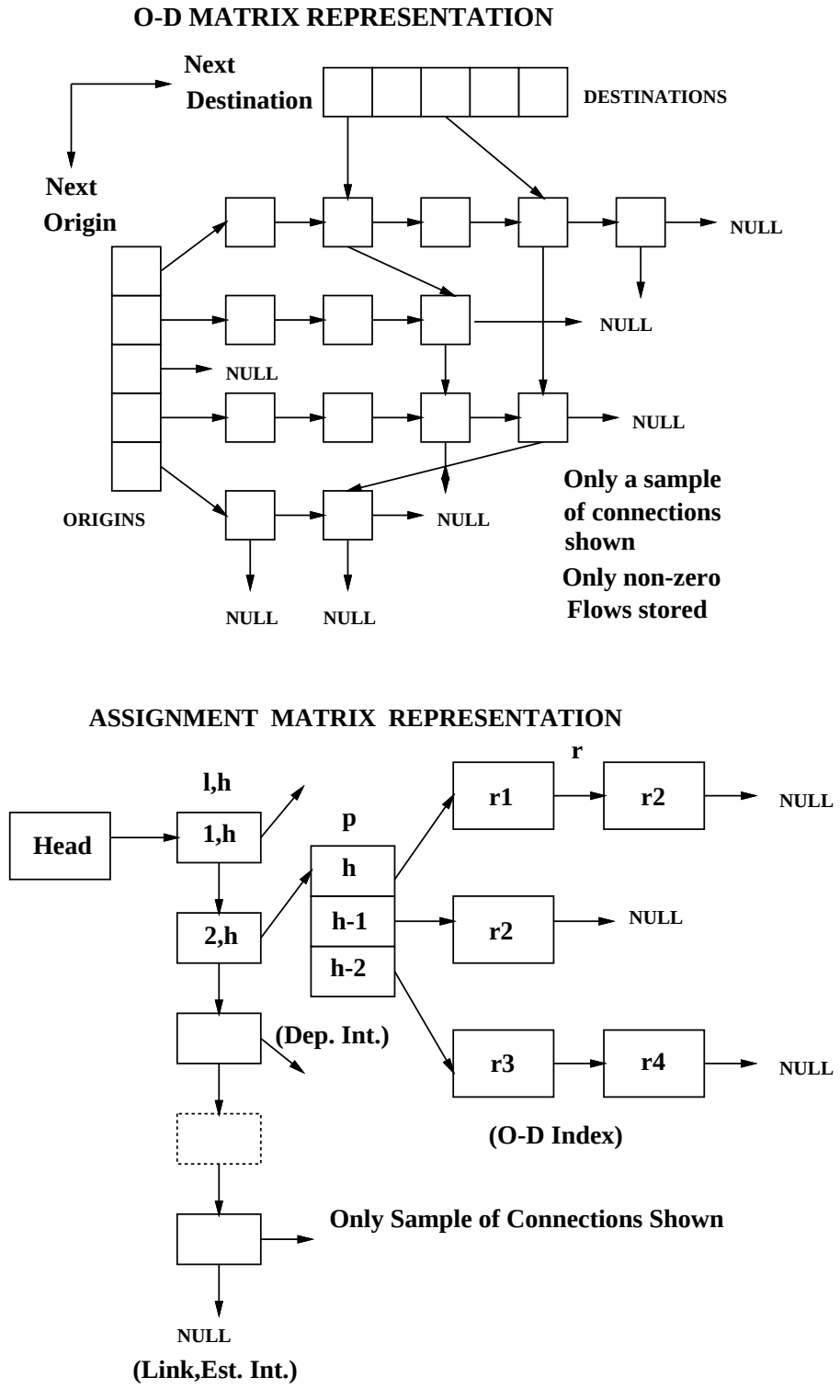

Figure 6-1: Proposed Data Structures

6.3 Further Research

6.3.1 O-D Prediction and Traveler Information

In a Dynamic Traffic Management System, pre-trip (prediction based) information could be available to travelers. This information would presumably affect departure time choice (as well as route-choice) over future time intervals. The prediction of O-D flows in various models proposed in this thesis is by means of an autoregressive process on deviations (in either O-D flows or trips and shares) that does not explicitly account for the effect of information[5]. Arguably, the prediction could benefit from a knowledge of the information about future traffic conditions provided to travelers during the current time-interval.

Modeling the impact of information on O-D prediction is non-trivial. One way of handling this would be to include additional terms in the transition equation that reflect the aggregate effect of switches from the "desired" departure times of all the trips in the network due to information. While this is the most direct approach to the problem, it is unclear what the form of the additional terms should be. Moreover, it is difficult to isolate the effects of the additional terms from the already existing terms[6]. And finally, the information itself could be based partly on the O-D predictions resulting in a fixed point problem.

A second technique to accommodate the effect of information might be to adaptively estimate the auto-correlation fractions in real-time[7]. This can be thought of as a purely "statistical" approach to the problem. The primary disadvantage of this approach is that it vastly increases the size of the state vector to be estimated. Also, Equations (2.15) and (3.7) now become non-linear, again increasing the computational load.

[5] If the nature and content of information does not vary by much day over day and in the absence of incidents or other unexpected events, the models *indirectly* capture the effect of information. This is because the autocorrelation matrices are calibrated from historical O-D flows that, in turn, are estimated from historical link volumes. The link volumes, in turn, reflect the effect of information provided (on the historical day).

[6] See previous footnote.

[7] The fractions estimated during prior intervals (and days) would be used as measurements in this approach.

A third technique to incorporate the effect of information is to adjust the historical matrices $\mathbf{x}_h^H$ by a separate process *before* the estimation and prediction. This technique is being employed in the application described in Section 6.2.5. The adjustment is done by means of discrete choice models that compute, for all travelers departing over a future rolling horizon, the probability of switching departure times from their habitual departure intervals. These probabilities are then used to adjust the historical database over the rolling horizon. This procedure is flexible in that it can also handle decisions regarding changes in mode (to public transportation), trip cancellation, habitual route, etc. in response to information. It however requires information on individual characteristics (such as socio-economic characteristics) for each traveler in the historical database, or at least a distribution of these over the driver population. Moreover, the autoregressive coefficients would have to be calibrated on the "adjusted" historical matrices.

6.3.2 The Assignment Matrix

Additional forms should be investigated for the transition equations (4.29), (4.30) and (4.31) that describe the temporal evolution of assignment fractions, or of the fundamental parameters constituting these such as the speeds, travel times, and route-choice fractions. It might also be possible to use information from turning fractions at intersections (which might be based on empirical observation) in estimating these fractions.

6.3.3 Mode Choice

This thesis has been about *vehicle trips*. No distinction has been made between individual and car-pool trips or between auto and public transportation. For the purpose of implementing an O-D prediction model within a DTA, it might also be desirable to incorporate the effect of mode choice. This need arises because O-D predictions could be affected by travelers switching from auto to public transportation (or vice-versa) in response to information provided by a DTA. It might be possible,

therefore, to realize improvements in prediction by enhancing the framework in this thesis with a mode choice model. This could be done by adjusting the historical matrices exactly in the same manner as for departure time choice.

6.3.4 Empirical Testing on Urban Networks

Due to limitations of existing data, the framework and models developed in this thesis remain to be applied to urban networks. Urban networks present unique problems unlike any encountered in the freeway case studies in this thesis. Some of the distinguishing characteristics are as follows:

- Absence of a clear set of origins and destinations: Unlike in the freeway case studies in this thesis, there is no obvious or easy way of defining origins and destinations in a general urban network. Clearly, it is impractical to define every parking lot or office building as an origin or destination. Equally, a very aggregate definition would involve significant approximations and could result in ignoring short trips.

- Large number of O-D pairs coupled with a relatively sparse surveillance system.

- Need for a *parking* model: On urban streets, vehicles often circle urban blocks in search of parking spaces. This might lead to vehicles being counted multiple times resulting in over-estimation of O-D flows.

- Larger set of routes for each O-D pair, more route-switching opportunities and therefore, a greater need for a fully calibrated route-choice model.

6.3.5 Evaluation of the Model System

Another important set of research issues to be addressed relate to the evaluation of O-D estimation and prediction models. As we have mentioned before, true O-D flows are seldom observed. Moreover, in most planning and traffic management applications, estimation and prediction of O-D flows is not an end in itself, but a necessary step to determining network performance measures such as link flows,

queue lengths, travel times, fuel consumption, etc. Given these objectives, one way of evaluating the accuracy of the O-D models is a *joint* test of their performance with a DTA, i.e., comparing the estimates and predictions of link flows, travel times, etc. from the joint O-D/DTA models with those observed by the surveillance system.

This might be taken a step further. Together, O-D Estimation/Prediction and DTA constitute an overall traffic prediction capability within a dynamic traffic management system. The other important component of such a system is Traffic Control. Prediction and Control are clearly interdependent and contribute *collectively* towards the performance of the system. An important research question that has not been addressed completely yet, pertains to the sensitivity of performance of this collective system to errors in traffic prediction[8] – these errors could be either in O-D prediction or in the DTA.

A promising approach to evaluating different O-D estimation and prediction models jointly with a DTA and control system is use of a *simulation laboratory*([48],[49]). Yang et al.[49] use a microscopic simulator, MITSIM, to track and move individual vehicles on the network. They also use a mesoscopic simulator that is similar in structure to a DTA as a traffic predictor (given predicted O-D flows as input). Such an evaluation framework can be directly applied to gauge the sensitivity of the performance of a dynamic traffic management system to the extent of estimation/prediction errors in time-dependent O-D flows.

6.4 Conclusion

A comprehensive framework for estimation and prediction of time-dependent O-D flows has been presented in this thesis. The models developed here have been tested using actual traffic data from different sources. Results obtained thus far are encouraging and indicate that the model system is robust with respect to quality of inputs and is ready for prototype implementation.

[8]Ben-Akiva et al.[8] provide some results for route-guidance based on an analytic study of a two route network. Van Toorenburg et al.[45] investigate conditions under which some types of predictive control are useful.

Appendix A

State Space Modeling

In this appendix we give a brief overview of State Space Modeling and the Kalman Filter. For more extensive coverage of the material, the reader is referred to Gelb[20].

A.1 The model

The State Space model typically consists of two equations – the *measurement*[1] equation and the *transition*[2] equation.

$$\textit{Measurement Equation} : \mathbf{y}_h = \mathbf{A}_h \mathbf{x}_h + \mathbf{v}_h \tag{A.1}$$

$$\textit{Transition Equation} : \mathbf{x}_{h+1} = \mathbf{F}_h \mathbf{x}_h + \mathbf{w}_h \tag{A.2}$$

where $\mathbf{x}_h$ is the vector that represents the latent "true state" of the system during interval h. $\mathbf{y}_h$ is a vector of observations made in interval h. $\mathbf{A}_h$ and $\mathbf{F}_h$ are known matrices. $\mathbf{w}_h$ and $\mathbf{v}_h$ are vectors of random errors.

We typically make the following assumptions about the model:

1. $\{\mathbf{v}_h\}$ and $\{\mathbf{w}_h\}$ are *independent*, *zero mean*, *gaussian*[3] processes with

$$E[\mathbf{v}_h \mathbf{v}_l'] = \mathbf{R}_h \delta_{hl} \text{and } E[\mathbf{w}_h \mathbf{w}_l'] = \mathbf{Q}_h \delta_{hl} \quad \delta_{hl} = 1 \text{ if } h = l \text{ and } 0 \text{ otherwise.}$$

[1] Or *observation*
[2] Or *system*
[3] Normality assumption made only for ease of exposition.

2. Initial system state $\mathbf{x}_0$ is *gaussian*[4] with mean $\bar{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$ and is *independent* of $\mathbf{v}_h$ and $\mathbf{w}_h \forall h = 0, 1, ...$

Given these assumptions, the *filtering* problem is to estimate the quantity $\hat{\mathbf{x}}_{h|h} = E[\mathbf{x}_h|\mathbf{Y}_h]$ where $\mathbf{Y}_h$ denotes the set $\{\mathbf{y}_0, \mathbf{y}_1, ... ,\mathbf{y}_h\}$. A *one-step prediction* problem is to estimate the quantity $\hat{\mathbf{x}}_{h|h-1} = E[\mathbf{x}_h|\mathbf{Y}_{h-1}]$. Finally, the *smoothing* problem is to estimate the quantity $\hat{\mathbf{x}}_{h|T} = E[\mathbf{x}_h|\mathbf{Y}_T]$ where $T > h$.

The applications of such models are extremely diverse, ranging from spacecraft orbit determination(See Campbell[12] for example) to predicting cattle populations[32]. In the oft-quoted example of satellite tracking, the state vector $\mathbf{x}$ could consist of the position (in terms of a spherical coordinate system with center at the center of the earth) and velocity of the satellite. Since it is not possible to measure these directly, the measurements (the $\mathbf{y}$'s) made are angles and distances from tracking stations around the surface of the earth. The laws of geometry that map these angles and distances into the state coordinates are embedded in the matrix $\mathbf{A}$. The errors in measuring the $\mathbf{y}$'s are modeled by $\mathbf{v}$. Also, the laws of physics for orbiting bodies predict the movement of the satellite with time – these laws are incorporated in the matrix $\mathbf{F}$. Deviations from these laws due to for example the non-uniform gravitational field of the earth are allowed for by the error term $\mathbf{w}$.

A.2 The Kalman Filter

A.2.1 Derivation

Given a prior estimate of the state of the system at time h denoted by $\hat{\mathbf{x}}_{h|h-1}$, we wish to obtain an updated estimate $\hat{\mathbf{x}}_{h|h}$ after measurement $\mathbf{y}_h$ is known[5]. Further to avoid storing past measurements, we seek an estimator in the following form.

$$\hat{\mathbf{x}}_{h|h} = \mathbf{K}_h^1 \hat{\mathbf{x}}_{h|h-1} + \mathbf{K}_h^2 \mathbf{y}_h \tag{A.3}$$

[4] Again an assumption not strictly necessary.

[5] The filtering equations shall be derived for the *discrete-time* case. The presentation here is based on [20]. Readers interested in a more rigorous derivation are referred to [28].

where $\mathbf{K}_h^1$ and $\mathbf{K}_h^2$ are two time-varying weighting matrices as yet unspecified. Let us define *estimation errors* before and after the measurement $\mathbf{y}_h$ by the following relations.

$$\begin{aligned} \hat{\mathbf{x}}_{h|h} &= \mathbf{x}_h + \tilde{\mathbf{x}}_{h|h} \\ \hat{\mathbf{x}}_{h|h-1} &= \mathbf{x}_h + \tilde{\mathbf{x}}_{h|h-1} \end{aligned} \tag{A.4}$$

where $\tilde{\mathbf{x}}_{h|h-1}$ represents the error in the estimate of $\mathbf{x}_h$ made *prior* to recording the measurement $\mathbf{y}_h$ and $\tilde{\mathbf{x}}_{h|h}$ the error in the estimate of $\mathbf{x}_h$ made *after* the measurement. Substituting the measurement equation and equation (A.3) into equations (A.4), we obtain

$$\tilde{\mathbf{x}}_{h|h} = (\mathbf{K}_h^1 + \mathbf{K}_h^2\mathbf{A}_h - \mathbf{I})\mathbf{x}_h + \mathbf{K}_h^1\tilde{\mathbf{x}}_{h|h-1} + \mathbf{K}_h^2\mathbf{v}_h \tag{A.5}$$

where $\mathbf{I}$ is the identity matrix. By assumption, $E[\mathbf{v}_h] = 0$. Also, if $E[\tilde{\mathbf{x}}_{h|h-1}] = 0$ the estimator we desire will be unbiased (i.e., $E[\tilde{\mathbf{x}}_{h|h}] = 0$) *only* if

$$\mathbf{K}_h^1 = \mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h \tag{A.6}$$

implying that the estimator can be written as

$$\hat{\mathbf{x}}_{h|h} = (\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)\hat{\mathbf{x}}_{h|h-1} + \mathbf{K}_h^2\mathbf{y}_h \tag{A.7}$$

or alternatively,

$$\hat{\mathbf{x}}_{h|h} = \hat{\mathbf{x}}_{h|h-1} + \mathbf{K}_h^2(\mathbf{y}_h - \mathbf{A}_h\hat{\mathbf{x}}_{h|h-1}) \tag{A.8}$$

Error Covariance Updates

Before proceeding to choose the weighting matrix $\mathbf{K}_h^2$, let us consider the error covariance updates. Using the definition $\mathbf{\Sigma}_{h|h} = E[\tilde{\mathbf{x}}_{h|h}\tilde{\mathbf{x}}_{h|h}^{'}]$ and equation (A.8),

$$\begin{aligned} \mathbf{\Sigma}_{h|h} &= E\{(\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)\tilde{\mathbf{x}}_{h|h-1}(\tilde{\mathbf{x}}_{h|h-1}^{'}(\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)^{'} + \mathbf{v}_h^{'}\mathbf{K}_h^{2^{'}}) + \\ &\quad \mathbf{K}_h^2\mathbf{v}_h(\tilde{\mathbf{x}}_{h|h-1}^{'}(\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)^{'} + \mathbf{v}_h^{'}\mathbf{K}_h^{2^{'}})\} \end{aligned} \tag{A.9}$$

Using the definitions $\mathrm{E}[\tilde{\mathbf{x}}_{h|h-1}\tilde{\mathbf{x}}_{h|h-1}^{'}] = \mathbf{\Sigma}_{h|h-1}$, $\mathrm{E}[\mathbf{v}_h\mathbf{v}_h^{'}] = \mathbf{R}_h$ and the fact that

$$E[\tilde{\mathbf{x}}_{h|h-1}\mathbf{v}_h^{'}] = E[\mathbf{v}_h\tilde{\mathbf{x}}_{h|h-1}^{'}] = 0^6 \tag{A.10}$$

we obtain the following expression for the error covariance.

$$\mathbf{\Sigma}_{h|h} = (\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)\mathbf{\Sigma}_{h|h-1}(\mathbf{I} - \mathbf{K}_h^2\mathbf{A}_h)^{'} + \mathbf{K}_h^2\mathbf{R}_h\mathbf{K}_h^{2^{'}} \tag{A.11}$$

Choice of $\mathbf{K}_h^2$

Let us first dispense with the superscript and simply write $\mathbf{K}_h^2$ as $\mathbf{K}_h$. The criterion for choosing this matrix is to minimize the sum of squared errors $\tilde{\mathbf{x}}_{h|h}$. In other words, we choose the function to be minimized as:

$$\mathbf{J}_h = trace[\mathbf{\Sigma}_{h|h}] \tag{A.12}$$

Partially differentiating $\mathbf{J}_h$ with respect to $\mathbf{K}_h$ and setting it to zero,

$$\frac{\partial}{\partial\mathbf{K}_h}(trace[\mathbf{\Sigma}_{h|h}]) = 0 \tag{A.13}$$

From equation (A.11) and using the relationship

$$\frac{\partial}{\partial\mathbf{A}}(trace[\mathbf{ABA}^{'}]) = 2\mathbf{AB} \tag{A.14}$$

for a symmetric matrix $\mathbf{B}$ we get

$$\frac{\partial}{\partial\mathbf{K}_h}\mathbf{J}_h = -2(\mathbf{I} - \mathbf{K}_h\mathbf{A}_h)\mathbf{\Sigma}_{h|h-1}\mathbf{A}_h^{'} + 2\mathbf{K}_h\mathbf{R}_h = 0 \tag{A.15}$$

which yields the result

$$\mathbf{K}_h = \mathbf{\Sigma}_{h|h-1}\mathbf{A}_h^{'}(\mathbf{A}_h\mathbf{\Sigma}_{h|h-1}\mathbf{A}_h^{'} + \mathbf{R}_h)^{-1} \tag{A.16}$$

[6] This can be shown to follow from the assumptions that the measurement error $\mathbf{v}_h$ is uncorrelated over time and uncorrelated with the error in the transition equation.

This matrix $\mathbf{K}_h$ is called the *Kalman gain* matrix. One can verify by differentiating Equation (A.15) that the Hessian of $\mathbf{J}_h$ (given by $\partial^2 \mathbf{J}_h / \partial \mathbf{K}_h^2$) is positive semidefinite confirming that this value of $\mathbf{K}_h$ does indeed minimize $\mathbf{J}_h$.

Substituting this value of $\mathbf{K}_h$ in (A.11), we obtain after some algebraic manipulation, a recursive formula for the error covariance as

$$\mathbf{\Sigma}_{h|h} = (\mathbf{I} - \mathbf{K}_h \mathbf{A}_h)\mathbf{\Sigma}_{h|h-1} \tag{A.17}$$

By using the transition equation, it is straightforward to show that the one-step predicted estimate for the state vector would be given by

$$\hat{\mathbf{x}}_{h+1|h} = \mathbf{F}_h \hat{\mathbf{x}}_{h|h} \tag{A.18}$$

and the corresponding error covariance behaviour by

$$\mathbf{\Sigma}_{h+1|h} = \mathbf{F}_h \mathbf{\Sigma}_{h|h} \mathbf{F}_h^{'} + \mathbf{Q}_h \tag{A.19}$$

This completes our derivation. Further, it may be stated that general k step predictions would be given by

$$\hat{\mathbf{x}}_{h+k|h} = (\mathbf{F}_h)^k \hat{\mathbf{x}}_{h|h} \tag{A.20}$$

and the corresponding variances would be computed by recursive application of equations (A.19) and (A.20).

Summary

To summarize the equations comprising the Kalman Filter,

$$\begin{aligned}
\mathbf{\Sigma}_{0|0} &= \mathbf{P}_0 \\
\mathbf{\Sigma}_{h|h-1} &= \mathbf{F}_{h-1} \mathbf{\Sigma}_{h-1|h-1} \mathbf{F}'_{h-1} + \mathbf{Q}_{h-1} \\
\mathbf{K}_h &= \mathbf{\Sigma}_{h|h-1} \mathbf{A}'_h (\mathbf{A}_h \mathbf{\Sigma}_{h|h-1} \mathbf{A}'_h + \mathbf{R}_h)^{-1} \\
\mathbf{\Sigma}_{h|h} &= \mathbf{\Sigma}_{h|h-1} - \mathbf{K}_h \mathbf{A}_h \mathbf{\Sigma}_{h|h-1}
\end{aligned} \tag{A.21}$$

$$
\begin{aligned}
\hat{\mathbf{x}}_{0|0} &= \bar{\mathbf{x}}_0 \\
\hat{\mathbf{x}}_{h|h-1} &= \mathbf{F}_{h-1}\hat{\mathbf{x}}_{h-1|h-1} \\
\hat{\mathbf{x}}_{h|h} &= \hat{\mathbf{x}}_{h|h-1} + \mathbf{K}_h(\mathbf{y}_h - \mathbf{A}_h\hat{\mathbf{x}}_{h|h-1}) \\
h &= 1, 2, \ldots
\end{aligned}
$$

A.2.2 Important Properties

Some of the important properties of the filter are:

- The filter produces the smallest Mean Square Error (MSE) covariance matrix among a class of *linear* estimators *whether or not* the gaussian assumptions hold. If the gaussian assumptions hold, then the filter produces the smallest MSE among *all* estimators – linear or nonlinear.

- The estimate is unbiased and orthogonal to its error i.e. $\mathrm{E}[\hat{\mathbf{x}}_{h|h}\tilde{\mathbf{x}}'_{h|h}] = 0$.

- The filter has significant computational advantages; because of its recursive form, all previous information need not be stored. All historical information is subsumed in the previous estimate.

- The filter can be applied to non-linear systems with some modifications. Essentially, the non-linear system is approximated by linearizations during each time interval about the latest state estimate. This algorithm is referred to as the *Extended Kalman Filter*.

- There are several ways in which the assumptions proposed here could be relaxed. One could for instance allow the measurement and transition errors to be correlated. One could also allow these errors to be correlated over time, thereby relaxing the white noise assumption. Under certain conditions, one could avoid having to perform a matrix inversion in every time-step, thus eliminating what could be a computationally challenging task.

- If the system has not been modeled properly, it is possible that the usefulness of the filter may be nullified by a phenomenon known as *Divergence*. In this

phenomenon, after an extended period of operation of the filter, the errors in the estimate eventually diverge to values entirely out of proportion to that predicted by theory. In such cases usually, the calculated covariance matrix becomes unrealistically small, so that undue confidence is placed in the estimates and subsequent measurements are effectively ignored. A good collection of literature in this area may be found in Sorenson[41].

- One of the most important concerns while considering use of the filtering technique is the requirement that $\mathbf{P}_0$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{A}$, $\mathbf{F}$ and $\bar{\mathbf{x}}_0$ be exactly known. Since this may be unrealistic in most practical applications, no filter design is really optimal. This then raises the question of whether one could deduce non-optimal behavior during operation and improve the quality of filter performance. Within certain limits, this is possible and is known in literature as *Adaptive Filtering*. Several investigators have considered the effect of errors in $\mathbf{Q}$ and $\mathbf{R}$ on the performance of the filter. Several others have proposed on-line schemes to identify $\mathbf{Q}$ and $\mathbf{R}$. And finally, there also exist some techniques to compensate for incorrect choice of $\mathbf{P}_0$ and $\bar{\mathbf{x}}_0$. A detailed discussion of these issues, except as they pertain to our problem, is beyond the scope of this thesis. The interested reader is referred to Sorenson[41] for details.

Appendix B

Equivalence of Kalman Filtering and Generalized Least Squares

In this Appendix, we point out the intimate connection between Kalman and Least Squares Estimation theory. More precisely, we show that the application of results obtained by classical Generalized Least Squares (GLS) to discrete stochastic linear processes leads to the Kalman Filter. While this may be shown in several ways, we rely here on the presentation by Genin[21].

B.1 Generalized Least Squares

Consider the following equation of measurements:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{v} \tag{B.1}$$

where $\mathbf{z}$ is a m-dimensional vector of measurements, $\mathbf{x}$ a n-dimensional ($n < m$) constant state vector to be determined, $\mathbf{H}$ a $m * n$ matrix of maximal rank and $\mathbf{v}$ a m-dimensional vector of errors with zero mean and positive definite covariance matrix $\mathbf{R}$.

Then, the unbiased estimate $\hat{\mathbf{x}}$ of the unknown state $\mathbf{x}$ which is a linear combination of the measurements $\mathbf{z}$ and has the smallest variance for each of its components,

is given by[1]

$$\hat{\mathbf{x}} = (\mathbf{H}'\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{R}^{-1}\mathbf{z} \tag{B.2}$$

with covariance matrix

$$\mathbf{P} = (\mathbf{H}'\mathbf{R}^{-1}\mathbf{H})^{-1} \tag{B.3}$$

B.2 Recursive Estimation

Next, consider a partition of the measurements vector $\mathbf{z}$ in two subvectors $[\mathbf{z}_{k-1}, \mathbf{z}_k]$. Partitioning $\mathbf{v}$ and $\mathbf{H}$ accordingly, Equation (B.1) may be written as

$$\begin{aligned} \mathbf{z}_{k-1} &= \mathbf{H}_{k-1}\mathbf{x} + \mathbf{v}_{k-1} \\ \mathbf{z}_k &= \mathbf{H}_k\mathbf{x} + \mathbf{v}_k \end{aligned} \tag{B.4}$$

and we assume the set $\mathbf{v}_{k-1}$ to be uncorrelated with the set $\mathbf{v}_k$, so that $\mathbf{R}$ has the form

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_k \end{bmatrix} \tag{B.5}$$

Let $\hat{\mathbf{x}}_{k-1}$ and $\mathbf{P}_{k-1}$ be the minimum variance unbiased estimate and its associated covariance matrix, defined on the subset $\mathbf{z}_{k-1}$ only. We can then prove that the minimum variance unbiased estimate $\hat{\mathbf{x}}_k$ defined on the set of measurements $[\mathbf{z}_{k-1}, \mathbf{z}_k]$ may be obtained without reprocessing the subset $\mathbf{z}_{k-1}$ and is given by

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k-1} + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k-1}) \tag{B.6}$$

with $\mathbf{K}_k$ defined by

$$\mathbf{K}_k = \mathbf{P}_{k-1}\mathbf{H}_k'(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}_k' + \mathbf{R}_k)^{-1} \tag{B.7}$$

while the covariance matrix $\mathbf{P}_k$ is obtained by

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k-1} \tag{B.8}$$

[1] A proof of this result may be found in any statistics text.

To demonstrate the above relations, note that, in view of Equation (B.3), the inverse covariance matrix is

$$\begin{aligned}
\mathbf{P}_k^{-1} &= \mathbf{H}'\mathbf{R}^{-1}\mathbf{H} \\
&= \mathbf{H}'_{k-1}\mathbf{R}_{k-1}^{-1}\mathbf{H}_{k-1} + \mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{H}_k \\
&= \mathbf{P}_{k-1}^{-1} + \mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{H}_k
\end{aligned} \tag{B.9}$$

This expression has a classical form in matrix algebra, known as the *Frobenius* form, so that the inverse may be readily obtained using the *matrix inversion lemma*[2]:

$$\begin{aligned}
\mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{P}_{k-1}\mathbf{H}'_k(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}'_k + \mathbf{R}_k)^{-1}\mathbf{H}_k\mathbf{P}_{k-1} \\
&= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k-1}
\end{aligned} \tag{B.10}$$

On the other hand, in view of Equation (B.2), the new estimate $\hat{\mathbf{x}}_k$ has the form

$$\begin{aligned}
\hat{\mathbf{x}}_k &= \mathbf{P}_k\mathbf{H}\mathbf{R}^{-1}\mathbf{z} \\
&= \mathbf{P}_k(\mathbf{H}'_{k-1}\mathbf{R}_{k-1}^{-1}\mathbf{z}_{k-1} + \mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{z}_k)
\end{aligned} \tag{B.11}$$

Combination of Equations (B.10) and (B.11) yields the desired result after some manipulation.

$$\begin{aligned}
\hat{\mathbf{x}}_k &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k-1}(\mathbf{H}'_{k-1}\mathbf{R}_{k-1}^{-1}\mathbf{z}_{k-1} + \mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{z}_k) \\
&= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\mathbf{x}}_{k-1} + (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k-1}\mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{z}_k \\
&= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\mathbf{x}}_{k-1} + \mathbf{P}_{k-1}\mathbf{H}'_k\mathbf{R}_k^{-1}\mathbf{z}_k \\
&\quad -\mathbf{P}_{k-1}\mathbf{H}'_k(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}'_k + \mathbf{R}_k)^{-1}(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}'_k + \mathbf{R}_k)\mathbf{R}_k^{-1}\mathbf{z}_k \\
&\quad +\mathbf{P}_{k-1}\mathbf{H}'_k(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}'_k + \mathbf{R}_k)^{-1}\mathbf{z}_k \\
&= \hat{\mathbf{x}}_{k-1} + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k-1})
\end{aligned} \tag{B.12}$$

[2]See for example Householder[24].

B.3 The Kalman Filter

We now show that the Kalman Filter is a direct consequence of the relationships in Sections B.1 and B.2, when applied to first order discrete linear systems.

Consider an n-dimensional non-constant state vector, taking the value $\mathbf{x}_i$ at time t_i and obeying the equation (i=1,2,...,N):

$$\mathbf{x}_i = \mathbf{\Phi}_{i-1}\mathbf{x}_{i-1} + \mathbf{w}_{i-1} \tag{B.13}$$

where $\mathbf{w}_{i-1}$ is a white noise random vector sequence with zero mean and positive definite covariance matrix $\mathbf{Q}_{i-1}$. At each time t_i, the state vector $\mathbf{x}_i$ is observed through the measurement equation:

$$\mathbf{z}_i = \mathbf{H}_i\mathbf{x}_i + \mathbf{v}_i \tag{B.14}$$

with the same definitions for $\mathbf{z}_i$, $\mathbf{H}_i$, and $\mathbf{v}_i$ as in Section B.2. It is further assumed that the two white noise random sequences $\mathbf{v}_i$ and $\mathbf{w}_{i-1}$ are uncorrelated.

Then, the problem may be formulated as follows. Find from the measurements $(\mathbf{z}_i, \mathbf{z}_{i-1}, ..., \mathbf{z}_1)$ the minimum variance unbiased estimate $\hat{\mathbf{x}}_i$ (covariance matrix $\mathbf{P}_i$) of the state $\mathbf{x}_i$, depending linearly on the measurements, assuming the minimum variance unbiased estimate $\hat{\mathbf{x}}_{i-1}$ (covariance matrix $\mathbf{P}_{i-1}$) to be known from the measurements $(\mathbf{z}_{i-1}, \mathbf{z}_{i-2}, ..., \mathbf{z}_1)$.

In order to apply the results of the preceding sections to the problem at hand, let us introduce the vectors $\mathbf{x}^i$,$\mathbf{z}^i$,$\mathbf{w}^{i-1}$,$\mathbf{v}^i$ defined by the following recurrence relations:

$$\mathbf{x}^i = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}^{i-1} \end{bmatrix}, \ \mathbf{z}^i = \begin{bmatrix} \mathbf{z}_i \\ \mathbf{z}^{i-1} \end{bmatrix}, \ \mathbf{w}^{i-1} = \begin{bmatrix} \mathbf{w}_{i-1} \\ \mathbf{w}^{i-2} \end{bmatrix} \text{ and } \mathbf{v}^i = \begin{bmatrix} \mathbf{v}_i \\ \mathbf{v}^{i-1} \end{bmatrix}$$

If we similarly define the variance-covariance matrices $\mathbf{Q}^{i-1}$ and $\mathbf{R}^i$ as well as matrices $\mathbf{H}^i$, $\mathbf{\Phi}^i$ and $\mathbf{F}^i$ given by[3]

$$\mathbf{H}^i = \begin{bmatrix} \mathbf{H}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{i-1} \end{bmatrix}, \ \mathbf{\Phi}^i = \begin{bmatrix} \mathbf{I} \\ \mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1} \end{bmatrix} \text{ and } \mathbf{F}^i = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1} & \mathbf{F}^{i-1} \end{bmatrix}$$

[3]The matrix $\mathbf{\Phi}_{i-1}$ is assumed to have an inverse $\mathbf{\Phi}_{i-1}^{-1}$.

the whole set of equations (B.13) and (B.14) for i running from 1 up to i can be globally written as

$$\mathbf{z}^i = \mathbf{H}^i\mathbf{\Phi}^i\mathbf{x}_i - \mathbf{H}^i\mathbf{F}^i\mathbf{w}^{i-1} + \mathbf{v}^i \tag{B.15}$$

or equivalently,

$$\mathbf{z}^i = \mathbf{H}^i\mathbf{\Phi}^i\mathbf{x}_i + \mathbf{e}_i \tag{B.16}$$

where

$$\mathbf{e}_i = -\mathbf{H}^i\mathbf{F}^i\mathbf{w}^{i-1} + \mathbf{v}^i \tag{B.17}$$

with zero mean and covariance matrix $\mathbf{C}_i$ given by:

$$\mathbf{C}_i = \mathbf{H}^i\mathbf{F}^i\mathbf{Q}^{i-1}(\mathbf{H}^i\mathbf{F}^i)' + \mathbf{R}^i \tag{B.18}$$

Since Equation (B.16) has the form of Equation (B.1), the estimate $\hat{\mathbf{x}}_{i-1}$ may be written as

$$\hat{\mathbf{x}}_{i-1} = \mathbf{P}_{i-1}(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1})'\mathbf{C}_{i-1}^{-1}\mathbf{z}^{i-1} \tag{B.19}$$

$$\mathbf{P}_{i-1} = [(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1})'\mathbf{C}_{i-1}^{-1}(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1})]^{-1} \tag{B.20}$$

in view of Equations (B.2) and (B.3).

Before computing $\hat{\mathbf{x}}_i$, let us compute the minimum variance unbiased estimate $\hat{\mathbf{x}}_i^*$ (covariance matrix $\hat{\mathbf{P}}_i^*$) of the state $\mathbf{x}_i$ based upon the measurements $(\mathbf{z}_{i-1}, \mathbf{z}_{i-2}, ..., \mathbf{z}_1)$ only.

Introducing Equation (B.13) in the generalized measurement equation

$$\mathbf{z}^{i-1} = \mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{x}_{i-1} + \mathbf{e}_{i-1} \tag{B.21}$$

we obtain

$$\mathbf{z}^{i-1} = \mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1}\mathbf{x}_i + \mathbf{e}_i^* \tag{B.22}$$

where $\mathbf{e}_i^*$ is a new random vector variable:

$$\mathbf{e}_i^* = \mathbf{e}_{i-1} - \mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1}\mathbf{w}_{i-1} \tag{B.23}$$

with zero mean and covariance matrix $\mathbf{C}_i^*$:

$$\mathbf{C}_i^* = \mathbf{C}_{i-1} + \mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1}\mathbf{Q}_{i-1}(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1})' \tag{B.24}$$

Again, a Frobenius form is recognized in the above equation, therefore,

$$(\mathbf{C}_i^*)^{-1} = \mathbf{C}_{i-1}^{-1} - \mathbf{C}_{i-1}^{-1}\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1}[(\mathbf{\Phi}_{i-1}^{-1})'\mathbf{P}_{i-1}^{-1}\mathbf{\Phi}_{i-1}^{-1} + \mathbf{Q}_{i-1}^{-1}]^{-1}(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1})'\mathbf{C}_{i-1}^{-1} \tag{B.25}$$

so that the covariance matrix $\mathbf{P}_i^*$ which may be written in view of Equation (B.3),

$$(\mathbf{P}_i^*)^{-1} = (\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1})'(\mathbf{C}_i^*)^{-1}(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1}) \tag{B.26}$$

reduces to

$$\begin{aligned}(\mathbf{P}_i^*)^{-1} &= (\mathbf{\Phi}_{i-1}^{-1})'\mathbf{P}_{i-1}^{-1}\mathbf{\Phi}_{i-1}^{-1} \\ &\quad -(\mathbf{\Phi}_{i-1}^{-1})'\mathbf{P}_{i-1}^{-1}\mathbf{\Phi}_{i-1}^{-1}[(\mathbf{\Phi}_{i-1}^{-1})'\mathbf{P}_{i-1}^{-1}\mathbf{\Phi}_{i-1}^{-1} + \mathbf{Q}_{i-1}^{-1}]^{-1}(\mathbf{\Phi}_{i-1}^{-1})'\mathbf{P}_{i-1}^{-1}\mathbf{\Phi}_{i-1}^{-1} \\ &= [\mathbf{\Phi}_{i-1}\mathbf{P}_{i-1}\mathbf{\Phi}_{i-1}' + \mathbf{Q}_{i-1}]^{-1}\end{aligned} \tag{B.27}$$

as can be easily verified using Equations (B.20) and (B.25). Thus, $\mathbf{P}_i^*$ is given by

$$\mathbf{P}_i^* = \mathbf{\Phi}_{i-1}\mathbf{P}_{i-1}\mathbf{\Phi}_{i-1}' + \mathbf{Q}_{i-1} \tag{B.28}$$

A similar manipulation of $\hat{\mathbf{x}}_i^*$, which by Equation (B.2) is defined to be

$$\hat{\mathbf{x}}_i^* = \mathbf{P}_i^*(\mathbf{H}^{i-1}\mathbf{\Phi}^{i-1}\mathbf{\Phi}_{i-1}^{-1})'(\mathbf{C}_i^*)^{-1}\mathbf{z}^{i-1} \tag{B.29}$$

leads via Equations (B.19), (B.20), (B.25) and (B.28) to the result

$$\hat{\mathbf{x}}_i^* = \mathbf{\Phi}_{i-1}\hat{\mathbf{x}}_{i-1} \tag{B.30}$$

We can now derive the Kalman Filter equations in a straightforward manner since we can directly apply results of Section B.2 with

$$\mathbf{z}_k = \mathbf{z}_i,\ \mathbf{P}_{k-1} = \mathbf{P}_i^*,\ \hat{\mathbf{x}}_{k-1} = \hat{\mathbf{x}}_i^*,\ \mathbf{H}_k = \mathbf{H}_i,\ \mathbf{R}_k = \mathbf{R}_i$$

so that the estimate $\hat{\mathbf{x}}_i$ is immediately given by Equations (B.6), (B.7) and (B.8):

$$\hat{\mathbf{x}}_i = \hat{\mathbf{x}}_i^* + \mathbf{K}_i(\mathbf{z}_i - \mathbf{H}_i\hat{\mathbf{x}}_i^*) \tag{B.31}$$

$$\mathbf{K}_i = \mathbf{P}_i^*\mathbf{H}_i'(\mathbf{H}_i\mathbf{P}_i^*\mathbf{H}_i' + \mathbf{R}_i)^{-1} \tag{B.32}$$

$$\mathbf{P}_i = (\mathbf{I} - \mathbf{K}_i\mathbf{H}_i)\mathbf{P}_i^* \tag{B.33}$$

Equations (B.28), (B.30), (B.31), (B.32) and (B.33) constitute the equations comprising the solution to the discrete time linear Kalman Filter.

Bibliography

[1] K. Ashok. Dynamic Trip Table Estimation for Real Time Traffic Management Systems. S.M. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[2] K. Ashok and M. Ben-Akiva. Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems. In C.F. Daganzo, editor, *International Symposium on Transportation and Traffic Theory*, pages 465–484. Elsevier Science Publishing Company, Inc., 1993.

[3] M.G.H. Bell. The real time estimation of origin-destination flows in the presence of platoon dispersion. *Transportation Research*, 25B:115–125, 1991.

[4] M. Ben-Akiva, K. Ashok, and Q. Yang. Commentary on "A statistical analysis of the reliability of using RGS vehicles as estimators of dynamic departure rates" by B. Hellinga and M. Van Aerde. *IVHS Journal*, 2, 1994.

[5] M. Ben-Akiva and D. Bolduc. Aproaches to Model Transferability: Combined and Mixed Estimators. Unpublished Paper, Massachusetts Institute of Technology, Cambridge, 1985.

[6] M. Ben-Akiva and D. Bolduc. Aproaches to Model Transferability and Updating: The Combined Transfer Estimator. *Transportation Research Record*, 1139:1–7, 1987.

[7] M. Ben-Akiva, E. Cascetta, H. Gunn, S. Smulders, and J. Whittaker. DYNA: A Real-Time Monitoring and Prediction System for Inter-Urban Motorways. In

Proceedings of the First World Congress on Applications of Transport Telematics and Intelligent Vehicle-Highway Systems, December 1994.

[8] M. Ben-Akiva, A. de Palma, and I. Kaysi. The Impact of Predictive Information on Guidance Efficiency: An Analytical Approach. In *Proceedings of the TRISTAN Conference*, May 1994.

[9] Moshe Ben-Akiva. Methods to combine different data sources and estimate origin-destination matrices. In Nathan H. Gartner and Nigel H.M. Wilson, editors, *International Symposium on Transportation and Traffic Theory*, pages 459–481. Elsevier Science Publishing Company, Inc., 1987.

[10] Michel Bierlaire. Mathematical Models for Transportation Demand Analysis. Ph. D. Thesis, Department of Mathematics, FUNDP, Namur, Belgium,, 1996.

[11] G.J. Bierman. *Factorization Methods for Discrete Sequential Estimation.* Academic Press, New York, 1977.

[12] James K. Campbell, Stephen P. Synott, and Gerald J. Bierman. Voyager Orbit Determination at Jupiter. *IEEE Transactions on Automatic Control*, AC-28:256–268, March 1983.

[13] Ennio Cascetta, Domenico Inaudi, and Gerald Marquis. Dynamic Estimators of Origin-Destination Matrices using Traffic Counts. *Transportation Science*, 27(4):363–373, 1993.

[14] Ennio Cascetta, Agostino Nuzzolo, Francesco Russo, and Antonino Vitetta. A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and Some Calibration Results for Interurban Networks. Unpublished Paper, 1996.

[15] Gang-Len Chang and Xianding Tao. Estimation of Dynamic O-D Distributions for Urban Networks. In *Proceedings of 13th International Symposium on Transportation and Traffic Theory*, July 1996.

[16] G.L. Chang and J. Wu. Recursive estimation time-varying O-D flows from traffic counts in freeway corridors. *Transportation Research*, 28B, 1994.

[17] C.K.Chui and G.Chen. *Kalman Filtering with Real-Time Applications*. Springer-Verlag, 1991.

[18] M. Cremer and H. Keller. A new class of Dynamic methods for the identification of Origin-Destination Flows. *Transportation Research*, 21B(2):117–132, 1987.

[19] Nathan H. Gartner. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. *Transportation Research Record*, 906.

[20] Arthur Gelb, editor. *Applied Optimal Estimation*. M.I.T. Press, 1974.

[21] Y. Genin. Gaussian Estimates and Kalman Filtering. *Theory and Applications of Kalman Filtering*. Ed. C.T. Leondes NATO AGARDograph No. 139.

[22] William H. Greene. *Econometric Analysis*. Maxwell Macmillan International Publishing Group, 1993.

[23] P.K. Houpt, M. Athans, D.G. Orlhac, and W.J. Mitchell. Traffic Surveillance Data Processing in Urban Freeway Corridors using Kalman Filter techniques: Final Report. DOT-TSC-RSPA-78-18, November 1978.

[24] A.S. Householder. *The Theory of Matrices in Numerical Analysis*. Blaisdell, New York, 1964.

[25] D. Inaudi. Personal communication. 1992.

[26] D. Inaudi, E. Kroes, S. Manfredi, and S. Toffolo. The DYNA on-line Origin-Destination Estimation and Prediction Model. In *First World Congress on Applications of Transport Telematics and Intelligent Vehicle-Highway Systems*, December 1994.

[27] Pushkin Kachroo, Arvind Narayanan, and Kaan Ozbay. Investigating the Use of Kalman Filtering Approaches for Dynamic Origin-Destination Trip Table Estimation. Center for Transportation Research, Virginia Tech, Blacksburg, 1995.

[28] R.E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering(ASME)*, 82D:35–45, Mar 1960.

[29] I. Kaysi. Framework and Models for the provision of Real-Time Driver Information. Ph. D. Dissertation, Department of Civil Engineering, Massachusetts Institute of Technology, 1992.

[30] Hartmut Keller and Gerhard Ploss. Real-Time Identification of O-D Network Flows from Counts for Urban-Traffic Control. In Nathan H. Gartner and Nigel H.M. Wilson, editors, *International Symposium on Transportation and Traffic Theory*, pages 267–284. Elsevier Science Publishing Company, Inc., 1987.

[31] M. Khoshyaran. Dynamic O-D Estimation for a large network. Michigan State University Department of Civil Engineering, 1995.

[32] Brice G. Leibundgut, Andre Rault, and Francoise Gendreau. Application of Kalman Filtering to Demographic Models. *IEEE Transactions on Automatic Control*, AC-28:427–434, March 1983.

[33] Nancy L. Nihan and Gary A. Davis. Recursive estimation of Origin-Destination matrices from input/output counts. *Transportation Research*, 21B(2):149–163, 1987.

[34] Nancy L. Nihan and Mohammad M. Hamed. A fixed point approach to estimating freeway Origin-Destination Matrices and the effect of erroneous data on precision. University of Washington, Seattle, December 1990.

[35] I. Okutani and Y. Stephanades. Dynamic Prediction of Traffic Volume through Kalman Filtering Theory. *Transportation Research*, 18B:1–11, 1984.

[36] Markos Papageorgiou. *Application of automatic control concepts to traffic flow modeling and control.* Springer-Verlag, New York, 1983.

[37] MIT Intelligent Transportation Systems Program. A Proposal for Development of a Deployable DTA. Massachusetts Institute of Technology, January 1995.

[38] Herbert E. Rauch. Linear Smoothing Techniques. *Theory and Applications of Kalman Filtering.* Ed. C.T. Leondes NATO AGARDograph No. 139.

[39] Herbert E. Rauch, F. Tung, and C.T. Striebel. Maximum Likelihood Estimates of Linear Dynamic Systems. *AIAA Journal*, 3(8):1445–1450, August 1965.

[40] B.L. Smith. Short-term traffic flow prediction – a neural network approach. March 1994. Transportation Research Board.

[41] Harold W. Sorenson, editor. *Kalman Filtering: Theory and Application.* The Institute of Electrical and Electronics, Inc., New York, 1974.

[42] J. Sussman. Intelligent Vehicle Highway Systems. *OR/MS Today*, December 1992.

[43] Henri Theil. *Principles of Econometrics.* Wiley, New York, 1971.

[44] Nanne van der Zijpp. Dynamic Origin-Destination Matrix Estimation on Motorway Networks. Ph. D. Thesis, Department of Civil Engineering, Delft University of Technology, 1996.

[45] J.A.C. van Toorenburg and R.J.P. van der Linden. Predictive control in traffic-management. Technical Report, Rotterdam, Netherlands, March 1996.

[46] P.C. Vythoulkas. Alternative Approaches to Short Term Traffic Forecasting for Use in Driver Information Systems. In C.F. Daganzo, editor, *International Symposium on Transportation and Traffic Theory*, pages 485–506. Elsevier Science Publishing Company, Inc., 1993.

[47] J. Whittaker. A Kalman Filter for Network Travel Time Prediction. Unpublished Paper, Lancaster University, July 1991.

[48] Qi Yang. A simulation laboratory for evaluation of dynamic traffic management systems. Forthcoming Ph. D. Dissertation, Department of Civil Engineering, Massachusetts Institute of Technology, 1996.

[49] Qi Yang and Haris N. Koutsopoulos. A Microscopic Traffic Simulator for Evaluation of Dynamic Traffic Management Systems. *Transportation Research C*, 1996. Forthcoming.