November 2007

Final Report: ITS

Dynamic Estimation of OD Matrices for Freeways and Arterials

Authors:

Juan Carlos Herrera, Saurabh Amin, Alexandre Bayen, Samer Madanat, Michael Zhang, Yu Nie, Zhen Qian, Yingyan Lou, Yafeng Yin and Meng Li

PREPARED FOR:

California Department of Transportation
Sacramento

PREPARED BY:

Institute of Transportation Studies
UC Berkeley

Final Report: ITS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Origin-Destination (O-D) matrices provide information on flows of vehicles traveling from one specific geographical area to another, and are one of the critical data inputs to transportation planning, design and operations. Because it is very time consuming and labor intensive to obtain them through household interviews or roadside surveys, significant efforts have been made to develop mathematical models for estimating the matrices from link counts, which are relatively easier to obtain. So far, up-to-date commercial planning tools and simulation software have provided built-in O-D estimation modules. However, most of these O-D estimators are only capable of estimating *static* O-D matrices rather than *dynamic* or *time-dependent* O-D matrices. The latter are pre-requisites for short-term planning applications and traffic operations studies.

The goal of this study was to bridge the gaps between practice and theory in O-D estimation. In particular, this work planned to develop the methodologies for deriving time-dependent O-D matrices for linear networks and implement them in a computer tool (a linear network is a stretch of highway with multiple entries and exits, where there would be no route choices involved).

The study was divided into two parts: O-D table estimation on freeways and O-D table estimation on arterials. To achieve the goal, the following tasks were performed on each part:

O-D estimation on freeways:

1. A review and comparison for determining appropriate techniques and methods from the literature was performed. A methodology based on Kalman filtering techniques was chosen, where the state vector can be either the O-D flows for each O-D pair or its deviation from a historical value.
2. A methodology that identifies when traffic conditions vary and makes use of existing models to estimate OD flows in a linear network was implemented. The models were tested using real data collected from two different networks (I-80W, CA, and I-90E, MA) during free flow and congested conditions. An algorithm to detect traffic condition changes, which makes use of the speed measurements provided by loop detectors, was proposed.
3. The development of a computer O-D estimator tool with a user-friendly interface has been started. A preliminary version of this tool is provided.
4. A new O-D table estimation approach, based on variational inequalities, was proposed. This approach takes into account various levels of traffic information, such as link flow count, historical O-D tables, static planning O-D and observed path travel times. A portion of freeway SR-41 in Fresno, CA (16.7 miles) was used to test the approach. True O-D tables are not known, but synthetic time-dependent O-D tables were produced. Different demand patterns were tested and investigated.

O-D estimation on arterials:

1. A two-step approach for estimating time-varying O-D flows for actuation-controlled corridors with incomplete information about entering and exiting flows is proposed. At the first step, turning movements for each intersection are estimated, and then used at the second step to construct the measurement

equations to infer the corridor O-D flows. The proposed approach has been demonstrated and validated on a test-case corridor. The O-D estimator has subsequently been applied to a segment of El Camino Real, San Mateo, CA with three intersections: 28th , 27th and 25th Avenue.

2. An empirical analysis of the O-D demand structures and examination of their impacts on the O-D estimation (at a single intersection level) have been conducted. Data from the 34th Street-University Avenue intersection in Gainesville, Florida in 2001, was used for this purpose.

Software development:

1. The objective of this part of the project was to implement the models and algorithms developed into a computer tool to allow practitioners to apply the proposed models and algorithms. The work realized on variational inequalities was not implemented, since the corresponding results will need further developments before they become applicable in the form of a tool. All other results were implemented. The software currently compute the O-D flows for 30 seconds intervals using the Berkeley Highway Lab data used to test the freeway algorithm.
2. The current version of the software is attached to this report (CD).

The fundamental conclusions from this research are the following:

O-D estimation on freeways:

1. Models using the deviation of the O-D flow from an historical value (instead of the O-D flow directly) yield more accurate O-D estimates. The estimations match the actual data well.
2. Accurate travel time estimates are not essential for the accuracy of the proposed method. Preliminary evidence that the accuracy of travel time estimates does not impact the results greatly, and a distinction between free flow and congested conditions is enough for the purposes of this work was shown.
3. For the variational inequality-based approach, computational experiments showed that traffic counts are indispensable for O-D estimation, and that the quality of the results increases with the number of measurements. If the number of counting locations is small, the location of the sensors might be crucial for the estimation. Traffic counts alone are not sufficient to obtain accurate O-D demand estimates. Historical time-dependent O-D demands can drastically improve the quality of the estimates. Static O-D tables can also be used in the estimation process. Finally, counter to our expectations, path travel times do not contribute significantly to improving O-D demand estimates.

O-D estimation on arterials:

1. For a single intersection, the proposed two-step approach outperforms the conventional generalized least squares approach. Moreover, the former is more efficient as well. As in the freeway case, both approaches would benefit from accurate prior knowledge or partial O-D information.
2. At the corridor level, the estimator is able to track the trend of time-varying O-D flows and produce estimates relatively close to the actual values in an average sense. The estimator is less sensitive to the changes of the O-D flows and thus estimates are generally not as fluctuating as the true values. The reason is that the proposed approach at the corridor level only makes use of the localized

information. By doing so, the problem can be decomposed to simplify the formulation and improve largely the computational efficiency.

3. Since real O-D observations are not available from El Camino Real, the accuracy of the estimates can not be verified. However, the application does demonstrate that the estimator is able to readily work with actual field loop data.
4. The comparison of estimators with different state variables suggested that the estimator with state variable of O-D flow outperforms the others in the particular case investigated. We fully recognize that O-D patterns would be site-dependent, and the results of this case study should not be generalized.
5. Demands or flows at different O-D pairs may possess different structures, which are very often not first-order auto regressive. Incorporating all of these "true" structures into the Kalman filtering algorithm makes the model formulation very complicated. On the other hand, the simple first-order auto regressive assumption produces acceptable results in our empirical experiments and previous studies. Therefore, unless there are sufficient O-D data that suggest otherwise, one might simply use the state variable of O-D flows or splits and assume that they are first-order auto regressive.

Our recommendations for future work are as follows:

1. Alternative sources of measurements. Can measurements from probe vehicles (already available, for example FasTrak transponder data) be incorporated in our estimation algorithms? Partial O-D information is now directly available from FasTrak readers at a high penetration rate in the Bay Area and could increase the accuracy of O-D estimations.
2. Sensor placement algorithms. For limited numbers of sensors to be deployed, where should the sensors be deployed to provide maximal estimation accuracy?
3. GPS-based measurements. Investigation of the possibility of integrating market driven information such as cellular phone data into the estimation algorithms. The applicability of this new type of GPS-based information has implications which goes beyond O-D estimation, in particular for:
    a. Travel time estimation for changeable message signs
    b. Congestion estimation for ramp metering
    c. High quality information for HOV lanes

    The use of GPS based cellular phone measurements could:
    a. Improve the quality of estimations for the above quantities,
    b. Provide Caltrans with a richer database, additional to PeMS and FasTrak,
    c. Provide Caltrans with measurements where infrastructure is not available,
    d. Progressively become an alternative to costly loop detector deployment or maintenance.

    This recommendation is in our view the most important, as it provides Caltrans with a cost efficient long term alternative to loop detectors. The progressive penetration of the cellular phone market by GPS equipped devices is driven by competition between cellular phone companies. Major cellular phone companies are already developing their own data retrieval infrastructure, from which Caltrans could potentially benefit everywhere (not only in urban or suburban areas).

CHAPTER 1

A DYNAMIC ORIGIN-DESTINATION MATRIX ESTIMATION ALGORITHM FOR FREEWAYS

Prepared by:

Juan Carlos Herrera, Saurabh Amin, Alexandre Bayen, and Samer Madanat

Institute of Transportation Studies

University of California, Berkeley

1.1 Introduction

Origin-Destination (OD) matrices[1] are needed for short-term planning applications and for operational studies. In particular, they can be used to develop some control strategies of the network under study in order to improve its performance. For instance, a better traffic signal coordination may be achieved on an urban corridor, an adaptive ramp metering strategy can be developed on a freeway, or they can be used to provide drivers with better information.

The aim of this project is to implement an algorithm to compute dynamic OD matrices on a linear network[2]. In particular, our focus is the freeway case.

The approach used here is based on the approaches developed in [1] and [2]. These approaches (which are quite similar) estimate OD flows assuming the network is in the same traffic condition. That is, there is no shockwave in the network. The approaches describe the system with a set of linear state-space equations. Some parameters of these equations will depend on the traffic conditions along the network (i.e. they will change if the network is under free-flow or congested conditions). For this reason, we first identify the traffic conditions, select the corresponding state-space equations and solve them until a change in the traffic conditions is detected.

If the traffic conditions are not the same along the whole network, the model will not work properly. The reason for this is that the system would not be linear, and thus, it would not be properly described by a set of linear state space equations. As a consequence, during this *transition* period[3] (which should not be very long) the model will probably not yield good estimates.

The rest of this document is organized as follows: approaches adopted to represent linear systems with state-space equations will be briefly described on Section 1.2. Section 1.3 describes the implementations of the approaches presented in Section 1.2 using real data and their results. These implementations assume the same traffic conditions all the time (i.e. there is no need to identify traffic conditions in order to select the appropriate state-space equations). Section 1.4 will address the identification of traffic conditions. A brief discussion regarding network size is given in Section 1.5. Finally, Section 1.6 presents the main conclusions of this work.

1.2 Methodology

This section presents the notation used and describes how the system is represented under the two approaches.

[1] Element ij of an OD matrix is the number of vehicles going from i to j during some time interval.

[2] A linear network is a network where there is no route choice for any OD pair. For instance, a highway with on- and off-ramps or an urban arterial can be modeled as a linear network.

[3] Period when a shockwave is in the network.

1.2.1 Notation

- N : Set of nodes in the network,
- L : Set of links in the network,
- n_L: Number of links in L that are equipped with counting stations (each station is denoted by $l \in L$),
- n_{OD}: Number of OD pairs,
- x_{rh}: Number of vehicles between the r-th OD pair that left the origin in time interval h,
- x_{rh}^H: historical estimate for x_{rh},
- Δx_{rh}: Deviation of x_{rh} from the corresponding historical estimate,
- x_h: $n_{OD} \times 1$ vector of all OD flows at time interval h,
- x_h^H: Associated historical OD flow vector,
- Δx_h: Vector of deviations $(= x_h - x_h^H)$,
- y_{lh}: Observed traffic counts at link station l at time interval h,
- y_h: $n_L \times 1$ vector of such counts,
- y_h^H: associated historical link count vector,
- Δy_h: Vector of deviations of all counts $(= y_h - y_h^H)$,
- f_{qh}: $n_{OD} \times n_{OD}$ matrix of coefficients describing the effects of flows at time interval q on flows at time interval h,
- w_h: $n_{OD} \times 1$ vector of random errors in the state transition equation (defined in the following subsection),
- Q: corresponding covariance matrix,
- p : maximum order of the autoregressive model in the state transition equation,
- a_{qh}: $n_L \times n_{OD}$ assignment matrix describing how OD flows x_q contribute to link flows y_h,
- v_h: $n_L \times 1$ vector of measurement errors in the measurement equation (defined in the following subsection),
- R: corresponding covariance matrix,
- u: maximum number of time intervals needed to travel between any OD pair.

1.2.2 System Representation

The system can be described by the following equations [1]:

$$x_{h+1} = \sum_{k=h-p+1}^{h} f_{kh} \cdot x_k + w_h \tag{1.1}$$

$$y_h = \sum_{k=h-u}^{h} a_{kh} \cdot x_k + v_h \tag{1.2}$$

The first equation relates current OD flow with the previous ones, and it tries to capture the correlation over time of the OD flows. This kind of correlation arises from unobserved phenomena such as changes in traffic demand, and/or changes in the transportation network. This equation will hold true as long as traffic conditions change gradually (smoothly).

The second equation relates OD flows with link counts through the assignment matrices a_{kh}. These matrices contain the information of how the traffic evolves over time and space along the network. The dimension of each assignment matrix is $n_L \times n_{OD}$, and there are as many matrices as the maximum number of intervals needed to go from any origin to any destination point in the network. Each element lr of the matrix a_{qh} represents the contribution of link count l at period h to OD flow r that entered the network at $h - q$.

Ashok and Ben-Akiva [2] found that Equation (1.1) can only capture temporal interdependencies among OD flows, and does not represent structural information about trip patterns. Instead of using x_h as the state variable, they decided to use its deviation from a historical OD flow. By doing so, the estimation and prediction process would have taken into account (indirectly) all the experience gained over previous estimations and it would be richer in its structural content. Also, deviations can be either positive or negative, so they can be approximated by a normal distribution, which is also a useful property for a tool such as Kalman Filter (which will be used later to solve the equations). The corresponding equations are given by:

$$\Delta x_{h+1} = \sum_{k=h-p+1}^{h} f_{kh} \cdot \Delta x_k + w_h \tag{1.3}$$

$$\Delta y_h = \sum_{k=h-u}^{h} a_{kh} \cdot \Delta x_k + v_h \tag{1.4}$$

(Note that $y_h^H = \sum_{k=h-u}^{h} a_{kh} \cdot \ x_k^H$)

There are two aspects to note about the previous equations:

1. Assumptions: The following assumptions are made about the error terms in previous equations:

   - $E[w_h] = 0$, for all h,

- $E[w_h w_m^T] = Q_h \delta_{hm}$ where δ_{hm} is the Kronecker's delta and Q_h is $n_{OD} \times n_{OD}$ covariance matrix at time interval h,
- $E[v_h] = 0$, for all h,
- $E[v_h v_m^T] = R_h \delta_{hm}$ where R_h is $n_L \times n_L$ covariance matrix,
- $E[w_h v_m^T] = 0$, for all h, m, that is, transition and measurement errors are uncorrelated.

When used in practical applications, relationships between OD flows (or deviations) across different OD pairs are ignored.

2. Parameter estimation: The matrix f_{rh} can be estimated from historical data by estimating linear regression models for each OD pair and Q can be approximated from the residuals of these regressions. For Equation (1.3) and (1.4), vectors x_h^H for all h are obtained from a database of the OD matrix created off line (from previous estimations). The matrix R can be approximated from historical data. However, computation of the assignment matrix a_{ph} is a complicated exercise [3] as it is a nonlinear function of the route choice assumptions, network topology and travel time. We do not deal with the route choice problem because our network is linear (i.e. there is no route choice). When the travel times in the network are unobservable, the assignment matrix is endogenous to the model. In order to address the endogeneity of the assignment matrix, [2] and [1] use an iterative OD estimation and assignment matrix computation approach. In this approach, for the current OD estimate, a traffic simulation model is used to compute the assignment matrix which is then used to compute the new OD flow estimates by the filtering algorithm. As pointed in [3], such an approach does not guarantee convergence and can potentially lead to biased estimates. The authors of [3] also develop a rigorous approach to estimate the assignment matrix based on stochastic mapping between dynamic OD flows and link counts.

1.2.3 State augmentation: State-Space Equations

A discrete linear state-space system is often described as follows:

$$x_{h+1} = A \cdot x_h + B \cdot t_k \quad (1.5)$$

$$y_h = C \cdot x_h + D \cdot t_k \quad (1.6)$$

Equation (1.5) is referred to as the *system* or *transition* equation, while Equation (1.6) is known as the *output* or *measurement* equation. x_h is the state vector at time interval h, y_h is the output or measurement vector at time interval h, t_h is the control input at time interval h, and A, B, C, and D are matrices.

Equations (1.1)-(1.2), and Equations (1.3)-(1.4) suggest that state augmentation is needed to fully utilize all available information and to represent our system as a linear state space model. If $s = \max(u, p-1)$, then the augmented state vector should be $n_{OD}(s+1) \times 1$. We denote the $n_{OD}(s+1) \times 1$ augmented state and the historical state vectors by X_h and X_h^H , the $n_L \times n_{OD}(s+1)$ augmented assignment matrix by A_h, the $n_{OD}(s+1) \times n_{OD}(s+1)$ augmented (and appropriately modified) autoregressive parameter

matrix by F_h, the $n_{OD}(s+1) \times 1$ augmented error vector by W_h (with bottom n_{OD}'s elements as zeros) and the $n_{OD}(s+1) \times n_{OD}(s+1)$ covariance matrix by Q_h. In addition, we define following notation: $\mathcal{Y}_h := \Delta y_h$, $\mathcal{X}_h := X_h - X_h^H$, $\mathcal{B}_h := A_h \cdot X_h^H - y_h^H$. Following [2], the augmented state transition and measurement equations can be written as follows (an example of how state augmentation is done can be found in Appendix A):

$$\left.\begin{array}{rcl} \mathcal{X}_{h+1} & = & F_h \cdot \mathcal{X}_h + W_h \\ \mathcal{Y}_h & = & A_h \cdot \mathcal{X}_h + \mathcal{B}_h + v_h \end{array}\right\} \tag{1.7}$$

In Equation (1.7) both expressions are in state-space form. In the implementations that will be presented in Section 1.3 we assume that matrices F, W, A, and v are time invariant, and then the subindex can be omitted. Equation (1.7) can now be directly fed into a Kalman filtering algorithm [4] to give minimum least squares estimates of the state variable $\mathcal{X}_{h+1}$.

Depending on the state variable chosen (x_h or Δx_h), we would end up with two different state space representations (in the form of Equation (1.7)). Solution techniques such as Kalman filtering (used in [1], [2], and [3]) and recursive least-square approaches (such as [5]) are based on one of these state-space representations[4].

Note that Equation (1.7) imply that each OD flow will be estimated $s+1$ times during each time interval. However, when n_{OD} is large (e.g., in a large network) and/or when s is large (e.g., when the network is in congested regime), this procedure might become computationally intensive. To address this problem, [2] proposes an approximation scheme based on the assumption that much of the information about an OD flow is incorporated the first time it is counted.

Finally, one remark regarding the observability of the system should be made. Complete observability refers to the ability to uniquely determine the initial state vector from a given set of measurements. The factors affecting the observability of the system are: (i) the ratio n_L/n_{OD}, (ii) degree of linkage between OD flows and station counts or the rank of the assignment matrix and (iii) the degree of linkage between the OD flows over time or the rank of the transition matrix.

1.3 Implementation of the Algorithm

This section describes an actual implementation of the model described in Section 1.2.3 (Equation (1.7)) and its main results. We present two implementations performed during free-flow traffic conditions (BHL[5] and MT[6]), and one during congested conditions (BHL). Both the methodologies with and without deviations were tried in two of the implementations in order to compare performance.

For the implementation, the algorithm requires a historical OD matrix, vehicles

[4]In some parts of this document we will refer to the model derived from Equation (1.1) and (1.2) as the *methodology without deviations*, and to the model derived from Equation (1.3) and (1.4) as the *methodology with deviations*.

[5]Berkeley Highway Laboratory.

[6]Massachusetts Turnpike.

counts, and travel times between detector stations as inputs. The output is the estimated OD matrix, which –for validation purposes only– will be compared against the real OD matrix to assess the performance of the algorithm. Some indexes of performance can be also computed in order to compare alternative approaches. Figure 1.1 presents an outline of the method followed in this study.

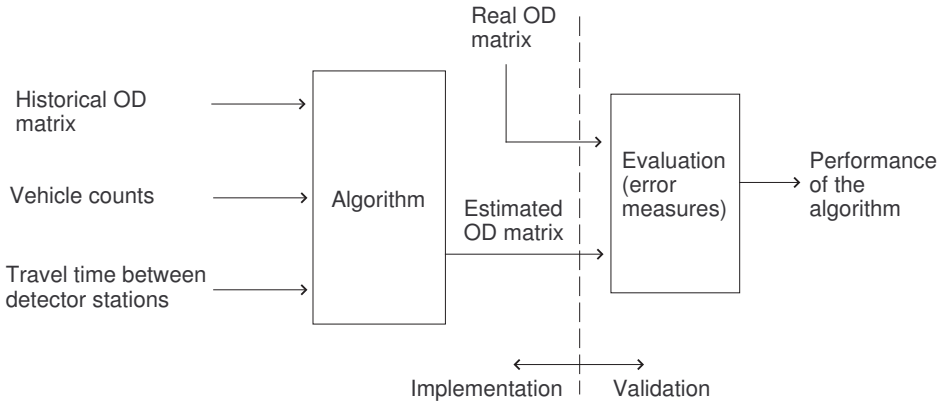


Figure 1.1: Implementation and validation of the algorithm.

The historical OD matrix will be used to obtain the transition matrix (by estimating a linear regression models for each OD pair, see Section 1.2.2). Depending on the methodology used, this historical matrix will be also used as part of the state variables (the methodology with deviations defines the state variable as the deviation of the estimated OD flow from the historical one). Travel times between detector stations are needed to compute the assignment matrices and vehicle counts corresponding to the output vector y_h. In all the cases, we are dealing with situations where the travel time is constant (i.e. same traffic conditions over the simulation period).

1.3.1 Implementations under free-flow conditions

The study with the first network (BHL) is implemented using the methodology without deviations, while both methodologies (with and without) deviations will be used on the second network (MT)[7]. For the second network, however, we will present here only the results from the methodology with deviations. Results from the methodology without deviations were only used to compare both methodologies (Section 1.3.3).

Interstate-80 Westbound (BHL)

Site description and data collection. The network chosen has two origins and two destinations (four OD pairs) and corresponds to highway I-80W between Ashby and Powell. Figure 1.2 depicts the geometry of the site.

The OD flows were numbered in the following way:
1-3 $\Rightarrow$ OD pair 1

[7]This is due to the lack of historical data for BHL (unlike the MT case).

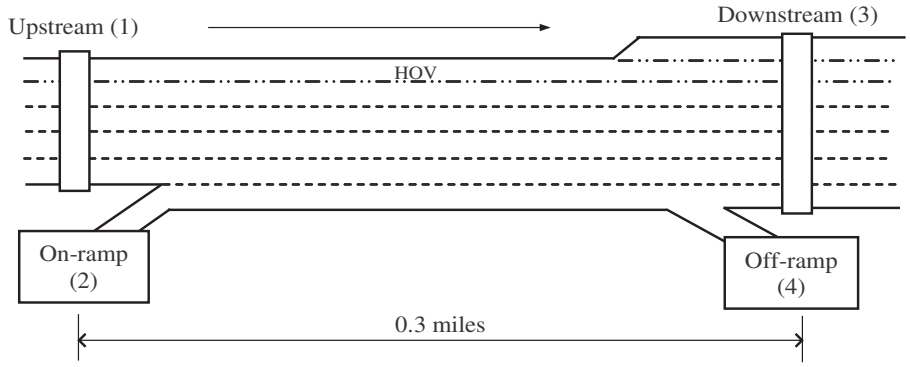


Figure 1.2: BHL section of I-80W used for the first part of this study.

1-4 $\Rightarrow$ OD pair 2
2-3 $\Rightarrow$ OD pair 3
2-4 $\Rightarrow$ OD pair 4

In the same way, link numbers correspond to the detector station numbers.

The data was collected using BHL's cameras installed on the roof of a building located next to the highway. All the data needed to run the algorithm were extracted from the video recordings. This process is time consuming.

Data were recorded during one hour under free flow conditions (10:00 to 11:00am) for two consecutive weekdays. The time interval chosen was 30 seconds (i.e. a new OD matrix estimation is done every 30 seconds). The first day of data was used to extract the "historical" OD matrix, and the second one was used to implement the algorithm. That is, for the second day –and in addition to the extraction of the real OD matrix– we had to count vehicles every 30 seconds at each entry and exit point of the network. A constant travel time of 18 seconds (from upstream/on-ramp to downstream/off-ramp) was used to compute the assignment matrices.

Transition equation. A fourth order autoregressive model (i.e. $p = 4$) for each of the OD pairs was used to compute the coefficients of each matrix f_{qh} (and then F_h). Other orders were tried, but $p = 4$ provides the best fit (i.e. lowest sum of squared residuals). After state augmentation (Section 1.2.3), the matrix F_h has dimensions 16×16 (u turns out to be 1, then $s = \max(u, p-1) = 3$). Random errors were computed as described in Section 1.2.2.

Measurement equation. Assignment matrices were computed as follows. Since the travel time is 18 seconds[8], each vehicle entering the network during the first 12 seconds of interval h will exit the network during the same interval h. All the vehicles entering the network in the last 18 seconds of interval h will exit the network in the next interval $h + 1$. Assuming a uniform distribution of the flow during 30 seconds (which is a fairly reasonable assumption), 0.4 (=12/30) of the vehicles entering in h leave the network

[8]The travel time was computed by inspecting the videos. Five randomly selected vehicles, every five minutes were used to compute the average travel time.

during h and the rest of the vehicles entering in h (0.6=18/30) leave the network in $h+1$. For this reason, there are two assignment matrices ($u = 1$).

$$A_h^h = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0.4 & 0 & 0.4 & 0 \\ 0 & 0.4 & 0 & 0.4 \end{pmatrix} \quad A_h^{h-1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0.6 & 0 \\ 0 & 0.6 & 0 & 0.6 \end{pmatrix} \tag{1.8}$$

Based on the amount of error that might have occurred while counting vehicles from video data, the value of the measurement errors (v_h) are 8, 4, 8, and 4 for y_{1h}, y_{2h}, y_{3h}, and y_{4h}, respectively. That is, for each observation every 30 seconds, we expect an error of the order of ± 8 vehicles in the count at a given station (including all lanes). Since the counts are approximately 80 vehicles every 30 seconds, this means an error of about 30%. The same reasoning is applied to the ramp counts, but assuming a larger error.

Results. Even though the data is available for one hour, to run the algorithm we used only the first 35 minutes (70 intervals) of data (mainly because the data-extraction-process is cumbersome). Figures 3 to 6 show the estimated (dotted line) and the real (solid line) OD flows for the first 35 minutes of the second day for every OD pair (please note that different figures are at different scales).

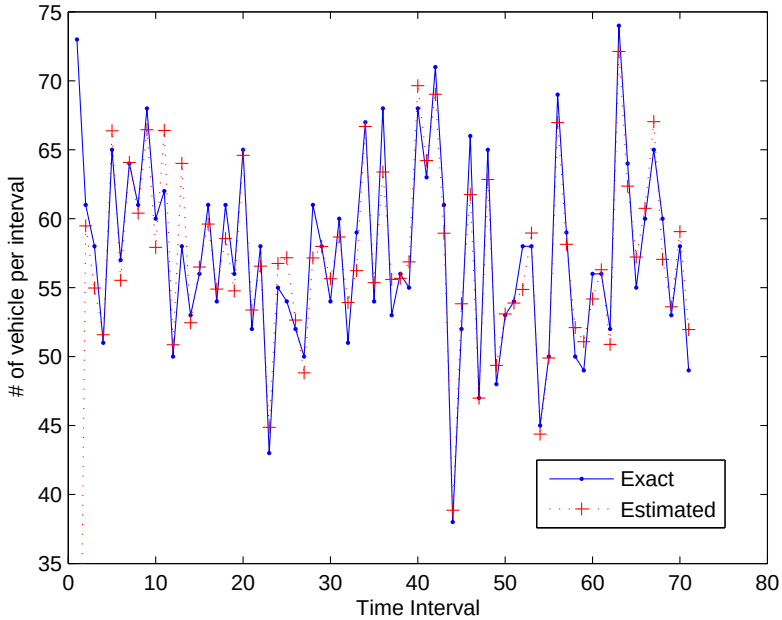


Figure 1.3: Exact vs. Estimated OD flow for OD pair 1.

Figure 1.5 shows a very good agreement between exact count and estimate. The fits are good for the other OD flows, but they are not as good as the case of the third OD pair.

It has to be noted that the real OD flow for OD pair 4 takes only three values (0, 1, and 2) (see Fig.1.6). In fact, most of the time, the flow is zero. For this reason, the

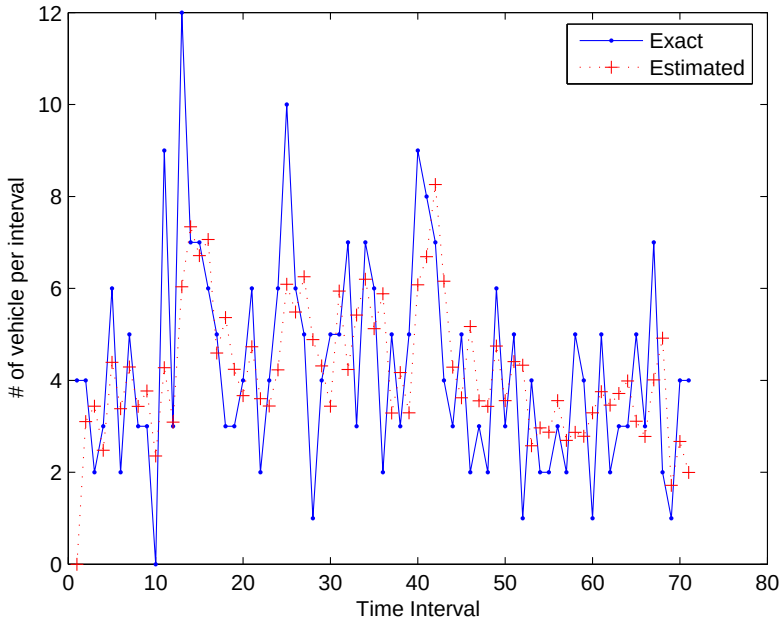


Figure 1.4: Exact vs. Estimated OD flow for OD pair 2.

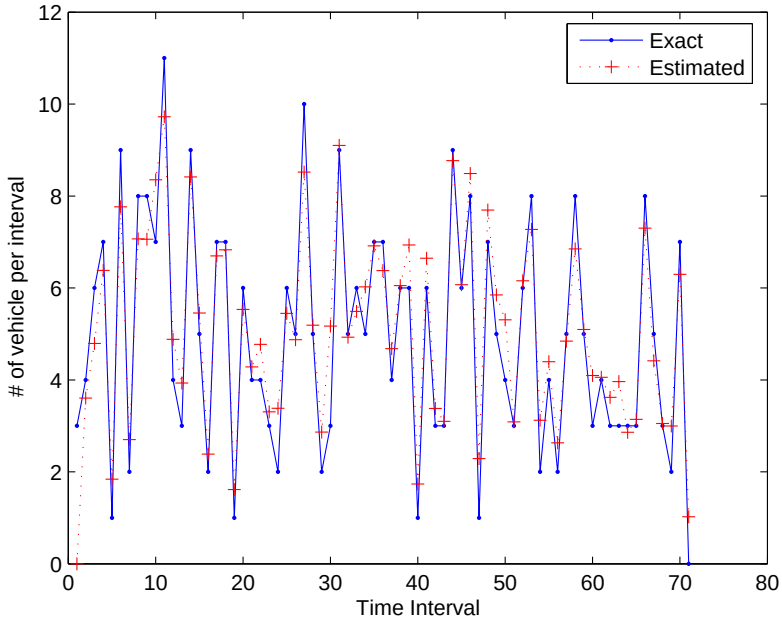


Figure 1.5: Exact vs. Estimated OD flow for OD pair 3.

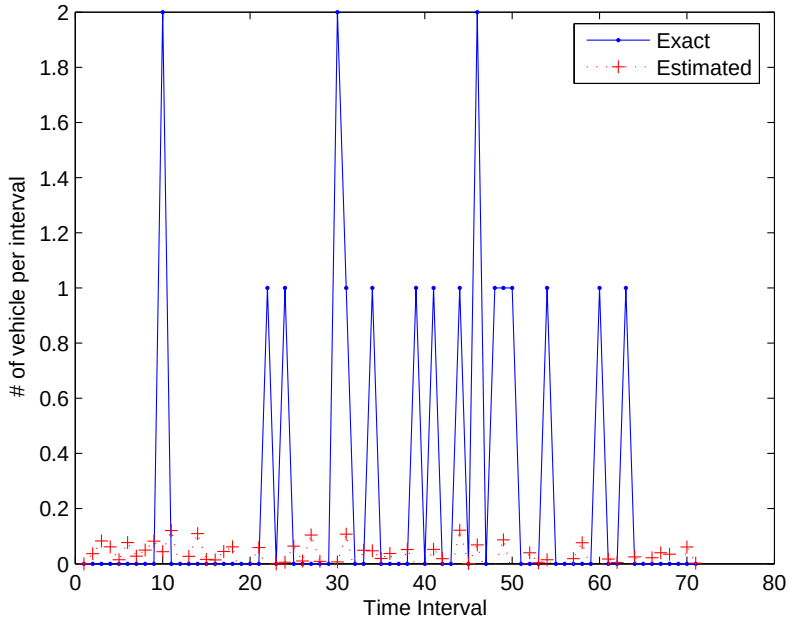


Figure 1.6: Exact vs. Estimated OD flow for OD pair 4.

estimated values for this OD pair are always less than 1. This fact might explain the very good agreement for the third OD pair.

Interstate-90 Eastbound (Massachusetts Turnpike)

Site description and data collection. The length of the network is 75.6 miles and it contains 10 origin/destination points (see Figure 1.7). The first one is only an origin point, and the last two are only destination points. The other 7 points in between are origin and destination points. That is, there are 8 entry points and 9 exit points, which yields 44 OD pairs (i.e. $n_L = 17$ [9] and $n_{OD} = 44$).

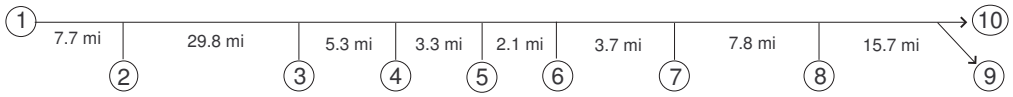


Figure 1.7: I-90E - Massachusetts Turnpike

The data consists of the OD flows every 15 minutes, during 6 days, for each OD pair. That is, the OD matrix for 6 days is known, but the counts at the entry and exit points are not. The data has to be manipulated in order to obtain vehicle counts at these locations. The count at any entry point can be easily computed from the original data, but assignment matrices are needed to estimate count at exit points. A constant speed of 55 mph was assumed to compute these matrices[10].

[9] We are assuming that there are no detectors along the freeway. We have assumed that there are detectors only at the entry and exit points.

[10] Since the section under study rarely get congested, this is a reasonable assumption.

For this network, both methodologies (with and without deviations) were tested. In this section, however, we will show only the results obtained with the methodology with deviations (the other implementation was performed for comparison purposes).

Transition Equation. The first day was used as the *historical* day, and the transition matrix was computed using data from the first two days. That is, for each OD pair an autoregressive model using the deviation of OD flows of the second day from those of the first one was performed in order to compute the transition matrix. The order of the autoregressive model is not the same for every OD pair, and it depends on the value which gives the best fit from an statistical point of view (i.e. least residuals). In our case, p (maximum order) is 4. As we will see later, $s = \max(u, p-1) = 6$, and then the augmented transition matrix is an square matrix of dimensions 308×308 ($n_{OD}(s+1) \times n_{OD}(s+1)$ (see Section 1.2.3).

The covariance matrix Q was computed using the residual of the autoregressive models (see Section 1.2.2), and it has the same dimensions as the augmented transition matrix.

Measurement Equation. Assignment matrices were computed assuming a constant speed of 55 mph in the whole section[11]. The travel time for the "longest" OD pair is about 82.5 minutes, which means that 5.5 intervals[12] is the maximum time that any trip will take in our network. Then, $u = 6$ and seven assignment matrices are needed ($s = \max(u, p-1) = 6$).

The dimensions of each one of these seven assignment matrices are 17×44 ($n_L \times n_{OD}$). Since $s = 6$ the augmented assignment matrix (see Section 1.2.3) has dimensions 17×308.

These assignment matrices were used to compute all the exit counts. Since some elements of the assignment matrices are not integer, we will obtain fractional counts. If we work with these fractional counts, then the covariance matrix R would be zero (because the difference between y_h^H and $\sum_{k=h-u}^{h} a_{kh} \cdot x_k^H$ would be zero). However, we have decided to round the count to the nearest integer[13]. The differences were then used to compute the 17×17 ($n_L \times n_L$) matrix R. Note that detectors on the entry points have no error.

Results. The third day was used to test how the algorithm works. The result of the implementation of the algorithm is a matrix that contains the *deviations* of the estimated OD flow from the historical one (OD flow for the first day) for every OD pair and for every time interval. Then, the estimated OD flow is just the sum of the deviation and the historical OD flow.

Figures 1.8 to 1.10 show the agreement between the exact and estimated curve for three OD pairs that have different OD flow levels. Figure 1.8 corresponds to OD pair number 9, which is the largest one (i.e. from the first entry to the last exit). Pair 17

[11] Distances between origin/destination points were also known.

[12] We are using 15 minutes intervals.

[13] Since loop detectors always give an integer number of vehicle count, this seems to be a reasonable thing to do.

(Figure 1.9) is from the second entry to the last exit, and pair 31 (Figure 1.10) goes from the fifth entry to first exit after it.

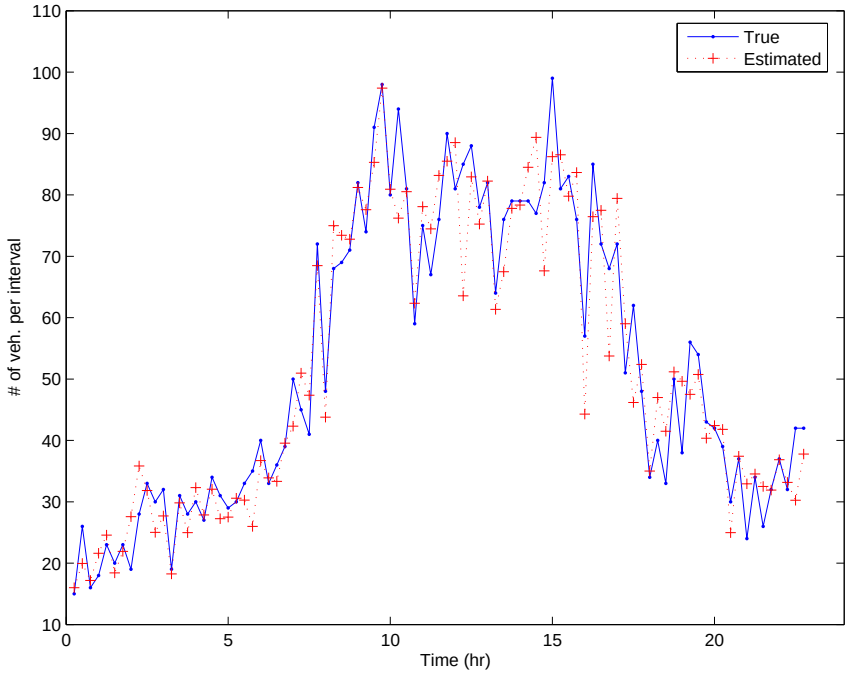


Figure 1.8: True vs. Estimated OD flow for OD pair 9.

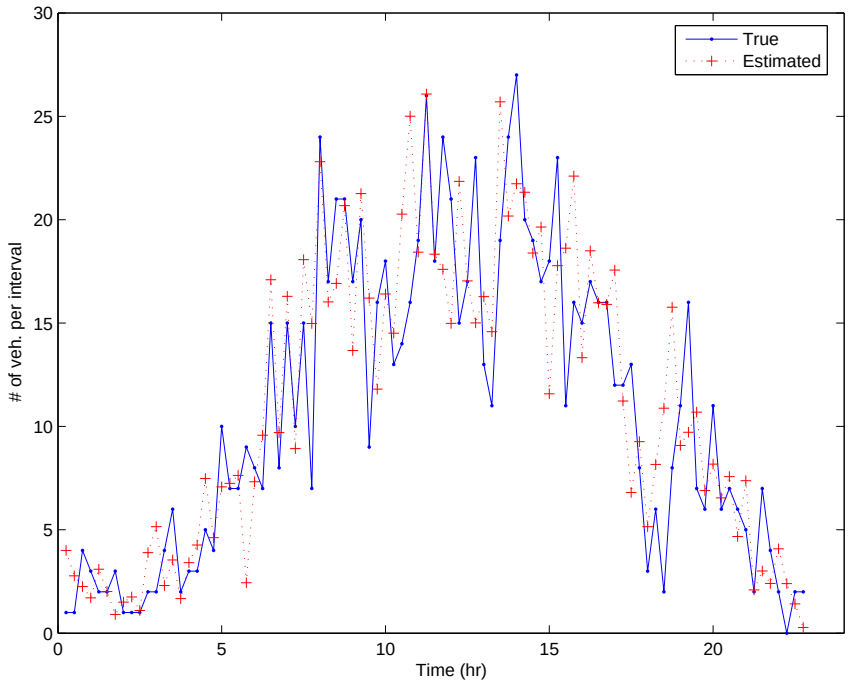


Figure 1.9: True vs. Estimated OD flow for OD pair 17.

It is important to note that the scale for the three graphs is not the same. If fact, Figure 1.11 shows both the true and estimated curves for OD pair 9 and 17 in the same graph.

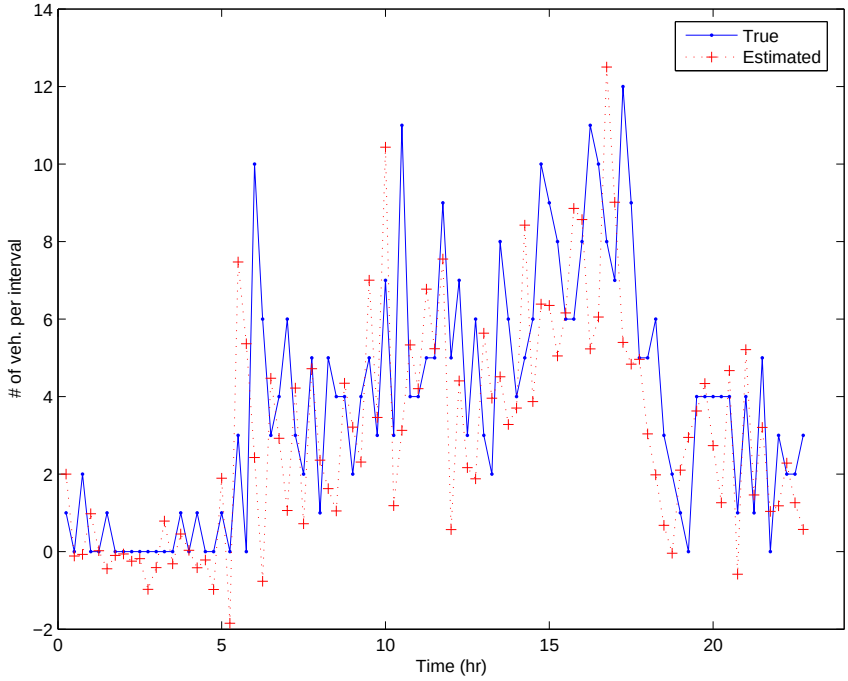


Figure 1.10: True vs. Estimated OD flow for OD pair 31.

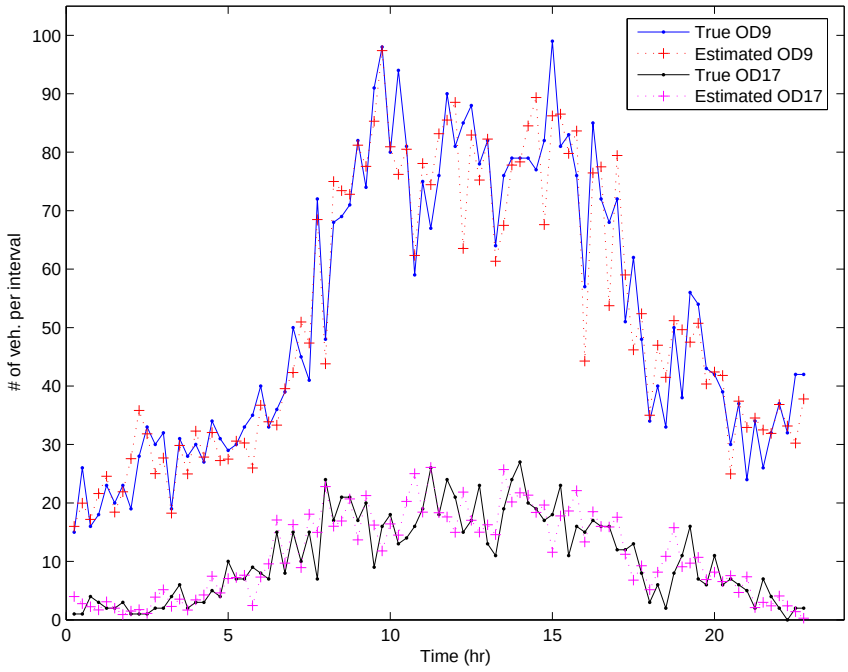


Figure 1.11: Comparison of two different levels of OD flows.

Unlike the BHL case, this new network has several OD pair, which makes the estimation process a little bit harder than before. The algorithm, however, still performs very well and the OD flows estimated follow the trend of the true OD flows.

1.3.2 Implementation under congested conditions

The BHL network (described in Section 1.3.1) was used to validate the algorithm under congested conditions. Three days of video data from 3-6pm were recorded. For our purposes, however, only 30 minutes were implemented, and time intervals of 30 seconds were used. As with the MT case, both methodologies with and without deviations were implemented on these data. We will present here, however, only the results of the methodology with the deviations (results using the other methodology were used to compare the two methodologies, see Section 1.3.3).

Again, the first day was used to extract the *historical* OD matrix, the second day was used to compute the transition matrix F (using the deviation of this OD matrix from the *historical* one). The algorithm was then implemented on the third day.

The assignment matrix A was computed in the same way as before (see Section 1.3.1), but now the travel time is larger (because of congestion). After inspection of the video data, it is important to note that travel time during the period under analysis –and across lanes– is not constant, and it has a significant variability. As a first approach, however, we have decided to run the algorithm assuming a constant travel time of 1 minute. This means that a vehicle entering in time interval h leaves the network in time interval $h+2$.

Results. Figures 1.12 to 1.15 show both the real and the estimated OD flow for each OD pair (note that the graphs use different scales).

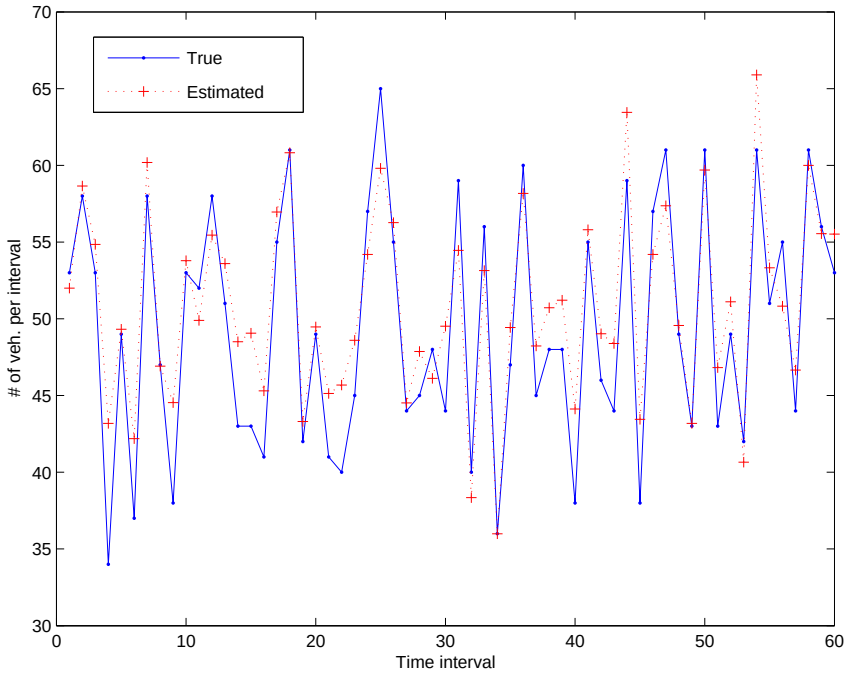


Figure 1.12: OD pair 1, under congested conditions.

Unfortunately OD pair 4 (Figure 1.15, from the on-ramp to the off-ramp) has only one interval with one vehicle. As we saw in the free flow case, this might make the estimation process easier.

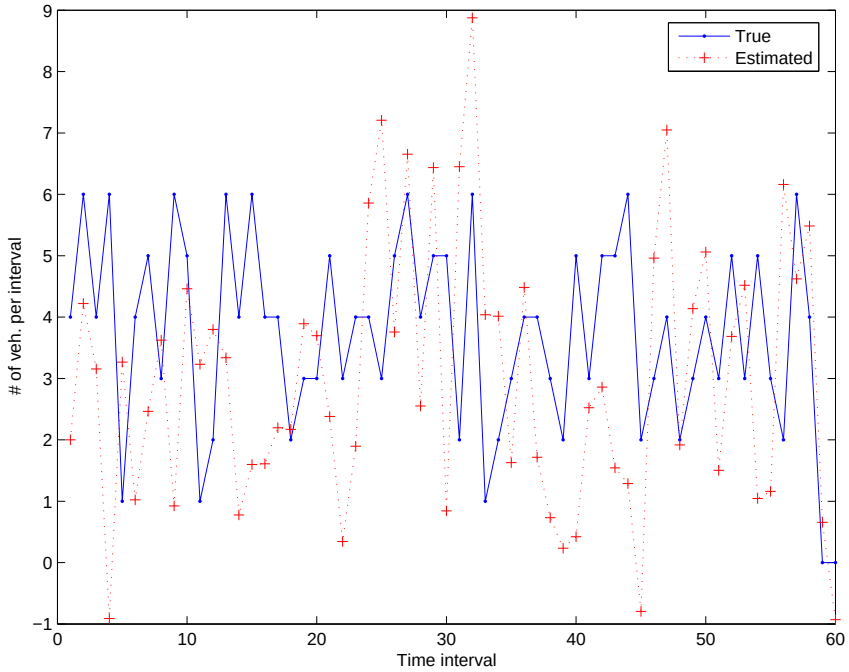


Figure 1.13: OD pair 2, under congested conditions.

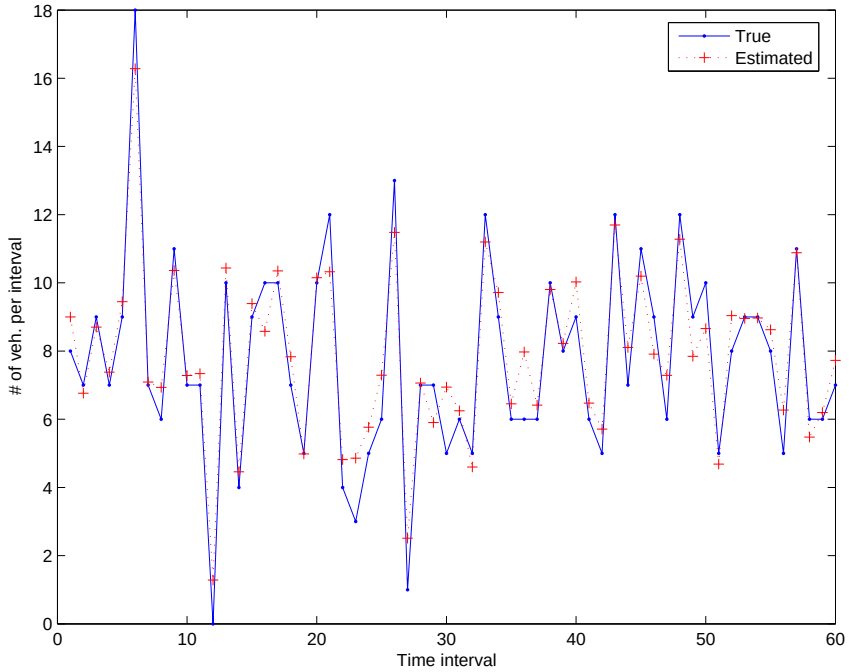


Figure 1.14: OD pair 3, under congested conditions.

1.3.3 Comparison of the two methodologies

Free-flow data from MT and congested data from BHL were implemented with both methodologies. Two error measures were computed in order to compare performances between the methodologies for each case [2]:

- Root mean square error (RMS) = $\frac{\sqrt{\sum_i (x_i - \hat{x}_i)^2}}{N}$

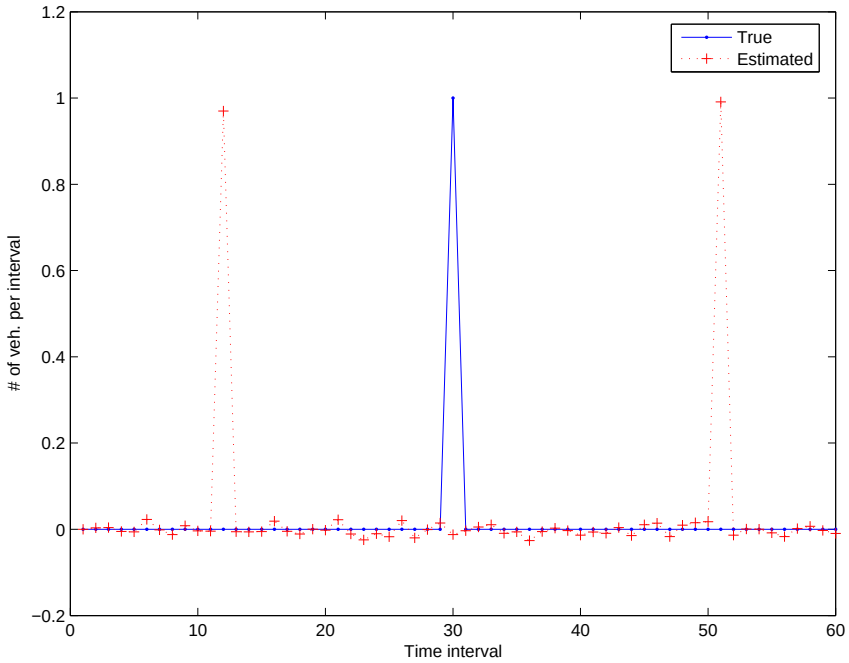


Figure 1.15: OD pair 4, under congested conditions.

- Root mean square error normalized (RMSN)= $\frac{\sqrt{N \cdot \sum_i (x_i - \hat{x_i})^2}}{\sum_i x_i}$

where, x and $\hat{x}$ are true and estimated OD flows respectively and the summation ranges over all OD pairs and all intervals over which analysis is performed. Table 1.1 shows the results.

Table 1.1: Comparison of both methodologies

	MT (free flow)		BHL (congested)	
	W/O deviations	With deviations	W/O deviations	With deviations
RMS	4.775	4.694	6.282	2.215
RMSN	0.476	0.468	0.449	0.146

In both cases the methodology with the deviations provides better results. Event though the statistics for the two methodologies are close in the MT case, Figure 1.16 shows that the methodology with deviation provides better results. The figure shows the agreement between the true and the estimated OD flows, for OD pair 9 (the same as in Figure 1.8), using the two methodologies. Clearly the methodology with deviations agrees much better with the real data.

Because of the reason stated in Section 1.2.2, this result is not surprising.

1.4 Transition period: how to detect when traffic conditions change

Travel time is a good indicator of the traffic conditions on a section of highway: the larger the travel time is, the more congested is the section. Assignment matrix A in Equation

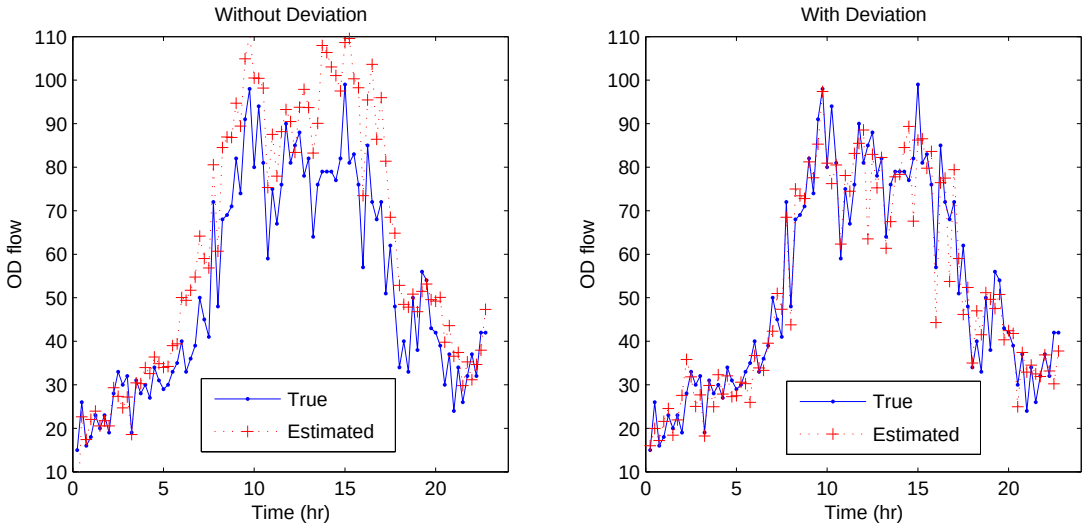


Figure 1.16: Visual comparison between the methodologies

(1.7) depends on travel times (and thus, on traffic conditions). If we knew travel times between stations, we would be able to directly compute the assignment matrix A and run the algorithm. Two questions arise at this point:

- How accurate does the travel time estimation need to be? If the model is very sensitive to the matrix A, the travel time estimation should be very accurate. On the other hand, if the matrix A does not affect the model in a significant manner, we can rely in less accurate (but easier-to-implement) methods to estimate travel time.
- How can we obtain travel times between stations (in real time)? This, of course, is going to depend on the answer to the first question.

1.4.1 How matrix A affects the model: sensitivity analysis

In order to determine how accurate the matrix A (and then, travel times) needs to be, a sensitivity analysis was performed. That is, implementations of the algorithm using different travel times were performed, which yield different assignment matrices A. The idea is to see how worse the result would be by assuming a wrong travel time.

For the BHL network six different travel times were tried (in the congested regimen): 18 (free flow), 30, 50, 70, 90 and 120 seconds[14]. As was stated in Section 1.3.2, the actual travel time is 60 seconds. The 18 seconds travel time is tried to see what the estimation results would be if free flow conditions are assumed when traffic is actually congested. The results in terms of the error measures described in Section 1.3.3 (RMS and RMSN) are shown in Table 1.2.

There are three interesting points to mention here. First, 90 seconds travel time gives the best error measures. This might be due to the fact that travel time was not constant

[14]In a 0.3 miles stretch of highway, these times imply the following speeds respectively: 65 (free flow), 40, 24, 17, 13, and 10 mph.

Table 1.2: Error measures for different travel times

	60 sec (real)	18 sec (ff)	30 sec	50 sec	70 sec	90 sec	120 sec
RMS	2.215	3.284	2.531	2.334	2.165	2.026	2.035
RMSN	0.146	0.216	0.166	0.154	0.142	0.133	0.134

during the period under study (as mentioned in Section 1.3.2), and 90 seconds was a better estimation of the travel time.

The second point to note is that error measurements –for travel times that assume certain degree of congestion (i.e. greater or equal to 30 seconds)– do not vary too much from one simulation to another. In fact, the visual agreement between true and estimated curves are quite similar in all cases (Figure 1.17 a) and b) show the cases when travel time is 30, and 120 seconds for the first OD pair).

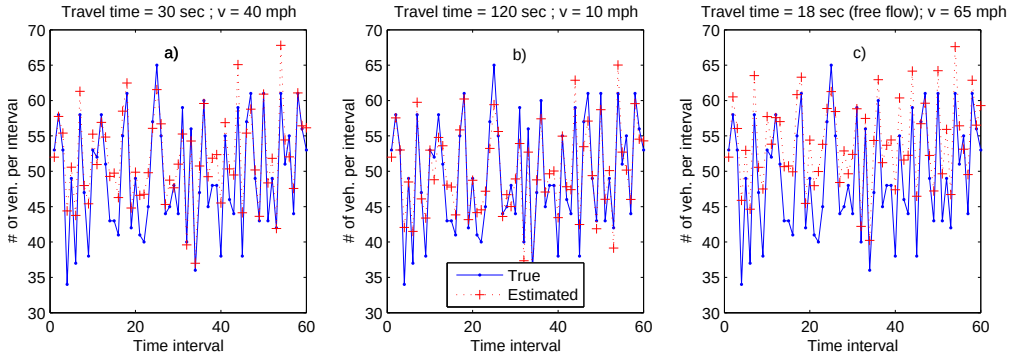


Figure 1.17: Visual agreement between true and estimated curves for first OD pair (BHL) and for different travel times.

Lastly, if free flow conditions are assumed (travel time 18 seconds), the RMS and RMSN increase significantly. As can be seen from Figure 1.17 c), in this case the visual agreement between true and estimated curves is not good. That is the reason for the large RMS and RMSN observed in Table 1.17.

The second and third points suggest that we need to be able to distinguish between free flow and congested regimes. However, if there is congestion, an accurate travel time estimation is not needed (because different matrices A will yield quite similar results). The importance of this result will become clear in the next section.

1.4.2 Travel time estimation

Different methods to estimate travel time can be found in the literature. The following are three different approaches:

- If we know the uniform speed along the whole section, we can use $t = \frac{d}{v}$ (t is time, d is distance, and v is speed) to estimate travel times. The speed could be directly measured from the detectors (using dual-loop detectors, for instance). We can then assume that this speed is the same along the segment. However, assuming that the speed at the detector station location ("point speed") corresponds to the speed

over the whole section does not seem to be a good approximation (see for instance Table 1 in [6]).

- Another option is to construct cumulative curves from consecutive loop detector stations (N-t curves, see [7] or [8]). The horizontal difference between two cumulative curves represents the travel time between two stations. This approach assumes FIFO behavior (which seems reasonable) and conservation of vehicles between stations (no ramps). The latter assumption can be treated in some way (by adding ramp counts to the mainline counts for instance), but it would introduce some error in the estimation. The estimation quality will also depend on the frequency of speed measurements.

- A third option consists in using a discrete version of the q-k diagram. Instead of an infinite number of possible states (infinite number of k or speed), we recognize three or four states (or *modes*). Each one of these states has a travel time associated with it, and thus an assignment matrix A too. For instance, if we are running the algorithm on the BHL network, we would use matrices computed in Section 1.3.1 when free-flow conditions are detected, and matrices from Section 1.3.2 when congestion arises. This example assumes only two modes: the whole section freely flowing or the whole section congested at the same level. In [9], the author assumes that the mode cannot be directly observed from the data. The mode jumps, however, follow a discrete-time Markov chain, with certain transition probability. Here, the algorithm would jointly estimate the mode and the OD flows at each time interval.

Given the results shown in Section 1.4.1, it seems like we do not need an accurate travel time estimation. In light of the third approach discussed earlier, we could say that speeds within a certain range (and travel times corresponding to these speeds) yield the same assignment matrix A. That is, the fundamental diagram $q - k$ has been *discretized* into a few modes. Then, the algorithm might use loop detector data (moving average of speed measurement) to determine under which mode the section is working, and use the corresponding matrix A.

We can identify as many mode as we want. For instance, one mode for the free-flow conditions, and two or three modes for different levels of congestion on the whole section[15] (and thus, different travel times and different A matrices). The first mode (free flow) would correspond to speeds greater or equal to v_1, the second mode would include speeds between v_1 and v_2, and the third mode would contain speeds less than or equal to v_2. Then, we would set the following rule for each observation $\bar{v}(k)$ (moving average of speed):

· If $v_1 \leq \bar{v}(k) \rightarrow$ mode 1 $\rightarrow$ pick A_1.

· If $v_2 < \bar{v}(k) < v_1 \rightarrow$ mode 2 $\rightarrow$ pick A_2.

· If $\bar{v}(k) \leq v_2 \rightarrow$ mode 1 $\rightarrow$ pick A_3.

[15]For reasons stated in Section 1.1, we will not consider modes where a shock is traveling the section. For instance, upstream congested and downstream freely flowing is not a possible mode. Those periods of time should be very short (especially if the network is short) and can be ignored.

Section 1.4.1 stated that estimations during congestion do not change too much for different travel times (as long as they actually correspond to congested conditions). This encourage us to think of only two modes, free flow and congested modes, and only one matrix A associated with each mode. Two issues remain unsolved so far: what should be the speed threshold v_{thr} that determines the limit between free flow and congested regimes; and what speed should be used to compute the matrix A in each mode. The work of Varaiya [10] can be used to address both questions. In that work, the author showed that, on a certain highways in California (I-10E), drivers spent very little time at transitional speeds (between 40-50 mph). On average, most of the time they drive at 30 mph (35% of the time) and 60 mph (65%). If this is extrapolated to other highways in California (which seems reasonable), the threshold speed v_{thr} could be set at 45 mph and matrices A may be computed using a speeds of 60 and 30 mph (for free flow and congested conditions respectively).

In order to test if $v_{thr} = 45$ mph is a good approximation or not, data from another detector station on BHL were collected. The data consist on the 30-seconds average speeds across four lanes (not including the HOV lane) collected from 4am to 8:30pm during a weekday. The moving average over 3 minutes[16] was computed in order to filter some variability of the data. Then, each filtered observation was compared against the threshold $v_{thr} = 45$ in order to determine the mode. Figure 1.18 shows both the speed profile read by the detector (lower curve) and the mode predicted using the algorithm just described (uper curve). Using $v_{thr} = 45$ mph yields a mode sequence that is in very good agreement with what happens in reality.

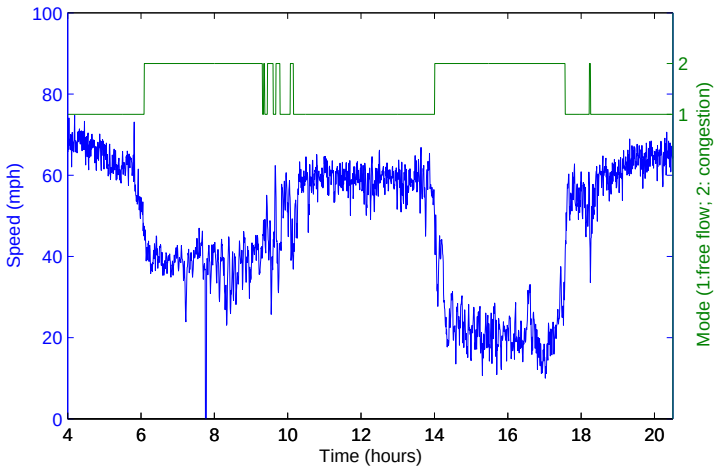


Figure 1.18: Speed measured from detector and mode predicted by setting $v_{thr} = 45$ mph.

From the speed profile, morning and evening rush periods can be clearly identified (6-10am and 2:30-5:30pm respectively). The average speed is around 40 mph in the morning and 30 mph in the evening. If we would have set the threshold at 40 mph, the morning rush period would have been jumping between the two modes. Section 1.4.1, however, provided reasons to believe that for our purposes, morning and evening rush

[16]That is, the filtered observation at time k is given by: $\bar{v}(k) = \frac{\sum_{i=k-6}^{k} v(i)}{7}$, where $v(i)$ is the average speed at interval i across all lanes.

periods are the same. Because of this, it would be preferable to set the threshold not below 45 mph.

In summary, given that the estimation results are not too sensitive to different travel times during congestion, only two matrices A should be computed for each network: one for the free flow regime (using a speed of 60 mph) and the other one for the congested regime (with a speed of 30 mph). Using speed measurements from detectors, we have proposed a very simple way to determine which of those matrices should be used (i.e. to detect when the traffic mode changes).

1.5 Consideration of network size

OD flows on a highway can be useful to implement control strategies (such as ramp metering) or to make operational studies in certain area (for instance, on weaving sections). For this type of use, we really need to know how many vehicles are going to make use of every ramp. For example, in Figure 1.19 we do not really need to know how many vehicles are going from point 1 to $n-1$ or n. We just need to know how many vehicles are going to take one of the next off-ramps (points 4 and 6) and how many are going through the highway (and the same is valid for those vehicles coming from point 2).

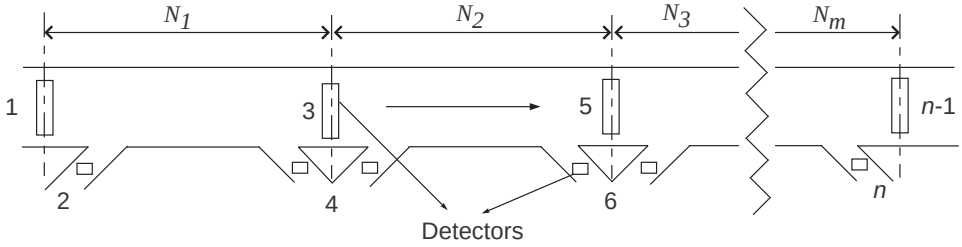


Figure 1.19: Imaginary network

For this reason, we think that it might be reasonable to divide the highway into smaller networks (for instance, N_1, N_2,,... N_m in Figure 1.19), and then run the algorithm on each one of these networks. What are the *pros* and *cons* of doing that?

- *Pros*:

  → Transition period (period when there is shock traveling through the network, which creates two modes) in each network: the shorter the network, the shorter the transition period on that network is. In Section 1.4, we stated that these periods will be ignored. If we consider the whole network in Fig.1.19 as one network (using $n = 8$ and $m = 3$), a shock with speed of 12 mhp would take 15 minutes[17] to cover the whole section (i.e. the algorithm would not work during 15 minutes). However, if we divided the network into N_1, N_2, and N_3, the algorithm would not work only during 5 minutes for each network[18].

[17]This assumes N_1, N_2 and N_3 are 1 mile long

[18]Each network, in this case, is independent on the adjacent network's estimation. That is, we can run our algorithm in N_1 and N_3, without information from N_2.

Clearly, for larger networks (i.e. larger n) the transition period cannot be ignored.

→ Assignment matrix A estimation: since the network is short, there would be few OD pairs and few link detector stations. Then, the estimation of the A matrix would be easier.

→ BHL network is like any of the short networks in Fig.1.19, and we have already validated the algorithm for both free flow and congested conditions in this network.

- *Cons*:

  → In Fig.1.19 we would not know the flows from point 1 to 6, $n-1$ or n, or from point 4 to $n-1$ or n, and so on. However, as we stated before, this is information that we do not really need for control purposes.

1.6 Conclusions

Real-time OD matrix estimation methods reported in the literature work under stationary conditions, i.e. their performance when traffic conditions change abruptly is not good. We have implemented a methodology that identifies when traffic conditions change and then makes use of existing models to estimate OD flows in a linear network. The models were tested using real data collected from two different network and during free flow and congested conditions. The estimations match the exact data well.

One parameter of the model used depends on traffic conditions (more precisely travel times). We have shown preliminary evidence to support the idea that an accurate estimation of the travel time is not needed, and a distinction between free flow and congested conditions is enough for our purposes. A very simple algorithm to detect these conditions, which makes use of the speed measurements provided by loop detectors, was proposed.

We have not mentioned the fact that matrix F in Equation (1.7) also depends on traffic conditions. Matrix F assumes that previous OD flows affect the current one in a linear way (autoregressive form in Equation (1.1) and (1.3)). If we have evidence to believe that these influences or effects are different from free flow to congested regime, we will end up with two matrices F instead of one. Estimation of each one of these matrices would be as described in Section 1.2.2.

A practical implementation of the methodology proposed should not be hard. Suppose we are interested in computing OD flows for a given network from 4am to 9pm on a weekday. The data needed to implement the methodology would consist of:

- Two sets of historical counts at every entry and exit point of the network: the first set of counts will be used to compute a first OD matrix based on optimization models described in the literature (such as the one described in [2]). If there exist a historical OD matrix for the network, these counts would not be needed. This first OD matrix will be used to compute matrix F in Equation (1.7). The second set is to implement the methodology without deviations in the network using the matrix F already estimated. As a result, we will have two OD matrices.

- Counts and speed measurements at every entry and exit point of the network for the period of interest: this is the information needed to run the whole methodology and to obtain estimations.

Finally, these data might be fed into a computational tool that will estimate the OD flows for the period under analysis. Based on the results reported here, these estimations should be accurate enough.

Bibliography

[1] J.Krogmeier S.R.Hu, S.Madanat and S.Peeta. Estimation of dynamic assignment matrices and OD demands using adaptive kalman filtering. *ITS Journal*, 6:281–300, 2001.

[2] K.Ashok and M.Ben-Akiva. Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. *Transportation Science*, 34(1):21–36, 2000.

[3] K.Ashok and M.Ben-Akiva. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. *Transportation Science*, 36(2):184–198, 2002.

[4] B.Anderson and J.Moore. *Optimal Filtering*. Prentice Hall Inc., 1979.

[5] M.Bierlaire and F.Crittin. An efficient algorithm for real-time estimation and prediction of dynamic od flows. *Operations Research*, 52(1):116–127, 2004.

[6] B.Coifman. Estimating travel times and vehicles trajectories on freeways using dual loop detectors. *Transportation Research A*, 36(4):351–364, 2002.

[7] G.F.Newell. *Applications of Queueing Theory*. Chapman & Hall, London, 2nd edition, 1982.

[8] C.F.Daganzo. *Fundamentals of Transportation and Traffic Operations*. Elsevier Science Inc., New York, 1997.

[9] X. Sun. *Modeling, Estimation, and Control of Freeway Traffic*. PhD thesis, University of California, Berkeley, 2005.

[10] P.Varaiya. Reducing highway congestion: an empirical approach. *European Journal of Control*, 11:301–309, 2005.

.1 State augmentation

This appendix aims to explain and show how the state augmentation process (described in Section 1.2.3) works. For this purpose we will do the whole process using a small network.

The network in Figure 20 contains one origin (A) and two destinations (B, and C). There are three OD pairs, which are labeled as follows:

· From A to B: OD pair 1.

· From A to C: OD pair 2.

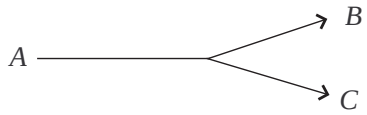


Figure 20: Small network with 2 OD pairs

In this case, $n_{OD} = 2$ and $n_L = 3$. For simplicity with the notation, let us assume that $p = 2$ and $u = 1$ (and then $s = \max(u, p-1) = 1$) and that we are working with the methodology without the deviations. Equations (1.1) and (1.2) are as follows:

$$x_{h+1} = f_{hh} \cdot x_h + f_{h-1,h} \cdot x_{h-1} + w_h \quad (9)$$

$$y_h = a_{hh} \cdot x_h + a_{h-1,h} \cdot x_{h-1} + v_h \quad (10)$$

Each term in previous equations are described next:

- State vector at time interval h: $x_h = \begin{pmatrix} x_{1h} \\ x_{2h} \end{pmatrix}$
- Two 2×2 matrices describing the effect of previous OD flows on the current one:
  $f_{hh} = \begin{pmatrix} f_{11}^h & 0 \\ 0 & f_{22}^h \end{pmatrix} \quad f_{h-1,h} = \begin{pmatrix} f_{11}^{h-1} & 0 \\ 0 & f_{22}^{h-1} \end{pmatrix}$
  Note that these forms assume that OD pair 1 does not affect OD pair 2 and viceversa.
- Random error: $w_h = \begin{pmatrix} w_{11} \\ w_{21} \end{pmatrix}$
- Vector of counts: $y_h = \begin{pmatrix} y_{Ah} \\ y_{Bh} \\ y_{Ch} \end{pmatrix}$
- Two 3×2 matrices describing how OD flows affect counts:
  $a_{hh} = \begin{pmatrix} 1 & 1 \\ a_{21}^h & 0 \\ 0 & a_{32}^h \end{pmatrix} \quad a_{h-1,h} = \begin{pmatrix} 0 & 0 \\ a_{21}^{h-1} & 0 \\ 0 & a_{32}^{h-1} \end{pmatrix}$
- Measurement errors: $v_h = \begin{pmatrix} v_{11} \\ v_{21} \\ v_{31} \end{pmatrix}$

State augmentation is done in order to bring Equations (9) and (10) into the form of Equations (1.5) and (1.6), respectively. The new representation will be given by:

$$X_{h+1} = F_h \cdot X_h + W_h \quad (11)$$

$$Y_h = A_h \cdot X_h + v_h \quad (12)$$

Each term now is as follows:

- State vector at time interval h: $X_h = \begin{pmatrix} x_{1h} \\ x_{2h} \\ x_{1h-1} \\ x_{2h-1} \end{pmatrix}$

- Square matrix $F_{4\times4}$ describing the effect of previous OD flows on the current one:

$$F_h = \begin{pmatrix} f_{hh} & f_{h-1,h} \\ I_{2\times2} & 0_{2\times2} \end{pmatrix} = \begin{pmatrix} f_{11}^h & 0 & f_{11}^{h-1} & 0 \\ 0 & f_{22}^h & 0 & f_{22}^{h-1} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

- Random error: $W_h = \begin{pmatrix} w_h \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} w_{11} \\ w_{21} \\ 0 \\ 0 \end{pmatrix}$

- Vector of counts (same as before): $Y_h = \begin{pmatrix} y_{Ah} \\ y_{Bh} \\ y_{Ch} \end{pmatrix}$

- Matrix $A_{3\times4}$ describing how OD flows affect counts:

$$A_h = \begin{pmatrix} a_{hh} & a_{h-1,h} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ a_{21}^h & 0 & a_{21}^{h-1} & 0 \\ 0 & a_{32}^h & 0 & a_{32}^{h-1} \end{pmatrix}$$

- Measurement errors (same as before): $v_h = \begin{pmatrix} v_{11} \\ v_{21} \\ v_{31} \end{pmatrix}$

Even though we have used subindex h in matrices F, W, A, and v, in practice we assume that these matrices are time invariant. Finally, Equations (11) and (12) are the ones that are finally fed into the Kalman filter to obtain OD flows estimates.

CHAPTER 2

ESTIMATING TIME-DEPENDENT FREEWAY O-D DEMANDS WITH VARIATIONAL INEQUALITIES

Prepared by:

Michael Zhang, Yu Nie and Zhen Qian

Department of Civil & Environmental Engineering

University of California, Davis

2.1 Introduction

2.1.1 The Dynamic O-D Estimation Problem

Traditionally, an O-D table concerns trips made over a relatively long time period (e.g., morning peak time) within which the traffic condition is assumed to be homogeneous. Such O-D demands are intended to be used with “static” travel forecasting models. However, it is widely recognized that static models are inadequate to predict the evolution of traffic pattern over time of day. Traffic congestion is essentially a dynamic phenomenon. First of all, travel demands do fluctuate over time of day. During morning commute, for example, demand levels change substantially as travelers adapt to time-varying traffic conditions by routing and scheduling of departure times. Recurrent traffic congestion often seen in urban areas is mainly a result of the way such fluctuations take place in space and time. Namely, imbalanced distribution of demand causes the shortage of road supplies during peak time at various locations (bottlenecks), where queues develop and spread over the network.

Therefore, an O-D table that reasonably reflects temporal distribution is often indispensable for dynamic travel forecasting models, which target a wide spectrum of applications ranging from short-term planning to within-day traffic control/management. However, getting reliable dynamic travel demands is notoriously difficult. In typical travel diary data, temporal information (e.g., starting time and duration) is only available for trips of certain purposes (mainly home-based work and school). The dynamic distribution factors used in practice are often aggregated based on these trips only, thus not necessarily representative for trips of other types, such as shopping and recreation trips. To extract temporal information for those trips call for activity-based travel demand models, which remains a state-of-the-art until very recent, we note that the distribution of demands in time can naturally arise from forecasting models themselves when individuals' departure time choices are endogenized, as first shown by William Vickery . However, a demand pattern established from such a DTA model highly depends on individuals' preferred arrival time windows and how they price unpunctual arrivals. Not surprisedly, these behavior parameters are difficult to calibrate, and the assumptions that intend to simplify the problem are often too strong to be realistic. Consequently, although existing DTA models with departure time choice may provide useful insights to macroscopic policy analysis, they hardly yield a dynamic O-D table more than conceptually meaningful. Moreover, commuters' scheduling of departure time may not fully explain demand fluctuations during the rush hour.

In a nutshell, determining the up-to-date time-varying travel demand pattern in a highway network remains a challenging and to some extent unresolved issue. On the one hand, travel demands obtained from large-scale surveys not only come with high prices (in terms of monetary, time and labor costs), but also are likely to be out-of-date. More importantly, household surveys based on travel diaries do not typically provide temporal trip information with a resolution adequate for dynamic travel forecasting. On the other hand, although travel forecasting models may be used to establish the dynamic pattern of travel demands, the outputs substantially depend on individuals'

travel behaviors and thus may not be reliable. This explains why substantial research efforts have been invested in "estimating" O-D demands from various traffic surveillance data, which can be automatically collected at relatively low costs. Among different traffic data, link flows (e.g., traffic counts from loop detectors) are most widely used.

In the next few sections, we will present a new Dynamic OD Estimation (DoDE) problem for freeways based on variational inequalities. The freeway networks we consider are similar to thee one shown Figure 2.1, where each O-D pair has a unique path.

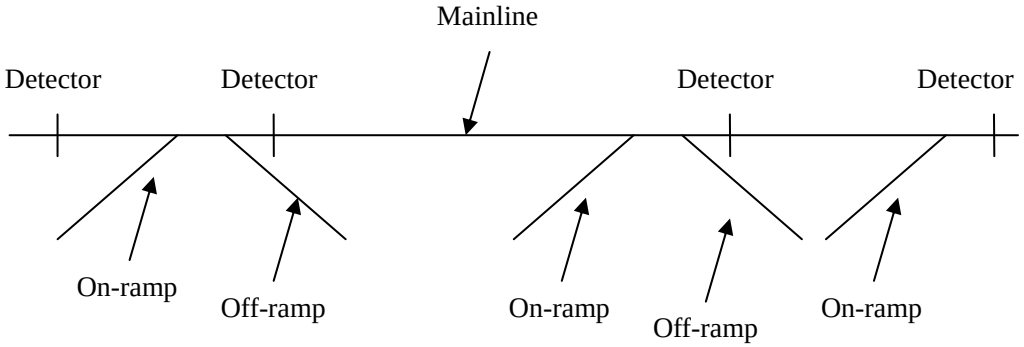


Figure 2.1 Simplified highways

Consider a freeway network $G(N, A)$, where N and A are the sets of nodes and links respectively. Nodes are the locations where traffic flow will merge into the mainline or leave it. Links are made up of ramp and mainline links. Let R and S represent the set of origins (the start nodes of On-ramps) and destinations (the end nodes of Off-ramps. The end of the freeway is considered a special off-ramp) respectively. The cardinalities of the sets nodes, links and O-D pairs are denoted as $|N| = m$, $|A| = n$ and $|R \times S| = o$ respectively. Let $[0, T]$ be an assignment horizon (i.e., the analysis period). The network is assumed to be empty at $t = 0$, and only travel demands departing within the assignment horizon are considered. Corresponding to the assignment period, we define a loading horizon $[0, T']$, where T' marks the time when all traffic clears the network. Let $q_{rs}(t)$ be the travel demand between O-D pair rs departing at time t, and the total demand for the whole assignment horizon is

$$q_{rs} = \int_0^T q_{rs}(t)dt$$

Let c_{rs}^t denotes the time-dependent travel time for OD pair *rs*, f_{rs}^t the departure flow rate for OD pair *rs*, all during the assignment interval t. Now, suppose that loop detectors are installed at the entrances of a set of selected links, A_o, so that traffic flow

entering any link $a \in A_o$ at any time $t \in [0, T']$, $\bar{x}_a(t)$, can be measured. Let $n_o = |A_o|$ be the number of the observed links, ϕ_a be an assignment interval, a discrete duration during which the departure flow rate for any O-D pair is assumed to be constant,(m_a, the number of assignment intervals is given by $T = m_a \phi_a$), ϕ_m be the measurement interval, a discrete duration for which the measured traffic quantities is aggregated and recorded (a loading horizon consists of $m_m = T'/\phi_m$ measurement intervals of uniform length), and ϕ_l be the loading interval, a discrete duration during which network conditions are assumed to be stationary (loading horizon consists of m_l loading intervals of uniform length, i.e., $T' = m_l \phi_l$), then by conservation, we have the following relationship between O-D flow and measured link flow, assuming there are no measurement errors:

$$\sum_r \sum_s \sum_{t=1}^{m_a} p_{rsa}^{th} f_{rs}^t = \bar{x}_a^h, \forall r \in R, s \in S, a \in A_o, h = 1, 2, ...m_m$$

where

$$p_{rsa}^{th} = \frac{1}{i_a} \sum_{d=1}^{i_a} \delta_{rsa}^{t_d h}$$

and

$$\delta_{rsa}^{t_d h} = \begin{cases} 1 & \text{if traffic departing } r \text{ at the } d\text{th loading interval of the } t\text{th assignment interval heading for destination } s \text{ arrives link } a \text{ during the } h\text{th measurement interval} \\ 0 & \text{otherwise} \end{cases}$$

and $i_a = \phi_a / \phi_l$. In vector form, it becomes

$$\mathbf{Pf} = \bar{\mathbf{x}} \tag{2.1}$$

We emphasize that the mapping $\mathbf{P}$, which relates time-dependent O-D flow to time-varying link flow measurements, is endogenously determined. That is, $\mathbf{P}$ has to be updated in accord with the change of the path flow pattern $\mathbf{f}$ in the estimation process. This in turn requires time-dependent link travel times to be computed from a given $\mathbf{f}$, a problem known as dynamic network loading (DNL). Underlying the DNL process is a network traffic model that describes the evolution of traffic flow. Dictated by this traffic model, the mapping between $\mathbf{P}$ and $\mathbf{f}$ is typically quite convoluted and not in a closed, analytical form.

Generally, (2-1) is underdetermined and thus has many solutions. To resolve this problem, additional information should be supplied. Such information can be roughly categorized into two types:

1) A partial or complete “base” O-D table, which is often established from existing survey data, called a historical O-D table.

2) Path travel times obtained from probe vehicles.

In the remainder of this report, we'll present a new framework for DoDE problem based on variational inequalities (VI). In this framework, the DoDE problem is first transformed into a VI problem, then a DNL based on the LWR model is used to evaluate the mapping **P** and path travel times, finally two solution algorithms, the basic projection algorithm and the method of successive averages (MSA) are suggested to solve the VI based DoDE problem.

2.2 The Proposed Estimation Framework and Solution Algorithms

2.2.1 The DoDE Formulation

Our objective is to obtain a time-dependent O-D demand pattern in terms of time-dependent path flows that, once loaded onto the network, can reproduce observed link traffic counts and path travel times as closely as possible. This can be cast into a psudo generalized least squares problem of the following form:

$$\min z(\mathbf{f}) = 0.5w_x(\mathbf{Pf} - \bar{\mathbf{x}})^2 + 0.5w_q(\mathbf{Mf} - \bar{\mathbf{q}})^2 + 0.5w_p s_p\left(\mathbf{c}(\mathbf{f}) - \bar{\mathbf{c}}\right)^2 \tag{2.2}$$

subject to:

$$\mathbf{f} \geq 0$$

where the matrices P and M map time-dependent path flows (here also O-D flows) into time-varying link traffic counts and historical O-D flow rates, respectively, and the positive scalars w_x, w_q and w_p are weights placed on traffic counts, historical O-D demands, and path cost observations, respectively, and the positive scalar s_p is added to properly scale two different types of quantities, flow and cost in the objective function.

The optimality conditions of (2.1) are:

$$\begin{aligned} \langle \nabla z(\mathbf{f}), \mathbf{f} \rangle &= 0 \\ \nabla z(\mathbf{f}) &\geq 0 \\ \mathbf{f} &\geq 0 \end{aligned} \tag{2.2*}$$

Because the mapping P (which depends on f) and c(f) are non-linear and possibly non-convex, the above problem is extremely difficult to solve directly. With the introduction of count, O-D and cost deviations, however, the above optimality conditions can be case into a variational inequality (VI), hence allowing VI solution algorithms be employed to solve the DoDE problem.

For any given path flow pattern $\mathbf{f} \geq 0$ and a set of observed link traffic counts $\bar{\mathbf{x}} > 0$, the count deviation dx_a^h is defined for each link a as

$$dx_a^h = \begin{cases} \bar{x}_a^h - \sum_r \sum_s \sum_{t=1}^{m_a} p_{rsa}^{th} f_{rs}^t, h = 1,2,...m_m & a \in A_o \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where

$$p_{rsa}^{th} = \frac{1}{i_a} \sum_{d=1}^{i_a} \delta_{rsa}^{t_d h}$$

and

$$\delta_{rsa}^{t_d h} = \begin{cases} 1 & \phi_m (h-1) < e_a^{t_d} \le \phi_m h \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

where $e_a^{t_d}$ is the time when a vehicle would enter link a of the path connecting an O-D pair and depart at $(t-1)\phi_a + (d-0.5)\phi_l$. However $\delta_{rsa}^{t_d h}$ is not a continuous function of the entry time $e_a^{t_d}$. The loss of continuity may cause non-existence of solutions. To resolve this issue we replace the dynamic path-link incidence relationship (2.4) with the following one:

$$\delta_{rsa}^{t_d h} = \max\left\{0, \frac{\phi_m - |\, e_a^{t_d} - \phi_m (h-0.5)\,|}{\phi_m}\right\} \tag{2.5}$$

For any given path flow pattern $\mathbf{f} \ge 0$ and a set of historical O-D demand pattern $\bar{\mathbf{q}} > 0$, the O-D deviation d_{rs}^t is defined for each O-D pair rs as

$$dq_{rs}^t = \begin{cases} \bar{q}_{rs}^t - f_{rs}^t, t = 1,2,..., m_a & \text{If } \bar{q}_{rs}^t \text{ is given for O - D pair } rs \text{ at } t \\ 0 & \text{otherwise} \end{cases}$$

For any given path flow pattern $\mathbf{f} \ge 0$, and a set of observed path travel times $\mathbf{c} = \left\{\bar{c}_{rs}^t \mid r,s,t\right\}$ the cost deviation is defined as

$$dc_{rs}^t = \sum_{r,s,t} \frac{\partial c_{rs}^t(\mathbf{f})}{\partial f_{rs}^t} [\bar{c}_{rs}^t - c_{rs}^t(\mathbf{f})]$$

With count, O-D and cost deviations, we can now define the path deviation d_{rs}^t as follows:

$$d_{rs}^t = w_x \sum_a \sum_{d=1}^{i_a} \delta_{rsa}^{t_d h} \frac{dx_a^h}{i_m} + w_q dq_{rs}^t + w_p s_p dc_{rs}^t \tag{2.6}$$

and its matrix form is given by

$$\mathbf{d}(\mathbf{f}) = \mathbf{P}^T w_x (\bar{\mathbf{x}} - \mathbf{P}\mathbf{f}) + \mathbf{M}^T w_q (\bar{\mathbf{q}} - \mathbf{M}\mathbf{f}) + w_p s_p J_c (\bar{\mathbf{c}} - \mathbf{c}) \tag{2.7}$$

where J_c is the Jacobian matrix of $\mathbf{c}(\mathbf{f})$ with respect to path flow **f**.

It can be shown that the optimality conditions for (2.2) can be transformed into the variational inequality (VI):

$$\text{Find } \tilde{\mathbf{f}} \in R_+^{|o \times m_a|} \text{ such that } \langle -\mathbf{d}(\tilde{\mathbf{f}}), \mathbf{f} - \tilde{\mathbf{f}} \rangle \ge 0 \text{ for all } \mathbf{f} \in R_+^{|o \times m_a|} \tag{2.8}$$

which lends the problem to easy solution through various VI solution algorithms.

It should be noted that J_c does not have a closed form and thus its evaluation usually requires numerical approximation. For every element in J_c, we have:

$$\frac{\partial c_{rs}^{t}(\mathbf{f})}{\partial f_{rs'}^{t'}} = \frac{\sum_{d'=1}^{i_a}\sum_{d=1}^{i_a}\frac{\partial c_{rs}^{t_d}(\mathbf{f})}{\partial f_{rs'}^{t'_{d'}}}}{i_a^{\,2}} \tag{2.9}$$

where *d* is an loading interval in the *t*th assignment interval, and *d'* is an loading interval in the *t'*th assignment interval.

Consider the marginal travel time of adding one more unit vehicle between any two loading intervals in the loading horizon. Let *t''* be the departure time of a vehicle departing at the *t'*th interval between O-D pair *r's'* , without the additonal vehicle. Let *t'''* denote the time the queue on that link disappears after *t''* without the additional vehicle (if there is no queue at t'', then there is no delay on this link). Let t_a denote the arrival time of a vehicle departing at the *t*th interval between O-D pair *rs*, without the additional vehicle. Let t_b denote the departure time of a vehicle departing at the *t*th interval between O-D pair *rs,* without the additional vehicle. There are two conditions to consider in evaluating the marginal path travel time:

1) For O-D pair *rs*, if its on-ramp *r* is ahead of the on-ramp *r'* of O-D pair *r's'* (or *r*=*r'*), we shall first search the links along the path of *r's'* to get [*t''*, *t'''*], the effective congestion interval, for every link. Then, search every link along the path of O-D pair *rs* to obtain its arrival and departure times for flows along that path, and if these arrival times are earlier than their corresponding *t''* or later than their corresponding *t'''*, congestion due to the flow from *r's'* will not affect the flow of O-D pair *rs*. When the flow of O-D pair *rs* arrives at any link during that link's effective congestion interval [*t''*, *t'''*], then its contribution to the marginal travel time, i.e. the reciprocal of the capacity of that link, will be accumulated along its path.

2) Reversely, if its on-ramp *r* is after the on-ramp *r'* of O-D pair *r's'* , we again search the links along the path of *r's'* to get their effective congestion intervals [*t''*, *t'''*]. Then, from the list of *t'''*, find the one corresponding to the link along the path of *r's'* that merges with the path of O-D pair *rs.* If this link is not congested, then congestion due to flows on path *r's'* will not contribute to the marginal travel time of O-D pair *rs*. Otherwise, perform the same adjustment and computation as in 1).

On the other hand, the use of the path deviation function provides a flexible framework to fuse different observations together. The optimality condition, in the form of a VI, is quite general and may be applied even when the estimation problem cannot be cast as a mathematical programming problem.

To solve the VI problem (2-7), we need to evaluate the path deviation vector **d**(**f**), which requires dynamic network loading, a process of loading time-dependent O-D flows onto the network according to a given model of traffic flow dynamics. This process produces the time-dependent link flows and travel times, and is described in the next section.

2.2.2 The Dynamic Network Loading Process

As mentioned earlier, the DNL problem aims at obtaining, on a congested network and over a fixed time period, the link cumulative arrival/departure curves (hence time-dependent link/path travel times) corresponding to a given set of temporal O-D demands.

DNL is an underlying component of many dynamic network problems in which paths costs depend on temporal path flows in ways governed by traffic flow dynamics. In the past two decades, DNL has attracted a great deal of attention from transportation researchers, stimulated by the need of both simulating urban traffic and solving dynamic traffic assignment problems. According to how they model traffic flow dynamics, existing DNL processes may be classified into three groups: macroscopic, microscopic and mesoscopic processes. A macroscopic DNL process employs macroscopic traffic flow models to describe traffic dynamics, while a microscopic DNL process uses microscopic traffic models, such as car-following or particle-hopping models, to describe traffic dynamics. A mesoscopic DNL process falls in-between a macroscopic and a microscopic DNL process in the sense that it uses macroscopic models to describe traffic flow but keeps track of individual vehicular quanta like a microscopic DNL process does. A vehicular quanta is an indivisible flow element which is tracked in DNL like a vehicle in microscopic simulation. However the size of the vehicular quanta can be set arbitrarily small to replicate analytical results as closely as desired.

In this research, we make use of a polymorphic, mesoscopic DNL process (PDNL) developed over the years at UC Davis (e.g., W. L. Jin 2003, X. J. Nie 2003, Yu Nie 2006). Since the freeway network we consider in this research has a special structure, the general PDNL process is considerably simplified to gain computational efficiency.
In our PDNL process, we use the hydrodynamic traffic flow model of Lighthill and Whitham (1955) and Richards (1956), known as the LWR model, with the following speed-density (s-k) relationship:

$$s = \begin{cases} s_f & k \le k_c \\ \alpha s_f(\frac{1}{k} - \frac{1}{k_j}) & k > k_c \end{cases} \tag{2.10}$$

where k_c is called critical density, k_j the jam density and s_f free-flow travel speed. Here $\frac{1}{\alpha} = \frac{1}{k_c} - \frac{1}{k_j}$. This $q-k$ curve was employed in Newell (1993) for streamlining a graphical LWR solution and then adopted by Daganzo (1994) in his cell transmission model (CTM).

Our PDNL process models the on-ramp merges and off-ramp diverges using the supply-demand method of Daganzo (1994,1995). The demand of a link, D, is the maximum possible exit flow rate that wish to leave it

$$D = \min\{C, Q\}$$

And the supply of a link, S, is the maximum possible flow rate that the link can accommodate

$$S = \min\{C, R\}$$

where C is the flow capacity depending on road characteristics and/or control strategies;

Q is the rate of the flow that is ready to exit;

R is the maximum entry flow rate to the link permitted by the current traffic condition.

With the introduction of supply and demand functions, one can determine the freeway and ramp flows on links 1-3 and 2-3 (see see Figure 2.2),at time t, denoted as $v_{13}(t)$ and $v_{23}(t)$ from the following maximization problem (the time index *t* is dropped for brevity):

$$\max v = \sum_i v_{i3}, \text{ subject to } 0 \le v_{i3} \le D_i, i = 1, 2, \sum_i v_{i3} \le S_3$$

where

D_i is the demand of link $i, i = 1, 2$.

S_3 is the supply of link 3, the downstream link.

This program, however, does not have a unique solution when $D_1 + D_2 < S_3$.
Jin and Zhang (2003) proposed an alternative distribution scheme

$$a_{i3} = \frac{D_i}{\sum_i D_i}$$

which yields the following simple solution to the ramp merge problem:

$$v_{i3} = a_{i3} v, \forall i, v = \min\{D_1 + D_2, S_3\}$$

Similarly, for a ramp diverge with freeway and off-ramp links 1-2 and 1-3 (see Figure 2.2), diverging flows v_{12} and v_{13} can be determined by the following maximization problem (Daganzo 1995):

$$\max v = v_{12} + v_{13}, \text{ subject to } a_{12} v \le S_2, a_{13} v \le S_3, v \le D_1$$

where a_{1i} is called turning proportion. In the simplest case (e.g., for evacuation applications), turning proportions a_{1i} can be determined exogenously. However, in the general context of network loading, a_{1i} are dependent on traffic composition, and vary with time and the demand pattern. Thus, turning proportions have to be derived from the destinations of the vehicles ready to advance into each diverging branch at any given time. The solution to the above mathematical program is simply:

$$v_{12} = a_{12} v, v_{13} = a_{13} v, v = \min\{D_1, \frac{S_2}{a_{12}}, \frac{S_3}{a_{13}}\}$$

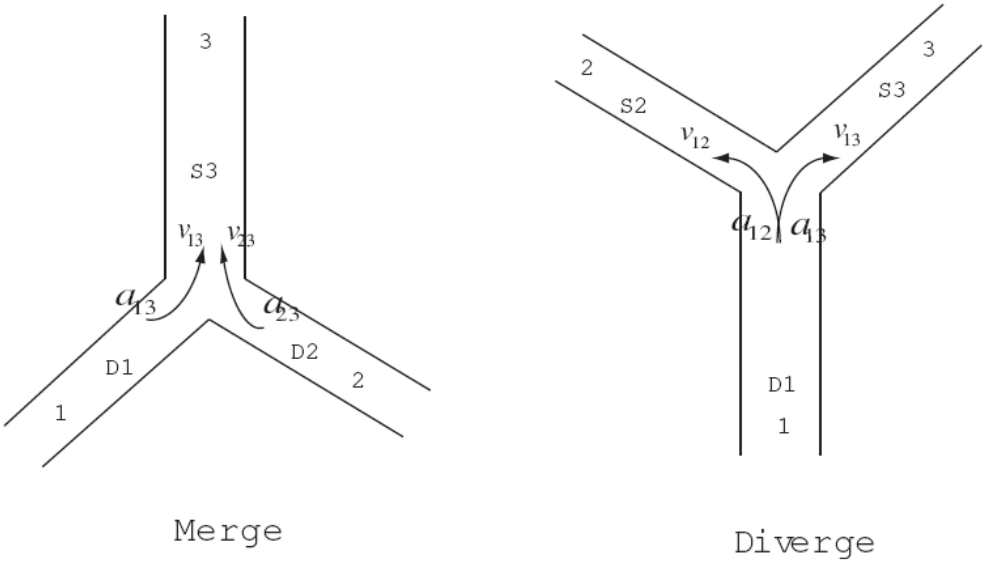


Figure 2.2 Merge and diverge nodes

By performing DNL, we can get a set of cumulative arrival and departure flow curves for every link in the loading horizon. Let us define $B_a(t)$ as the cumulative flows to pass the entrance of link a by time t, i.e.,

$$B_a(t) = \int_0^t u_a(t), t \in [0, T']$$

where $u_a(t)$ is the rate of flow entering link a at time t. Similarly, $E_a(t)$ denotes the cumulative flows to pass the exit of link a by time t, i.e.,

$$E_a(t) = \int_0^t v_a(t), t \in [0, T']$$

where $v_a(t)$ is the rate of flow leaving link a at time t. Figure 2.3 gives an example of cumulative arrival and departure curves.

These cumulative curves are very useful because most quantities of interest to the description of traffic flow can be derived from them. The traversal time that a vehicle would experience if it enters the link at time t is the horizontal separation between the two curves (note that the horizontal line crosses $B_a(t)$ at time t). Mathematically, $\tau_a(t)$ can be computed as

$$\tau_a(t) = \arg\min_{\tau \geq 0} \{B_a(t) \leq E_a(t+\tau)\} \tag{2.11}$$

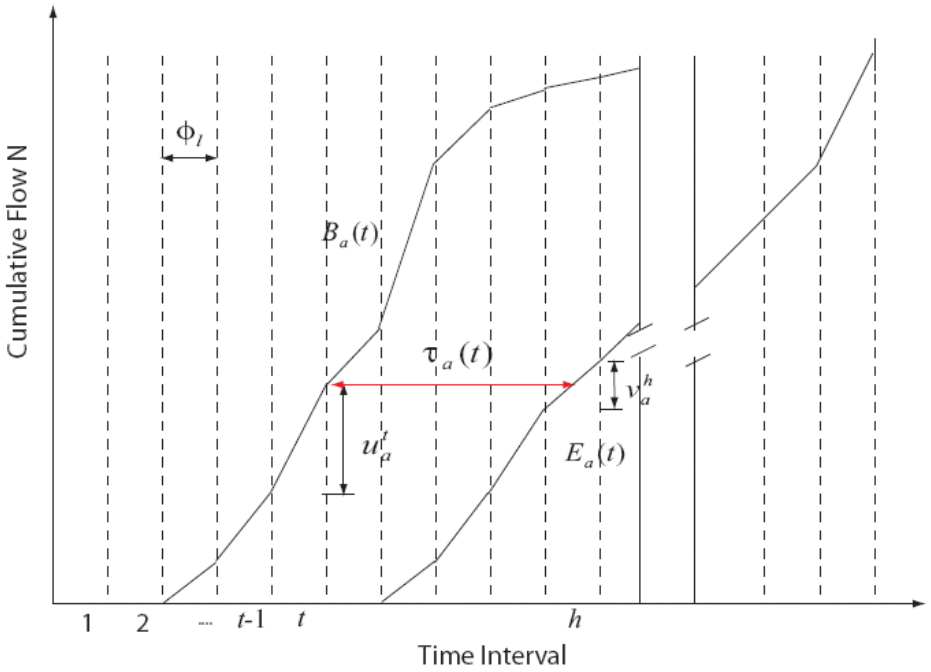


Figure 2.3 Cumulative curves constructed from DNL

We emphasize that (2.11) holds only when vehicles do not pass each other when traversing the link, known as the First-In-First-Out (FIFO) rule. To see this, note that the formula (2.11) is valid only if any vehicle will not exit the link until all vehicles present on the link at its entry time have left. Put it in another way, no vehicle can leave the link earlier than any vehicle which enters the link before it.

2.2.3 Algorithms for Solving Variational Inequality Problems

The DoDE problem cast as the VI problem (2-8) can be solved by a number of algorithms. Here we present two such algorithms, the basic project algorithm and the heuristic solution algorithm known as the method of successive averages (MSA). Before we proceed to describe these algorithms, we simplify the notation by using

$$f \equiv \mathbf{f}, W \equiv R_+^{om_a}, c(f) \equiv -\mathbf{d}(\mathbf{f})$$

and accordingly denote the VI problem by VI(c, W).

Let us define the merit (or gap) function of VI(c, W) as follows:

$$\rho(f) := \min_{g \in W} \langle c(f), g - f \rangle = \max_{g \in W} \langle c(f), f - g \rangle$$

f^* is a solution to VI(c, W) if and only if f^* solves the maximization program

1) Basic Projection Algorithm

Let r be a constant positive scalar, the main iteration of this algorithm centers around computing the following projection:

$$f^{k+1} = \Pi_W (f^k - \frac{1}{r} c(f^k))$$

This algorithm has been applied for solving the DTA problem in Wu et al (1998), where it was shown that the projection mapping $\Pi_W (f - \frac{1}{r} c(f))$ has to be a contraction to ensure the convergence of the above algorithm.

2) The Successive Averages Algorithm

In the method of successive averages, the main iteration of the algorithm concerns the following flow update:

$$f^{k+1} = (1-\lambda) f^k + \lambda g^*(f^k)$$

where $\lambda = \frac{1}{k}$, and $g_i^*(f) = \begin{bmatrix} -\kappa c_i(f) & c_i(f) < 0 \\ 0 & c_i(f) \geq 0 \end{bmatrix}, \kappa \geq 1$

In the next section, we will employ these algorithms to solve an example DoDE problem for a synthetic freeway network with eight O-D pairs, and compare the estimation results under different levels of traffic information.

2.3 Numerical Results

2.3.1 The Freeway Network and Related Data

The freeway network used in this study, which contains a northbound portion of Freeway SR-41 in Fresno, CA (from N. Fresno St. to S. Golden State Blvd, 16.7 miles), is shown in Figure 2.4. The network used in our experiment is a trimmed version which includes only the on/off ramps and mainline links of the freeway network (as illustrated in Figure 1-1). The trimmed network consists of eight interchanges, 12 off-ramps, 17 on-ramps, 31 freeway mainline links and 116 O-D pairs. There are three main bottlenecks in this network due to lane drops. The assignment horizon is a two-hour peak period, and the assignment, measurement and loading intervals are set to 15 minutes, 15 minutes and 6 seconds, respectively.

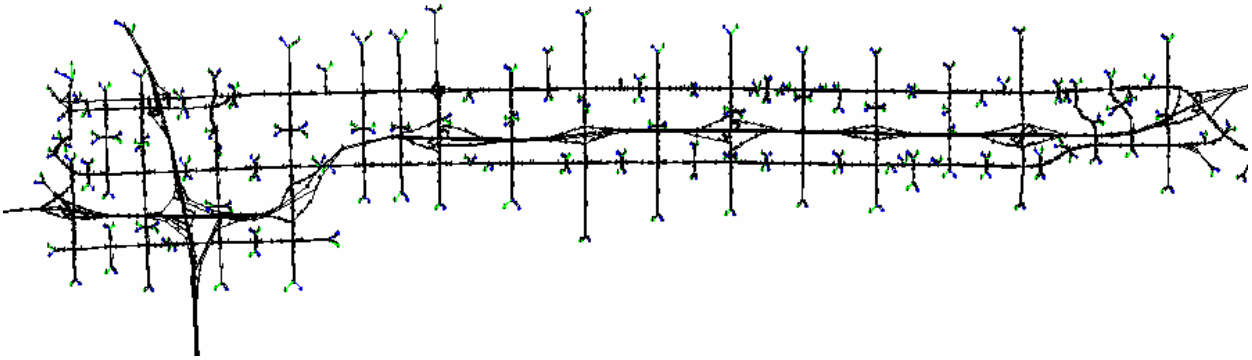


Figure 2.4 A real freeway network

The synthetic time-dependent O-D tables and traffic measurements in our experiments

are produced using the following procedure. First, real traffic counts provided by Caltrans District 6 are used to estimate an initial time-dependent O-D table. Second, total travel demands for the whole assignment horizon are obtained for each O-D pair by summing all time-dependent entries of the table. Third, various rules are applied to allocate the total demands to different assignment intervals following a flat, trapezoidal or a two-peak pattern (see Figure 2.5 and Table 2.1), and the results are used as the ground truth for time-dependent O-D demands (the "True" O-D table). Fourth, a dynamic network loading (DNL) based on the kinematic wave theory is performed to obtain traffic measurements including traffic counts and path travel times (note that traffic assignment equals network loading in the freeway case since no route choice is involved). Finally, the synthetic O-D table is uniformly perturbed by 20% (i.e., each entry of the synthetic OD table times 1.2) to generate a synthetic time-dependent historical O-D trip table. The total demands in the assignment horizon are used as planning (static) demands.

Since we have synthetic O-D demands which are assumed to be the true O-D flow in this case, we may obtain synthetic traffic counts for each link and path travel time for each OD pair by performing PDNL with the synthetic OD, as described above. Regarding the historical OD matrix, we uniformly perturbed the synthetic O-D by 20% (i.e synthetic OD times 1.2 uniformly). Moreover, we set the constant total O-D demands in two hours as the planning (static) O-D. There are all kinds of information we can have to conduct the O-D estimation, but sometime only some of them are available to us, or for example, we may only obtain the traffic counts information for very limited links (main links or ramps). Therefore, we also would like to see the accuracy of demand estimation given by partial information based on combinations of different information. Now we are able to create different scenarios to check which kind of information is more efficient in predicting O-D demands.

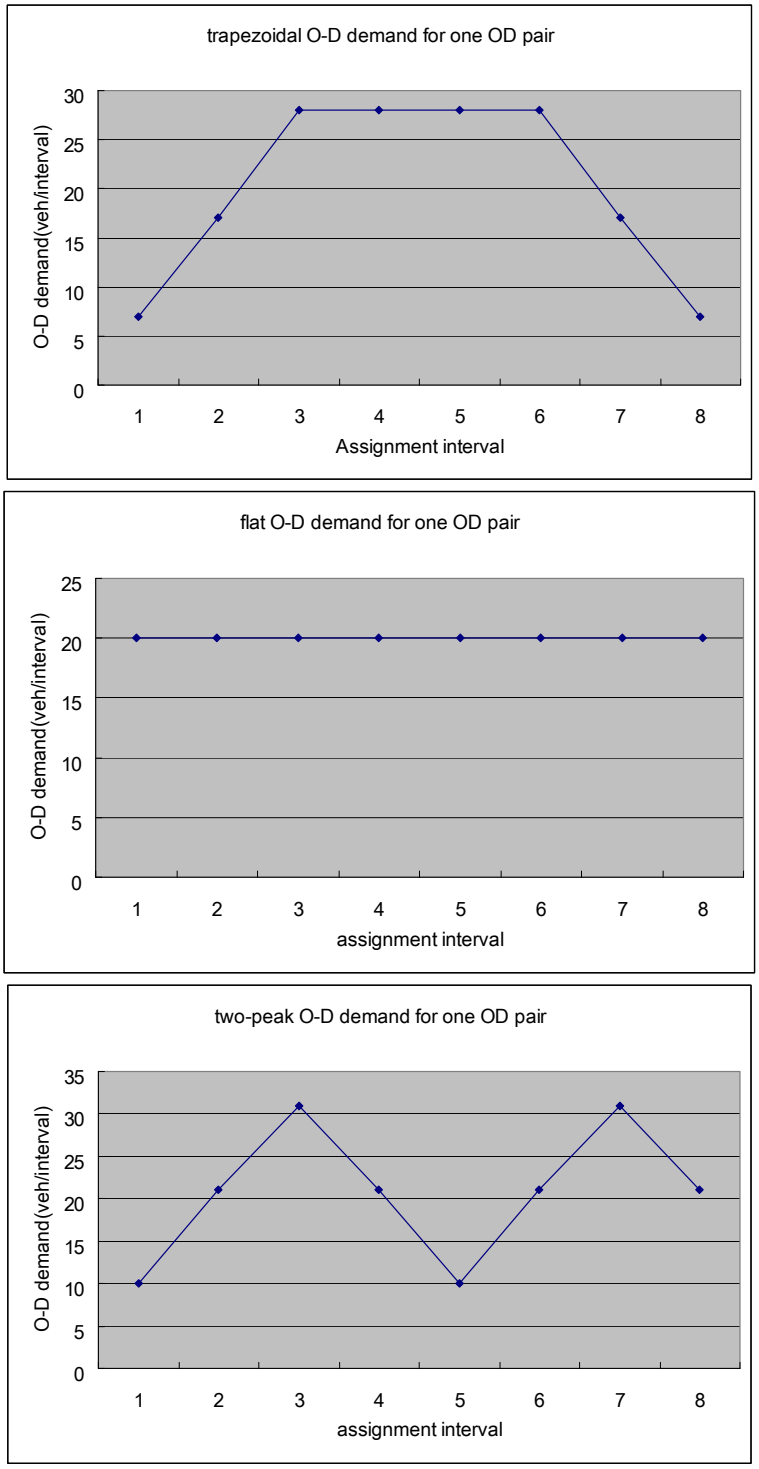


Figure 2.5 The three types of distributions of demand over time

Table 2.1 Allocation rules for three types of distributions of demand

Two-peak pattern

Interval	1	2	3	4	5	6	7	8
Proportion.	1/16	1/8	3/16	1/8	1/16	1/8	3/16	1/8

Flat pattern

Interval	1	2	3	4	5	6	7	8
Proportion.	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Trapezoidal one-peak pattern

Interval	1	2	3	4	5	6	7	8
Proportion.	1/23	5/46	4/23	4/23	4/23	4/23	5/46	1/23

The following three indices are used to measure the quality of the O-D estimation:
1) GEH: a metric proposed by British engineers to measure the quality of the estimates (Zhang et.al 2006).

$$GEH = \frac{\sum_{i=1}^{N} \sqrt{\frac{\left(V_{pi} - V_{ti}\right)^2}{\left(V_{pi} + V_{ti}\right)/2}}}{N}$$

where *Vpi* = the *i-th* value predicted by the model and *Vti* = the *i-th* field measurements. *N* is the total number of observations. A perfect match will result in a zero GEH value (Zhang et.al 2006).

2) Mean Error (ME):

$$ME = \frac{\sum_{i=1}^{N} \frac{\left|V_{pi} - V_{ti}\right|}{V_{ti}}}{N} \times 100\%$$

3) RMSE: the root mean square error.

$$RSME = \sqrt{\frac{\sum_{i=1}^{N} \left(V_{pi} - V_{ti}\right)^2}{N}}$$

Sixty scenarios are created to compare the estimation quality under various data coverage schemes. For each demand pattern (trapezoidal or one-peak, flat and two-peak), we consider eight different data coverage schemes:
1) Traffic counts
2) Traffic counts and partial path travel times
3) Traffic counts and path travel times
4) Traffic counts and historical O-D
5) Traffic counts, historical O-D and partial path travel times[1]
6) Traffic counts, historical O-D and path travel times
7) Traffic counts and planning (static) O-D
8) Traffic counts, planning (static) O-D and path travel times

As for traffic counts, to see how many links of counts are appropriate for DoDE in

[1] *Partial* means that path travel times are available only on some assignment intervals and/or O-D pairs.

terms of accuracy, we also tried to give three types of traffic counts

1) All traffic counts information (including all counts on mainlines and ramps)
2) Mainline traffic counts information
3) Random traffic counts information (where 30 out of 60 mainline links and ramps are randomly picked from a uniform distribution).

In all the experiments, BPA is terminated when the number of main iterations exceeds 50 or the standard deviation of estimated O-D tables in two consecutive iterations is less than 0.05, whichever comes first.

2.3.2 Testing Scenarios and results

Among the three demand patterns, the constant (flat) demand pattern produces no congestion in the network, so its flow characteristics are more predictable. Both the trapezoidal and the two-peak patterns cause traffic congestion at three disjoint bottlenecks. The congestion lasts half an hour in the two-peak pattern and one hour in the one-peak pattern, and never spreads to other sections beyond the bottleneck sections.

Table 2-2 reports the ME, GEH and RMSE statistics obtained from the DoDE estimation results in all 60 scenarios. Since the three statistical measures show consistent results, i.e., when one measure is lower in one case than in another case, the other measures also share the same trend, when will use only the ME statistics in our subsequent discussions.

The role of traffic counts

Among the various forms of traffic information, traffic counts are the most commonly available and hence provide the basic inputs to O-D estimation. The questions are: are they sufficient to obtain good O-D estimates, and how many counting locations are needed? Our experiments indicate that even under a full set of traffic counts, the mean error obtained still ranges from 22% to 53%, although the traffic counts themselves are closely reproduced by the model for all the three demand patterns. Reducing the amount of counting locations, on the other hand, would lead to even poorer results. For example, the ME for the one-peak demand pattern increases from 49% to 64% when about half of the links are counted. In either case, just traffic count alone seems inadequate to provide a reliable estimate of a time-dependent O-D trip table, because the mapping between time-dependent O-D demand and the observed link traffic counts is not one-to-one. In our experiments, the number of counting locations is reduced by half, and in one case, the 30 counting locations are placed on all freeway mainline links, and in anther case, they are randomly placed on either freeway or amp links. It seems that when there is sufficient number of counting locations, where to place them is not a critical issue. But this may change when the number of counting locations is much fewer, and is worth further investigation.

Table 2.2 The results of all the 60 scenarios

Scenarios			Trapezoidal demand			Flat demand			Two-peak demand		
			ME	GEH	RMSE	ME	GEH	RMSE	ME	GEH	RMSE
All counts	No PT	---	0.492	1.602	9.388	0.217	0.874	3.994	0.273	1.025	4.682
		His.	0.135	0.515	3.021	0.032	0.128	0.660	0.069	0.273	1.887
		SP	0.429	1.201	6.771	0.034	0.139	1.047	0.174	0.632	3.553
	PPT	---	0.503	1.689	9.741	0.217	0.874	3.994	0.289	1.031	4.721
		His.	0.136	0.518	3.037	0.032	0.128	0.660	0.070	0.278	1.902
	PT	---	0.525	1.737	10.497	0.217	0.874	3.994	0.288	1.037	4.725
		His.	0.139	0.526	3.041	0.032	0.128	0.660	0.070	0.288	1.933
		SP	0.428	1.211	6.078	0.034	0.139	1.047	0.176	0.632	3.561
Mainline counts.	No PT	---	0.645	1.815	11.491	0.263	1.092	4.933	0.335	1.366	6.612
		His.	0.153	0.671	3.608	0.035	0.139	0.708	0.075	0.313	2.064
	PPT	---	0.679	1.908	12.022	0.263	1.092	4.933	0.363	1.477	6.982
		His.	0.156	0.674	3.642	0.035	0.139	0.708	0.075	0.314	2.062
	PT	---	0.681	1.935	12.178	0.263	1.092	4.933	0.392	1.582	7.676
		His.	0.157	0.678	3.671	0.035	0.139	0.708	0.076	0.315	2.062
Random counts	No PT	---	0.703	2.243	15.731	0.276	1.168	5.536	0.343	1.439	7.356
		His.	0.175	0.815	3.815	0.053	0.225	1.228	0.084	0.332	2.347
	PPT	---	0.726	2.381	15.835	0.276	1.168	5.536	0.352	1.583	7.639
		His.	0.176	0.822	3.845	0.053	0.225	1.228	0.086	0.351	2.462
	PT	---	0.739	2.474	16.705	0.276	1.168	5.536	0.364	1.622	7.877
		His.	0.179	0.831	3.860	0.053	0.225	1.228	0.086	0.364	2.540

"His" Stands for historical O-D, "PT" stands for path travel time, "PPT" stands for partial path travel time and "SP" stands for static planning O-D

The role of historical O-D information

In all cases where historical O-D trip tables are used, the mean errors are reduced dramatically. Note that in our experiments the historical O-D demands always retain the "shape" (or profile) of the underlying travel demand pattern because they are generated from a uniform perturbation to the synthetic O-D that we are trying to estimate. The remarkable improvement we obtained with the addition of historical O-D should be largely credited to availability of this structural information. Including static O-D demands also improved marginally the estimates for both the one- and two-peak demand patterns, but the improvement is more sizable to the flat demand pattern. This is somewhat expected because a static O-D trip table contains the "shape" of the flat-demand pattern, but not the one- or two-peak patterns. These results highlight the importance of the knowledge of the profile of the demand, not the volume of demand in O-D estimation.

Although historical OD information that contains the profile of the O-D demands to be estimated is very useful in guiding the estimation process to find the right demand profile, historical O-D demands themselves may not be relied on at the later stages of the estimation because they can be far away from the actual O-D demands. Therefore forcing the demand deviation smaller actually could lead to larger estimation errors.

Figure 2.6 shows the change of ME, GEH and RMSE of counts, path travel time and O-D demand over all the iterations in the one-peak pattern where all three kinds of data are provided. Obviously, the estimation starts to move in the wrong direction after 15 iterations when the deviation of historical OD is comparatively large. We can eliminate this problem by using historical information in the beginning part of the estimation process to shape the demand profile, then discard it and proceed with other information such as traffic counts and/or path travel times. Figure 2.7 shows an example where the above procedure is carried out, and one can see that the values of ME, GEH and RMSE are nearly monotonically decreasing without the obvious upturn shown in Figure 2.6.

The role of path travel times

As expected, the use of travel times did not improve the quality of O-D estimates in the flat demand pattern, since there is no congestion in the network in that scenario. Surprisingly, our experiments showed that the use of travel times in the other two cases where there is congestion in the network also did not improve the O-D estimates. This is somewhat unexpected because unlike traffic counts, travel times can reveal more about traffic conditions on the network. Upon a more careful examination, however, an explanation can be found. Intuitively, if two O-D pairs go through the same bottleneck, it makes no difference to the queuing time (hence the path travel time) whether the additional vehicle that joins the queue is from one O-D pair or the other. That is, the mapping between O-D path flow and path travel time is also not one-to-one. Thus the use of path travel times will not eliminate the under-determined condition in the O-D estimation problem.

What complicates the problem even more is that under in the dynamic context, traffic counts and travel times are intricately related through the evolution of time, so the two pieces of information often overlap each other. In fact, if there are several bottlenecks in the network as in our case, the increase of one more unit of travel time on one path will affect all the travel times of O-D flows that go through the bottleneck that caused the one unit travel time increase. Therefore, it is difficult to identify the O-D pair with that additional vehicle that caused the travel time increase that affected the other O-D pairs.

It should also be noted that bringing in path travel time poses difficulties to the solution algorithm as well, since the mapping $c(\mathbf{f})$ is highly nonlinear and may not satisfy the monotone properties required by the basic projection algorithm. Moreover, the solution to the DoDE problem with the path travel time deviations considered may not have a unique solution.

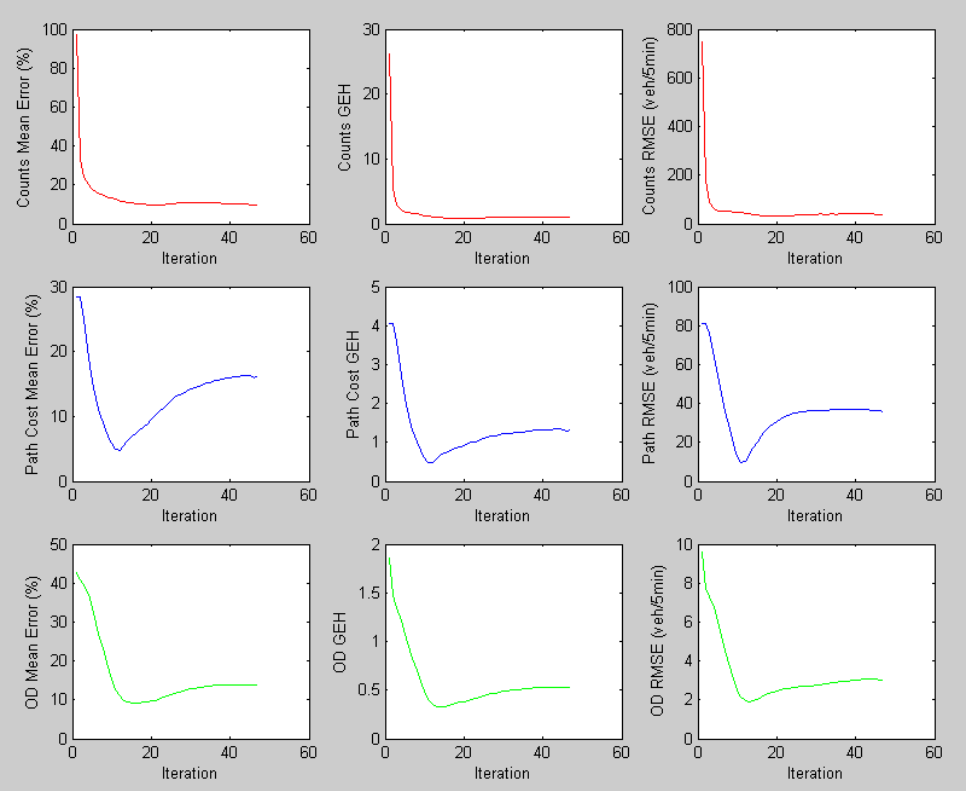


Figure 2.6 ME, GEH and RMSE of counts, path travel time and O-D demand for an original scenario

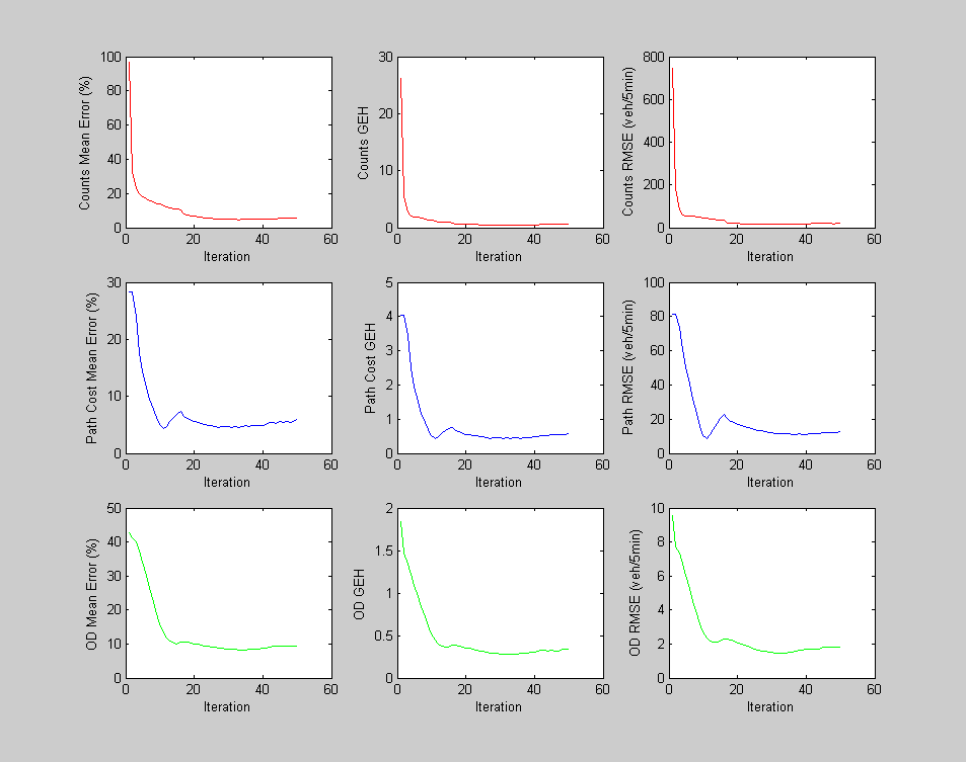


Figure 2.7 ME, GEH and RMSE of counts, path travel time and O-D demand, historical OD data discarded since the 16th iteration

Other findings

Our numerical experiments also revealed some interesting algorithmic convergence patterns. When only traffic counts are used, the ME, GEH and RMSE statistics for travel demand all converge monotonically, with a sharp drop in the 10~15 iterations and a long, flat tail after that. For all other cases, their convergence patterns are not monotonically decreasing: smallest ME, GEH and RMSE of O-D demands are always obtained around 12~15th iteration, but then the ME, GEH and RMSE starts to rise before the pattern gets flat again and the algorithm stops when the maximum number of iteration is reached. Take the case of the one-peak demand pattern with traffic counts, historical O-D and path travel times used in the estimation, the lowest ME (0.0915), GEH (0.33) and RMSE (2.05) values are reached at the 15th iteration, but rise to 0.138533 (ME), 0.525811 (GEH) and 3.04085 (RMSE) when the projection algorithm stops at the 47th iteration. The possible reasons for these different convergence patterns are threefold: 1) when only traffic counts are used, the DoDE problem is a quadratic optimization problem, hence a unique solution can be found through the basic projection algorithm, thus we have monotone convergence; 2) when both traffic counts and historical O-D demands are used, the DoDE is still a quadratic problem, but the historical O-D can be quite inaccurate and forcing the estimated O-D demands to approach the historical O-D demands by the projection algorithm would lead the solution away from the underlying O-D demand pattern in later iterations, when the O-D deviation is comparatively larger than the count deviation; and 3) when travel times are used, the DoDE problem becomes highly nonlinear, and the mapping $c(\mathbf{f})$ may not be monotone, therefore the basic projection algorithm may not converge or may converge to a local solution.

We also found that the quality of O-D estimates vary considerably across O-D pairs. For some O-D pairs, good estimates can be obtained regardless of the types of data used. For others, however, accurate estimates can only be obtained when multiple sources of data are used. Clearly, different O-D pairs have different dependence on data, but their temporal profile can be captured in most cases. As an evidence of this, Figure 2.8 shows the estimated O-D demands for four OD pairs (randomly picked in all the 116 OD pairs) for the one-peak demand pattern. As we can see from this figure, even when the magnitudes of demands are not estimated well, the temporal profiles of the O-D demands are closely followed by the estimates.

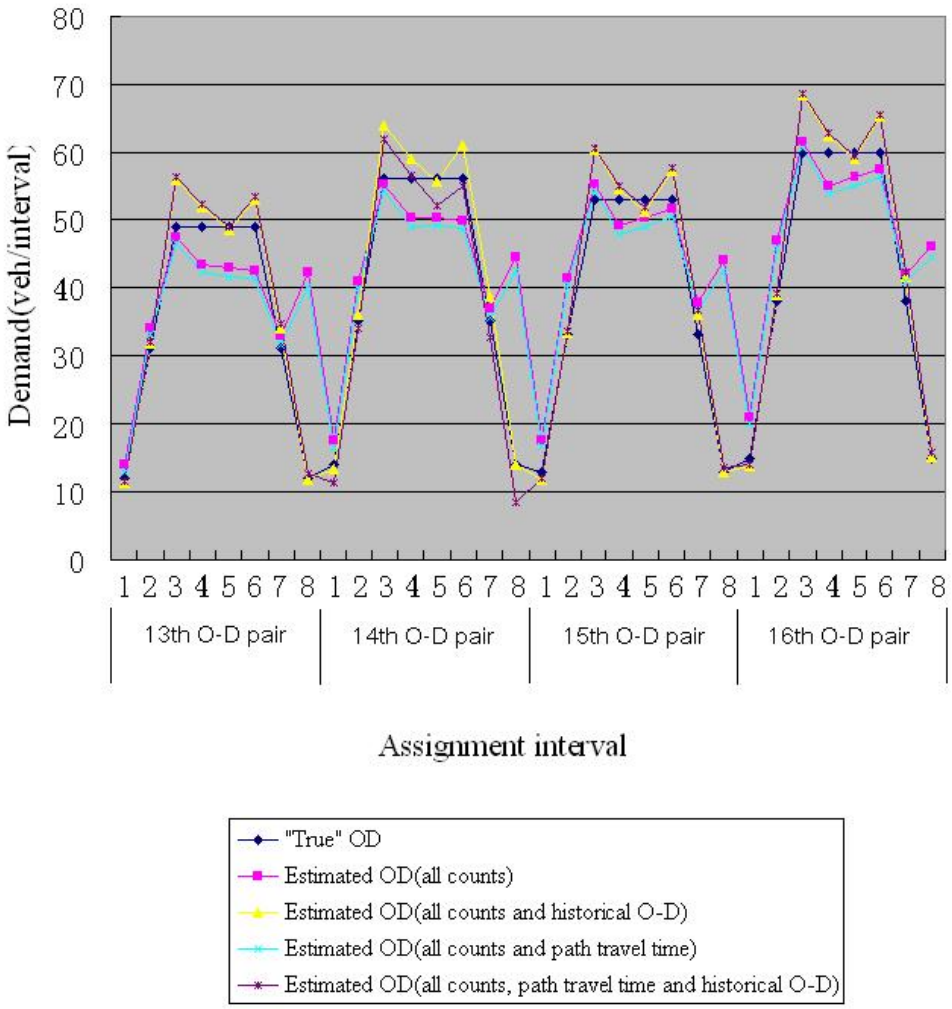


Figure 2.8 Estimated O-D demands for four OD pairs for trapezoidal pattern

2.4 Summary

We have proposed a variational inequality approach to estimate time-dependent O-D demands for a freeway network with entry and exit ramps. This approach takes into account various levels of traffic information, such as link flow count, historical O-D tables, static planning O-D and observed path travel times. The DoDE problem presented in this report exclusively targets off-line applications, i.e., estimating the temporal demand pattern that is relatively stable from day to day. Within day fluctuations can be handled through a Kalman filter process or a rolling horizon strategy as described in (Kang 1999) and (Zhou 2004).

From our computational experiments, some conclusions for data coverage can be drawn. First, traffic counts are indispensable to O-D estimation, and the more the better. But when the number of count locations is sufficiently large (more than half of the total number of links in the network in our case), it appears that counting locations does not matter. When the number of counting locations is small, however, where to place those counters may be vital and is worth of further investigation. Second, traffic counts alone are usually not sufficient to obtain accurate O-D demand estimates. Historical

time-dependent O-D demands, particularly those that reveal the temporal demand profiles of the underlying demand patterns, can drastically improve the quality of the demand estimates. When a planning (static) demand matrix is provided, one can apply a temporal profile to distribute the total demand into several time intervals, and use it as the historical O-D matrix, and finally, counter to our expectations, path travel times do not contribute much to improving O-D demand estimates. Moreover, due to the complex relation between path flow and path travel time, the use of path travel times also destroys the nice quadratic structure of the DoDE problem and brings convergence difficulties to the basic projection algorithm. Therefore, their use in dynamic O-D estimation is not recommended. We suspect that in the dynamic setting, traffic counts and their temporal distributions capture much more of the network conditions than in the static setting, rendering travel times less needed in the dynamic O-D estimation problem.

Our further work will be directed at how to place the traffic counters when their number is small compared with the total number of links in a network, and how to improve the computational efficiency of the DoDE solution procedure employed in this study.

References

Daganzo, C.F.(1994), 'The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory', Transportation Research 28B, 269-287.

Daganzo, C.F. (1995), 'The cell transmission model, part II: Network traffic', Transportation research 29B, 79-93

Jin, W.L. & Zhang, H.M. (2003), 'On the distribution schemes in the discrete kinematic wave model of merges', Transportation Research 37B, 521-540

Jin, W. L. (2003). *Kinematic Wave Models of Network Vehicular Traffic*. PhD thesis, University of California, Davis.

Kang, Y. (1999), 'Estimation and Prediction of Dynamic Origin-Destination (O-D) Demand and System Consistency Control for Real-Time Dynamic Traffic Assignment Operation', PhD thesis, The University of Texas at Austin.

Lighthill, M. J. & Whitham, J. B. (1955), "On kinematic waves. II. a theory of traffic flow on long crowded roads," Proceedings of the Roy Society A 229, 291-345.

Newell, G.F. (1993), 'A simplified theory of kinematic waves in highway traffic', Transportation Research 27B, 281-314

Nie, Xiaojian (2003). *The Study of Dynamic User-Equilibrium Traffic Assignmen.* PhD thesis, University of California, Davis.

Nie, Yu (2006). *A Variational Inequality Approach For Inferring Dynamic Origin-Destination Travel Demands.* PhD thesis, University of California, Davis.

Richards, P. I. (1956), `Shockwaves on the highway', Operation Research 4, 42-51.

Wu, J.H., Florian, M., Y. W. & Rubio, M. (1998), ‘The continuous dynamic network loading problem: a mathematical formulation and solution method’, Transportation Research 32B, 173-187.

Zhang, H.M Zhang, Ma, Jingtao, and Dong, Hu (2006), Calibration of Departure Time and Route Choice Parameters in Micro Simulation with Macro Measurements and Genetic Algorithm, Transportation Research Board Annual Meeting 2006 Paper #06-2376

Zhou, X. (2004), ‘Dynamic Origin-Destination Demand Estimation and Prediction for Off-Line and On-Line Dynamic Traffic Assignment Operation’, PhD thesis, University of Maryland, College Park.

CHAPTER 3

ESTIMATING TIME-DEPENDENT O-D MATRICES FOR ARTERIALS

Prepared by:

Yingyan Lou and Yafeng Yin

University of Florida

3.1 Literature Review

3.1.1 Background

Origin-Destination (O-D) matrices provide information on flows of vehicles traveling from one specific geographical area to another, and are one of the critical data inputs to transportation planning, design and operations. Because it is very time consuming and labor intensive to obtain O-D matrices through household interviews or roadside surveys, significant efforts have been made for decades to develop mathematical models for estimating the matrices from link counts, which are relatively easier to obtain. So far, up-to-date commercial planning tools (e.g. EMME/2) and simulation software (e.g., Paramics) have provided built-in O-D estimation modules. However, most of these O-D estimators are only capable of estimating *static* O-D matrices rather than *dynamic* or *time-dependent* O-D matrices. The latter are pre-requisites for short-term planning applications, traffic impact analyses, and operations studies. For example, as a new generation of planning tools, DynaMIT-P and DYNASMART-P overcome the limitations of static models by capturing the dynamics of congestion formation and dissipation associated with traffic peak periods. This enables the evaluation of a wide array of congestion relief measures, which could include both supply-side and demand-oriented measures (FHWA, 2004). To apply these tools, time-dependent O-D matrices should be supplied as inputs. As another example, in freeway corridor management, time-of-day ramp metering algorithms require O-D flow fractions, and adaptive ramp control strategies often need to know dynamic O-D flow in order to distribute expected flow reductions from a bottleneck to various metered ramps upstream.

Though the estimation of static O-D matrices is well researched, dynamic O-D estimation is a much more recent topic. For the static estimation, there are extensive literature concerning various formulations and solution methods, including information minimization, entropy maximization, maximum likelihood, Bayesian inference, and generalized least squares for networks without congestion, and bi-level programming for networks with congestion (see the report by Chen et al., 2004 for a recent survey of this topic). The static approach requires all trips started in the modeled period to be completed in the same period. This assumption may be appropriate for a long time horizon, but certainly not realistic at all for a very short time at a general network, say five to ten minutes with a long stretch of a highway corridor. In view of that real-time information of O-D flows or the O-D matrix for each short interval is an essential input for short-term planning applications and real-time traffic operations and management, especially in the context of intelligent transportation systems (ITS), a variety of models have been proposed to use time-series of link flow data to derive dynamic O-D matrices over the last two decades.

This section reviews the existing models for dynamic O-D estimation and then attempts to identify the issues that may need further research. The two prevailing approaches for open networks and closed networks respectively are summarized in Section 3.1.2. Since the major concern of the research project is on linear networks (a linear network is a stretch of highway with multiple entries and exits, where there would be no route choices involved) where entry and exit counts are more likely to be known (therefore closed networks), Section 3.1.3 further dwells on the closed-network-oriented approach,

elaborating the issues of dealing with constraints in the framework of recursive estimation, considering travel time and flow propagation, and incorporating multiple data sources. Section 3.1.4 summarizes the issues to be further addressed.

3.1.2 Existing Approaches

Generally speaking, dynamic O-D estimation is to use time-series of link flow data to derive time-dependent O-D matrices. Based on the kernel measurement relationship used, the existing approaches may be categorized into two classes: closed-network-oriented approach and open-network-oriented approach[2].

3.1.2.1 Closed-Network-Oriented Approach

"Closed" networks are networks where all the entry and exit counts are known during all measurement intervals. For a general network, this actually implies that time-varying departure rates (trip production rates) of all origins and arrival rates (trip attraction rates) of all destinations are known. Consequently, the key feature of this approach is the direct estimation of O-D splits from time-series measurements of network entry, exit counts and sometimes link flows. The fundamental idea is that the traffic flowing through a transport facility is treated as a dynamic auto-regressive process in which the sequences of real-time exiting counts depend, by causal relationships, upon the sequences of real-time entering counts. In this manner, additional information can be obtained, which can be used besides the conservations between the exit and entry flows, to identify the structure and size of the flows inside the facility without using further *a priori* information (Crème and Keller, 1987). Crème and Keller (1981) shall be credited for their first application of the above idea in identifying turning flows at isolated intersections.

The flow conservation equation is the basic system equation in this approach, representing the relationship between the real-time exiting counts of a certain destination j and real-time entering counts of all the related origins i:

$$y_j(t) = \sum_i b_{ij}(t) q_i(t) + e_j(t)$$

Here $y_j(t)$ and $q_i(t)$ are the exiting and the entering traffic counts, $b_{ij}(t)$ is the O-D split, the proportion of traffic flows entering at origin i and exiting at destination j, and $e_j(t)$ is a random error.

Since there are always much more unknowns, namely the O-D splits, than the relationships established, various system identification methods should be applied to estimate the unknowns. Crème and Keller (1987) proposed four methods:

[2] There exist other terms for classifying these models, such as "non-assignment-based" versus "assignment-based approaches" by Chang and Tao (1999), and "intersection-oriented" versus "network-oriented" by Chen et al. (2004). Here we follow the terminology by Ashok and Ben-Akiva (2000). Note that there is actually no clear dividing line between these two approaches, and the classification is more to facilitate the presentation of the ideas.

cross-correlation matrices, constrained optimization, recursive estimation and Kalman filtering for solving this problem. Concurrently, Nihan and Davis (1987) developed a recursive predictions error (RPE) estimator of tracking dynamic O-D parameters. These methods can all be interpreted as recursive or non-recursive least-squares methods, because they share the same assumptions and insights that can be traced back to Gauss's least-squares estimation theory (Sorenson, 1970). Also note that although the recursive algorithms mentioned above require an initial O-D matrix to start with, the dependency on this initial input in their estimates decreases as time passes. However, when major shifts in demand patterns occur (such as from peak to non-peak), the performance of these algorithms degrades.

These models ignore travel times in the facilities, which is justifiable for intersections and very small networks. Bell (1991) extended the models by permitting the distribution of travel times to span a number of different intervals. His first method employs a concept of platoon dispersion in representing the dynamic interactions between entry and exit flows while his second method assumes freely-distributed travel time to address travel time variability. Chang and Wu (1994) used nonlinear macroscopic speed-density-volume relations to estimate travel times and introduced link-use proportions to establish a new set of flow propagation constraints.

With the assumption that the O-D patterns are auto-regressive, these closed-network-oriented models do not need a target O-D table, although they require the entry and exit counts of the network at all time points. Li and Moor (2002) attempted to address the issue of incomplete observations by estimating the O-D flows rather than the O-D splits using a generalized least squares (GLS) approach, which however involves *a priori* target O-D matrix.

Apparently limited by the requirement of all entry/exit counts, this approach is not very practical for large-scale general networks.

3.1.2.2 Open-Network-Oriented Approach

This approach is intended for being used to estimate dynamic O-D matrices for general networks. The literature on this approach is rather limited, and the most noteworthy work known to us are those of Willumsen (1984), Okutani (1987), Cascetta et al. (1993), Ashok and Ben-Akiva (1993, 2000, 2002), Madanat et al. (1996), Bell et al. (1996), Sherali and Park (2001) and Hu et al. (2001).

This approach considers the estimation of time-varying O-D matrices as the inverse problem of dynamic traffic assignment (DTA) problem. Instead of using the simple flow conservation relationship, the assignment matrix from DTA model serves as the kernel measurement relationship. Define $\vec{y}$ as the vector of the measured link counts, $\vec{f}$ as the vector of O-D flows to be estimated, $\vec{e}$ as the random error vector and A as the assignment matrix from DTA model, the system equation is as follows:

$$\vec{y}(t) = \sum_{k=t-p}^{t} A(k,t)\vec{f}(k) + \vec{e}(t)$$

The element of the assignment matrix A is the link-use proportion, which is defined as

the proportion of a particular O-D flow departing its origin during interval *k,* prior to the current interval *t* by at most *p* intervals, contributes to the flow on link *l* during interval *t*. Since the resultant system of equations is highly under-determined, previous studies have taken two different paths to resolve the problem. Note that no matter which path is employed, the main difficulty is the determination of the assignment matrix.

The first path is to formulate optimization problems, normally constrained GLS problems, with using an *a priori* O-D matrix and then to select among the infinite number of potential candidates the one that is closest to the *a priori* O-D matrix. Willumsen (1984) and Cascetta et al. (1993) formulated very similar minimization problems with slight difference in their objective functions: Willumsen (1984) used the entropy function of O-D flows while Cascetta et al. (1993) used a combination of entropy and least squares (of link flows). Their major difference, however, lies in the way of computing link use ratios. In Willumsen (1984), link use proportions are not directly computed. Rather, a traffic simulation model (CONTRAM) is used to obtain an accumulation factor based on ratios between simulated and observed link counts, and uses this factor to update O-D flows and force convergence. On the other hand, Cascetta et al. (1993) estimated the link-use proportions through dynamic network loading, which requires the knowledge of route travel times. Rather than obtaining travel times based on estimated O-D demands, Cascetta et al. (1993) presumed that historical travel times are available and uses them to perform dynamic network loading. More recently, Sherali and Park (2001) generated the time-dependent link-use proportions (but did not explicitly specify how), and formulated a constrained GLS model whose objective function has an additional total-cost-driven component in order to avoid using an *a priori* O-D matrix. A column generation approach was developed to solve their model.

The second path is to assume that traffic dynamics is auto-regressive. Both Okutani (1987) and Ashok and Ben-Akiva (1993, 2000) use Kalman Filter to update time-varying O-D data with the assumption that the assignment matrix A is known from directly-measured travel times. Distinctive from all other work, Ashok and Ben-Akiva (1993) assumed that rather than the O-D data themselves, the deviations of current O-D data from historical O-D data are auto-regressive. Madanat et al. (1996) added a control equation to the model of Ashok and Ben-Akiva that accounts for the time-varying effects of traffic information on travel demands via a binary choice on route switching. This binary route switching decision is either to exit through the original destination, or to exit through an off-ramp before reaching the original destination. Hu et al. (2001) proposed an adaptive Kalman filtering that uses time-varying assignment matrices generated by DYNASMART. Recently, Ashok and Ben-Akiva (2002) further revised their model by using stochastic link-use proportions to address the uncertainty associated with the proportions. They also suggested an iterative O-D estimation as an alternative when the directly-measured travel times are not available. But this process has two major defects that the convergence is not guaranteed and that the resultant O-D flows could be biased. Most recently, Bierlaire and Crittin (2004) followed the formulation of Ashok and Ben-Akiva (2000) but suggested using the LSQR algorithm (first presented by Paige and Saunders, 1982) instead of the Kalman filtering to solve this sparse linear system.

The time-dependent path flow estimator proposed in Bell et al. (1996) is quite different

from the models reviewed above in the sense that it is only quasi-dynamic and steady state conditions are assumed within each period. More specifically, trips started in one period will always be completed within the same period unless inadequate road capacities prevent them from doing so. The queued vehicles, if any, will be carried from one period to the subsequent. Except this, the propagation of traffic flow and spatial and temporal evolution of congestion are simply ignored. Regardless of such limitations, Bell's time-dependent path flow estimator is comparably efficient and applicable to large-scale general networks.

3.1.3 Closed-Network-Oriented Approach

This project is concerned with deriving time-dependent O-D matrices for linear networks. In linear networks, entry and exit counts are often known. Therefore, the closed-network-oriented approach will be readily applicable. This section further elaborates the development of this approach, which has followed three major paths addressing issues of satisfying constraints of the O-D splits, taking account of travel time and flow propagation in a network and incorporating multiple data sources to increase the system observability.

3.1.3.1 Satisfying Equality and Inequality Constraints

As aforementioned, this approach is featured with direct estimation of O-D splits from time-series entry and exit counts. The O-D split $b_{ij}(t)$ is the proportion of the traffic entering at entry i at time interval t that leaves at exit j. Therefore, the parameters $b_{ij}(t)$ are obviously bounded between zero and one. Moreover, the sum of all the O-D splits from a specific entry i at a specific interval should be equal to one. These inequality and equality constraints cause some difficulties when applying the recursive least-square estimation and Kalman filtering methods because the basic structures of these methods do not allow for constraints. Consequently, extra caution should be exercised to guarantee the satisfaction of these constraints.

For the inequality constraints, note that the requirement of the O-D splits less than one is redundant if the non-negativity constraints and the equality constraints are satisfied. Nihan and Davis (1987) first proposed a truncation method for the RPE estimator to guarantee the non-negativity of the O-D splits. In the recursive estimator, the solution of the O-D split at the time interval t, $b_{ij}(t)$, is equal to $b_{ij}(t-1)$ plus a correction item $m_{ij}(t)$. The truncation method essentially ensures that the absolute value of the correction item is always less than $b_{ij}(t-1)$ by multiplying the correction item with a weighting factor. It can be seen that the truncation method actually leads to a loss of "optimality" in the recursive estimation. Bell (1991) suggested a more powerful constrained recursive least squares algorithm to handle the inequality constraints. His basic idea is to derive the Karush-Kuhn-Tucker (KKT) optimality condition for the least square program at each time interval, and apply an iterative process to determine the Lagrange multipliers. At the iteration if the non-negativity of the O-D parameter is violated, an adjustment to the associated Lagrange multiplier will be made until all the constraints are met.

The equality constraints are more difficult to deal with because they are applied to each row of the O-D matrix while the closed-network-oriented approach estimates the matrix column by column, in view of the fact that the traffic flow exiting from *j*th exit is only related to the *j*th column of the O-D matrix. A normalization method and a projection method after the unconstrained estimation process have been proposed by Nihan and Davis (1987). The normalization is to simply update each row element by dividing the row sum, while the projection method is to project the result of the unconstrained estimation onto the hyperplane defined by the equality constraints. Again, these two approaches are heuristic and seem provide no guarantee for an unbiased estimation[3]. More recently, Li and Moor (1999) proposed a recursive approach based on equality-constrained optimization to address the equality constraints. Instead of estimating the O-D matrix column by column, their formulation employs all the O-D parameters and thus the dimension of the problem is increased from the number of the origins to the product of the number of origins and that of the destinations. But the advantage is that the equality constraints can be handled explicitly while solving the least square problem. For the sake of saving computation time, the proposed approach only performs one step of iteration in Bell's algorithm to correct for inequality constraints.

Recently, Simon and Chia (2002) developed a constrained Kalman filtering method that can be applied to deal with the equality constraints in the Kalman-filtering-based algorithm. Their method is essentially a projection method that can be viewed as the generalization of the project method used in Nihan and Davis (1987). They provided a rigorous proof that the projection is an unbiased state estimator for any known symmetric positive definite weighting matrix, and then further presented a weighting matrix that has the smallest estimation error covariance.

Note that in order to handle the equality constraints it is inevitable to expand the dimension of the problem so that all the O-D parameters can be estimated at the same time.

3.1.3.2 Travel Time Consideration

The temporal and spatial dispersion of traffic is of great importance to determine either the casual relationship between the entering and exiting counts or the time-dependant assignment matrix. Indeed representation of flow propagation and estimation of travel time are tightly related. Incorrect representation of flow propagation will lead to wrong relationships between O-D parameters and link flows, resulting biased estimates of these parameters.

The first generation of the closed-network-oriented models focuses on turning movement identification for intersections where travel time can be safely assumed negligible. When extending these models to networks, representation of flow propagation is the first issue to resolve. Bell (1991) proposed two methods to allow for distributions of travel times through intersection or network that span more than one

[3] Based on the work by Simon and Chia (2002), the projection method turns out to be an unbiased state estimator, not the best one though.

interval. The first method assumes a geometrically distributed travel time for each exit. In this method, the parameters to be estimated are the platoon dispersion factor together with the O-D splits. The second method considers freely-distributed travel times, but assumes vehicles from any entrance should reach to a specified exit within three time intervals. Thus, the variables to be estimated at each time interval are exactly the O-D parameters for the three intervals before the current interval. Chang and Wu (1994) used a set of non-linear macroscopic traffic relations to estimate travel times, but assumed vehicles that reach one exit during an interval come from only two consecutive time intervals for each entrance. Moreover, the proposed method represents flow propagation by introducing link-use proportions to be estimated simultaneously with the O-D parameters. Consequently, not only the dimension of the problem increases, but also the problem itself becomes nonlinear. Extended Kalman filtering was adopted by Chang and Wu to identify the nonlinear system.

3.1.4 Incorporating Multiple Data Sources

Most existing dynamic O-D estimation models make use of time-series traffic counts at entries and exits in the network and some even require a target O-D matrix to guarantee the system observability. The traffic counts are supposed to be collected from sensors such as the inductive loop detectors, and the target O-D matrix is assumed to be available from a historical O-D data or a simple survey. Wu and Chang (1996) and Chang and Tao (1996) included constraints established from dynamic screenline and cordonline flows to increase the observability of the dynamic interactions between O-D patterns and the resulting link flow distributions. In order to obtain a more reliable estimate, Chang and Tao (1999) presented an integrated model that employs the intersection turning flow data to produce an additional set of constraints in identifying path flows from a DTA model.

The advent of automatic vehicle identification (AVI) technologies would benefit dynamic O-D estimation with providing sampled complete or incomplete vehicle trajectories. For example, if there are video detection systems installed at selected intersections, explicit turning movements, including right-turn, through and left-turn will be available. Although our literature search has not found a directly-relevant research that incorporates such a data source into the aforementioned modeling framework, many researchers have investigated the possibility of taking advantage of these AVI data in various O-D estimations. For example, electronic toll collection tag can provide partial trip trajectories of vehicles equipped with a tag. Due to the fact that only a fraction of tagged vehicles can be sampled, Kwon and Varaiya (2005) developed a statistical model to derive an unbiased estimator of the O-D matrix (essentially tag reader to tag reader interchange flows) based on the method of moments. As another example, area-wide AVI systems, such as in-vehicle global positioning system and cell phone tracking, can provide partial, but complete trajectories of the vehicles. It has been proposed by several researchers that together with the market share of such AVI equipments, an off-line O-D can be obtained from these AVI data (Asakura et al. 2000; Antoniou et al. 2004; Dixon and Rilett 2005; Eisenman and List 2004). The application of this idea to the Han-Shin expressway network in Japan (See Asakura et al. 2000) provided a practical case which indicates the effectiveness of this method.

3.1.5 Issues to Be Further Addressed

From the above review of relative literatures, we identify the following issues that may need further research:

1) Improvement of the usage of AVI data. Most studies treat the AVI data as an off-line resource for the target O-D matrix. However, the real-time AVI data may serve as another information source which will provide at least the origins of the vehicles and an approximation of the link volume. Also, other traffic information such as turning movements on the selected intersection from video detection systems may be employed to improve the system observability. It would be necessary to investigate how to fuse these different sources of AVI data to maximize the accuracy of the estimation.

2) Link use proportions. Because there is no route choice in linear networks, it is relatively easier to compute more accurately link-use proportions. With an exogenous reliable source of real-time travel time information, efficient ways should be investigated to compute link-use proportions. Without such an exogenous source, an iterative procedure between O-D estimation and flow propagation tracing should be conducted to provide endogenous estimates of travel time and link use proportions. Although such an iterative technique is expected to be quite difficult, it might be still feasible for linear networks.

3) Incomplete information. For linear networks, current models require all the entry and exit counts. However, the observation information is often incomplete. For example, it is unlikely to obtain exit counts from the typical settings of loop detectors for actuated signal control systems. It is necessary to investigate how to ensure the system observability in such a context, in addition to introducing a target O-D.

4) Measurement errors. Our experience with the traffic loop detectors suggests that the loop detectors generally have only 70%-80% accuracy. Such (systematic) measurement errors can not be represented by the error term in the Kalman Filtering. Therefore, research efforts need be made to mitigate the impacts of the inaccuracy of traffic data. One possible solution is to seek a robust counterpart of the O-D estimation optimization problem that will tolerate changes in the traffic data, up to a given bound known *a priori*.

3.2 Estimation of Origin-Destination Flows for Actuation-Controlled Intersections

3.2.1 Introduction

This section addresses the real-time estimation of O-D flows (splits or turning proportions) for isolated intersections. The purpose is to use time-series data of traffic counts to derive time-independent or time-varying O-D flows. Estimating O-D flows for isolated intersections is a starting point for the network O-D estimation. A single intersection can be viewed as the smallest network system with multiple origins and destinations. Moreover, individual intersection is the key element of a large-scale network. If turning movements of all the intersections in the network are known, the traffic situation within this system can be replicated or simulated. Therefore, intersection O-D estimation problem serves as the basis for system identification, monitoring and control.

A variety of estimators have been developed for estimation of dynamic intersection O-D, requiring all the entering and exiting counts at all time points. However, even as small as an isolated intersection, the observation information is often incomplete. For example, it is unlikely to obtain exiting counts from the typical settings of loop detectors for actuated signal control systems. Li and De Moor (2002) proposed a constrained generalized least squares (GLS) method to address the issue of incomplete observations. Their method requires *a priori* target O-D matrix, and the results will heavily depend on this initial value, which might lead to a biased estimate. This chapter presents a new two-step optimization procedure for problems with complete entering counts but incomplete exiting counts, a common information pattern from actuation-controlled intersections. The formulation is still based on the notion of GLS, but makes full use of the available information, thereby simplifying the computation and partially eliminating the dependence on the prior information. To facilitate the presentation of the idea on dealing with information incompleteness, we first focus on the situation where the O-D matrix to be estimated is constant and then extend the framework to track time-varying O-D matrices by modeling the O-D splits as a random walk process.

3.2.2 Problem Statement and Notations

3.2.2.1 Problem Statement

A big segment of intersections in the U.S. are actuation-controlled. Actuated signal controllers receive calls or actuations that request service for a particular movement, typically from inductive loop detectors cut into the pavement surface. For the purpose of signal operations, the controller does not need to determine if the call is due to a single

vehicle or a large platoon of vehicles. However, the advancement and deployment of telecommunication and ITS technologies have made traffic counts and occupancies more readily available from actuated control signal systems. For example, in California, second-by-second returns of signal status and loop detector data can be obtained for all phases.

Fig. 3.1 shows a typical loop layout for the major approach to an actuation-controlled intersection. There are two set of loops, advance loops and presence loops, installed for the through-movement, and the left-turn bay is equipped with loop detectors as well. Note that for the minor approach normally only presence loops are installed.

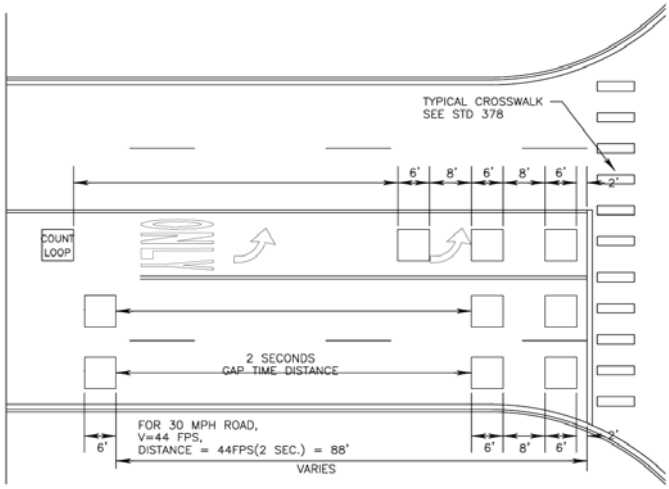


Fig. 3.1. A typical loop layout

With such a typical setting, all the entering counts can be obtained from either the advance or presence loops. However, exiting counts are generally not available. For the major approaches, we may estimate the exiting flows from advance loops for the downstream intersection with the assumption that there are no driveways in between. For the minor approach, it is less likely to even estimate the counts.

In summary, depending on the specific loop settings, there are multiple information patterns available from actuated control systems. In this chapter, we focus on a prevalent case in which all of the entering flows and exiting flows on major approaches are known while the exiting counts on minor approaches are missing.

3.2.2.2 Notations

Consistent with the notations by Nihan and Davis (1987) and Li and De Moor (2002), we let:

- J_i denote the set of exits j, which is permissible for vehicles entering at entrance i;
- I_j denote the set of entrances i, which permits vehicles to take exit j;
- OJ is defined as the index set of exits where the traffic counts are available;
- $y_j(k)$ and $q_i(k)$ denote the exiting counts of exit j and arriving counts of entrance i respectively, during time interval k;

- $\vec{y} = \begin{bmatrix} y_1 & \ldots & y_j & \ldots \end{bmatrix}^T, j \in OJ$
- $\vec{q} = \begin{bmatrix} q_1 & \ldots & q_4 \end{bmatrix}^T$
- $Q = diag(\vec{q}^T, \vec{q}^T, \vec{q}^T, \vec{q}^T)$
- b_{ij} denotes the O-D split, i.e. the proportion of traffic counts entering via entrance i and leaving via exit j. By definition, b_{ij} must satisfy:

$$\sum_{j \in J_i} b_{ij} = 1, \forall i \tag{3.1a}$$

$$b_{ij} \geq 0, \forall i, j \tag{3.1b}$$

- $\vec{b}_j$ is defined by the jth column of the O-D matrix;
- $\vec{b} = \begin{bmatrix} \vec{b}_1^T & \ldots & \vec{b}_4^T \end{bmatrix}^T$ is a column vector;
- $D = \begin{bmatrix} I & I & I & I \end{bmatrix}$ is a matrix with 4 rows and 16 columns facilitating the constraints (1a) to be rewritten in the vector form.

3.2.3 Conventional GLS Method

3.2.3.1 Formulation

For the exits where exiting counts are available, the measurement equation can be expressed as:

$$y_j(k) = \sum_{i \in I_j} b_{ij} q_i(k) + e_j(k), \forall j \in OJ \tag{3.2}$$

where $e_j(k)$ is the random measurement error with zero mean.

The framework of the estimation is to find O-D splits such that the mean squares of the error term are minimized. However, since some of the exiting counts are not observable, the corresponding b_{ij} become inestimable under the least square framework. To deal with the above issue, *a priori* information of the O-D splits to be estimated has to be introduced into the objective function of the GLS formulation. Essentially, the estimates of O-D splits are determined by the previous estimate O-D and then are corrected by the current observations.

At each time interval k, given the estimate of b_{ij} in the previous interval, $\hat{b}_{ij}(k-1)$, and the weighting factor ζ_{ij}, the GLS formulation can be written as:

$$\min \sum_j \sum_{i \in I_j} \zeta_{ij} \left[b_{ij} - \hat{b}_{ij}(k-1) \right]^2 + \sum_{j \in OJ} \left[y_j(k) - \sum_{i \in I_j} b_{ij} q_i(k) \right]^2 \tag{3.3}$$

subject to (3.1a) and (3.1b)

Note that the above model is proposed by Li and De Moor (2002) in a different but equivalent form.

3.2.3.2 Solution Algorithm

To handle the non-negativity constraints (3.1b) in the least squares formulation, Bell (1991) suggested a powerful iterative algorithm. The basic idea is to derive the Karush-Kuhn-Tucker optimality condition for the optimization problem at each time interval, and then apply an iterative process to determine the Lagrange multipliers. During the iteration if the non-negativity of the O-D parameter is violated, an adjustment to the associated Lagrange multiplier will be made until all the constraints are met. Li and De Moor (1999) further extended this method to deal with equality constraints (3.1a).

In this chapter, instead of following their procedure, we adopt Kalman filtering as the solution algorithm. In view of two facts that the weighting factor w_{ij} should be determined according to the covariance structure of both terms in the objective function in order to obtain the best linear unbiased estimator (Cascetta, 1984), and that the Kalman filtering represents essentially a recursive solution to the original least squares problem (Sorenson, 1970), we transform the original optimization problem (3.3) into the following equivalent constrained Kalman filtering problem.

Since $\hat{b}_{ij}(k-1)$ can serve as an available observation at each time interval *k*, we write the state-space equations as follows:

$$\begin{cases} \vec{b}(k) = \vec{b}(k-1) \\ \begin{bmatrix} \hat{\vec{b}}(k-1) \\ \vec{y}(k) \end{bmatrix} = \begin{bmatrix} I \\ Q(k) \end{bmatrix} \vec{b}(k) + \begin{bmatrix} \vec{w}(k) \\ \vec{e}(k) \end{bmatrix} \end{cases} \tag{3.4}$$

where $\vec{w}(k)$ and $\vec{e}(k)$ are the corresponding random error terms, subject to constraints (3.1a) and (3.1b):

$$D\vec{b}(k) = 1 \tag{3.5a}$$

$$\vec{b}(k) \geq 0 \tag{3.5b}$$

To apply the Kalman filtering algorithm, we can assume the covariance matrix of $\vec{e}(k)$ is known, denoted as $R(k)$ (this assumption is realistic since the measurement error of the loop detectors can be possibly estimated prior to the operation of the filter) and that $\vec{w}(k)$ and $\vec{e}(k)$ are independent. Noting that $\vec{w}(k)$ is actually the deviation of the estimate value from the true value at time interval (*k-1*), the covariance matrix of $\vec{w}(k)$ should be equal to $E[(\vec{b} - \hat{\vec{b}}(k-1))^T (\vec{b} - \hat{\vec{b}}(k-1))]$, which is denoted by $P(k-1)$ and can be estimated during the Kalman filtering procedure.

Let $Z(k)$ denote $\begin{bmatrix} \hat{\vec{b}}(k-1) \\ \vec{y}(k) \end{bmatrix}$ and $H(k)$ denote $\begin{bmatrix} I \\ Q(k) \end{bmatrix}$ and apply Kalman filtering with equality constraints using the maximum probability method (Simon and Chia, 2002) to this system, the solution can be expressed as:

$$\begin{cases} K(k) = P(k)H^T(k)\left[H(k)P(k)H^T(k) + V(k)\right]^{-1} \\ \tilde{\tilde{b}}(k) = \hat{\tilde{b}}(k-1) + K(k)\left[Z(k) - H(k)\hat{\tilde{b}}(k-1)\right] \\ \hat{\tilde{b}}(k) = \tilde{\tilde{b}}(k) - P(k)D^T\left[DP(k)D^T\right]^{-1}\left[D\tilde{\tilde{b}}(k) - 1\right] \\ P(k+1) = \left[I - K(k)H(k)\right]P(k) \end{cases} \tag{3.6}$$

where,

$$V(k) = E\left(\begin{bmatrix} \vec{w}(k) \\ \vec{e}(k) \end{bmatrix}\begin{bmatrix} \vec{w}(k) & \vec{e}(k) \end{bmatrix}\right) = \begin{bmatrix} P(k-1) & 0 \\ 0 & R(k) \end{bmatrix} \tag{3.7}$$

It should be pointed out that we discard (3.5b) here for the computation simplicity. The inequality constraint (3.5b) is very likely to be met since the O-D splits are constant and that their initial values can be well chosen to be all positive. If necessary, the truncation method proposed by Nihan and Davis (1987) can be applied to guarantee the non-negativity.

3.2.4 Improved Two-Step Method

3.2.4.1 Formulation

For the exits where the traffic counts are available, the corresponding column of the O-D matrix can be directly estimated from model (3.2), which is an unbiased estimator and its dependency on the initial input diminishes as time passes. However, model (3.4) is unbiased only if the mean value of the error term $\left[\vec{w}(k) \quad \vec{e}(k)\right]^T$ is zero. We can assume safely the measurement error $\vec{e}(k)$ has a zero mean, but whether $\vec{w}(k)$ meets this requirement depends on the choice of the initial values of the O-D splits. A poor initial value will always lead to a biased result for all b_{ij} even those could be unbiasedly estimated by using model (3.2).

Another limitation of conventional GLS model (3.4) is the computational difficulty. To obtain the recursive solution (3.6), an inversion of matrix should be conducted at each step. The dimension of the matrix depends on the size of $\vec{b}$ and $\vec{y}$. It may be acceptable for the isolated-intersection problem, but is not applicable for a large-scale network.

To improve the model in these two aspects, we propose a new two-step procedure. The procedure estimates the *j*th column of O-D matrix associated with the exit $j \in OJ$ using model (3.2), and then obtains the remaining unspecified O-D parameters using a constrained least squares model. Consequently, the first step estimator is unbiased and does not depend on the initial inputs. This two-step model is essentially applying the notion of GLS, but attempts to determine as many O-D splits as possible from the observations, different from the conventional way of using the observation as a correction to the prior O-D information. Such a decomposition scheme not only guarantees the accuracy of the first-step estimator, but also makes it possible to convert the original vector identification problem into several independent scalar problems in

both steps, which certainly improves the computation efficiency.

Given the estimate $\hat{b}_{ij}(k-1)$ in the preceding interval, the formulation at each time interval *k* is as follows:

- *Step 1*:

$$\begin{aligned} &\min\left(y_j - \sum_{i \in I_j} b_{ij} q_i \right)^2, \forall j \in OJ \\ &\text{subject to } b_{ij} > 0, \ \forall i \in I_j \end{aligned} \tag{3.8}$$

- *Step 2*:

$$\begin{aligned} &\min\left(b_{ij} - b_{ij}(k-1)\right)^2, \forall i, j \notin OJ \\ &\text{subject to } \sum_{j \in J_i} b_{ij} = 1; b_{ij} > 0, \forall j \notin OJ \end{aligned} \tag{3.9}$$

3.2.4.2 Solution Algorithm

Based on the same procedure of converting model (3.3) to model (3.4), the first-step model (3.8) can be represented by the following system equations:

$$\begin{cases} b_j(k) = b_j(k-1) \\ y_j(k) = \bar{q}^T(k) b_j(k) + e_j(k) \end{cases}, \forall j \in OJ \tag{3.10}$$

where the covariance of the random error $e_j(k)$ is assumed to be known, denoted as $r_j(k)$.

Since the result of this estimator is expected to be very close to the true value at convergence, we ignore the inequality constraints and apply the unconstrained Kalman filtering for each $j \in OJ$ separately to identify this system:

$$\begin{cases} K_j(k) = P_j(k)\bar{q}(k)\left[\bar{q}(k)^T P_j(k)\bar{q}(k) + r_j(k)\right]^{-1} \\ \hat{b}_j(k) = \hat{b}_j(k-1) + K_j(k)\left[y_j(k) - \bar{q}(k)^T \hat{b}_j(k-1)\right] \\ P_j(k+1) = \left[I - K_j(k)\bar{q}(k)^T\right] P_j(k) \end{cases} \tag{3.11}$$

It is apparent from (3.11) that after the decomposition, no matrix inversion is needed when computing the Kalman gain $K_j(k)$ since the item in the bracket is only a scalar. Thus, this formulation improves the computational efficiency to a great extent.

The second-step model is an equality-constrained least-squares problem for each row *i* of the O-D matrix, after omitting the non-negativity constraints. It is essentially a projection of the previous O-D parameters to a plane in the b_{ij} space governed by the result from step 1. Since the exiting counts for the two major approaches of a typical four-way intersection are known, two columns of the O-D matrix will be estimated from step 1. Therefore, the second-step formulation is at most a two-dimensional minimization problem for each row *i*, which can be easily solved.

If the left/right-turn splits on the major approach can be directly measured (or additional information of the O-D pattern is given, such as U-turn is known to be prohibited), the number of unknown variables can be further reduced by separately estimating those O-D parameters with available observations. As a consequence, the result is expected to be more accurate.

3.2.5 Numerical Example

In this section, a simulation example is provided to illustrate the two methods of estimating O-D splits for an actuation-controlled intersection with incomplete exiting counts.

Consider a typical four-way intersection prohibiting U-turn, as shown in Fig. 3.2. All entering counts are available, but the exiting counts are only available at major approaches, leg 1 and leg 3. The duration of the simulation was set as 100 time intervals.

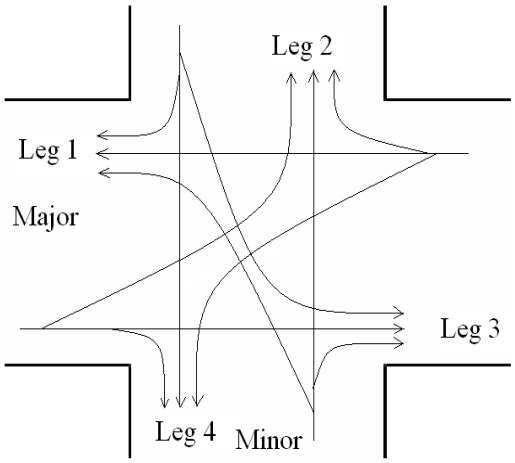


Fig. 3.2. A typical four-way intersection

To mimic the operation of the estimator, we adopted the hypothetical true O-D matrix used by Nihan and Davis (1987) shown as follows:

$$\begin{matrix} 0 & .2 & .7 & .1 \\ .05 & 0 & .8 & .15 \\ .3 & .2 & 0 & .5 \\ .1 & .8 & .1 & 0 \end{matrix}$$

The actual entering counts were generated randomly from 0 to 100 for each time interval, and observed exiting counts for exits 1 and 3 were generated based on (3.2). The measurement errors for both legs 1 and 3 were assumed to be independent normal random variables with zero mean; and the variance $r_j(k)$ is set equal to 15 percent of the actual exiting traffic.

The initial value of the O-D parameters was given as:

$$\begin{matrix} 0 & .33 & .33 & .34 \\ .33 & 0 & .33 & .34 \\ .33 & .33 & 0 & .34 \\ .33 & .33 & .34 & 0 \end{matrix}$$

And the P matrices were always initialized as identity matrices with zero elements on the diagonal corresponding to the O-D parameters $b_{ii} = 0$.

Five simulation experiments were conducted for both the conventional GLS method and the two-step method. The root mean squares (RMS) of the difference between the true and the estimated O-D parameters were computed as a measurement of estimator performance at each time interval. Table 3.1 shows the average RMS over the last 20 iterations for both estimators, leading to the conclusion that the two-step method outperforms the conventional GLS method.

Fig. 3.3 displays the RMS errors across iterations for experiment 2 and 3. And the convergence of the O-D splits for experiment 5 is plotted in Fig. 3.4 to Fig. 3.7.

Table 3.1 Average RMS Error over Last 20 Iterations

Experiment	Conventional GLS Method	Two-Step Method
1	0.0251	0.0184
2	0.0232	0.0184
3	0.0234	0.0185
4	0.0204	0.0183
5	0.0292	0.0185

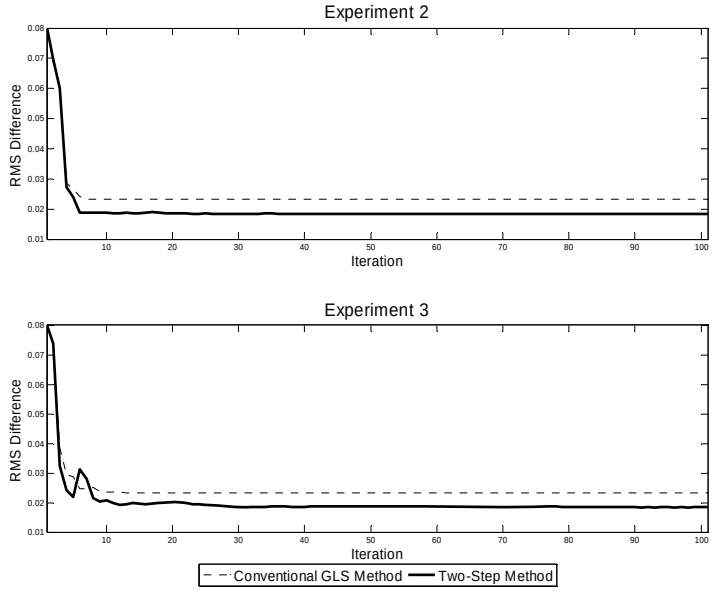


Fig. 3.3. RMS error between actual and estimated O-D parameters for Experiment 2 and Experiment 3

Fig. 3.4 and Fig. 3.6 show that for those exits with available traffic counts, the two-step

method provides more accurate estimates of their corresponding columns of the O-D matrix, which do not depend on the initial value. Note that b_{22} and b_{44} are known to be zero, leaving only one decision variable in the second step for both row 2 and row 4. Therefore, the accuracy in the first and the third column estimation of the O-D matrix will result in an unbiased estimation of b_{24} and b_{42} (See Fig. 3.5 and Fig. 3.7). More specifically, in the two-step method only four O-D parameters are dependent on the initial value while in the conventional GLS method, all the twelve splits to be estimated rely on the initial value.

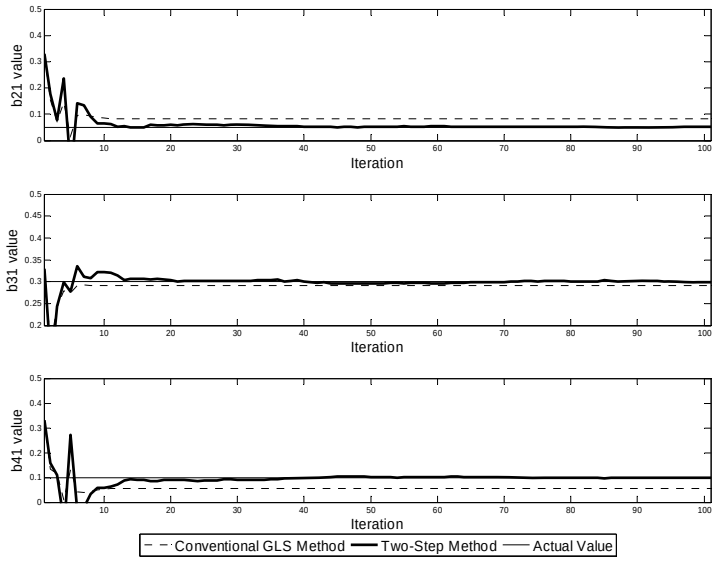


Fig. 3.4. Convergence of O-D parameters b_{i1} for Experiment 5

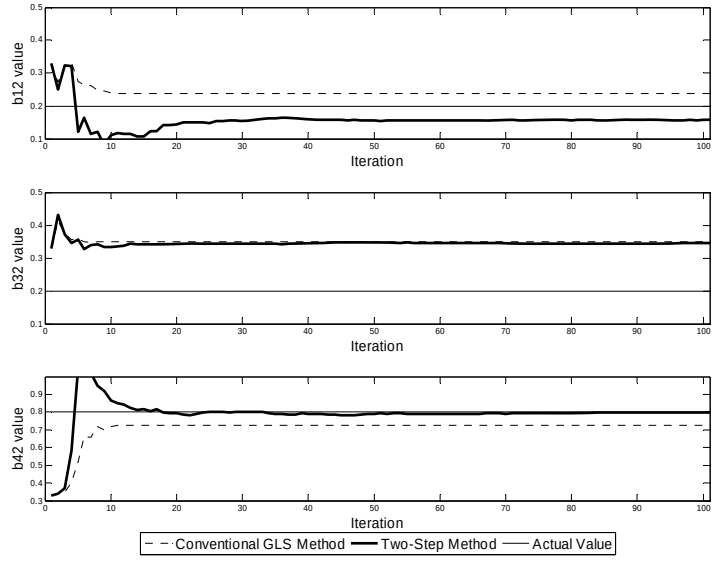


Fig. 3.5. Convergence of O-D parameters b_{i2} for Experiment 5

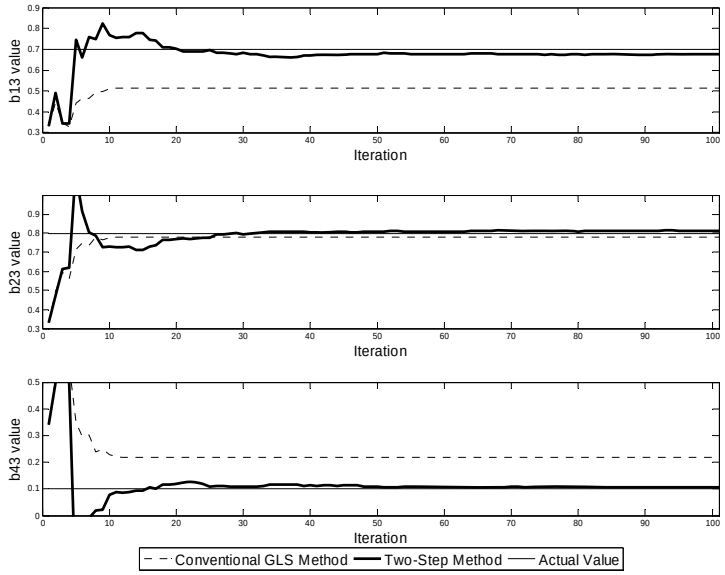


Fig. 3.6. Convergence of O-D parameters b_{i3} for Experiment 5

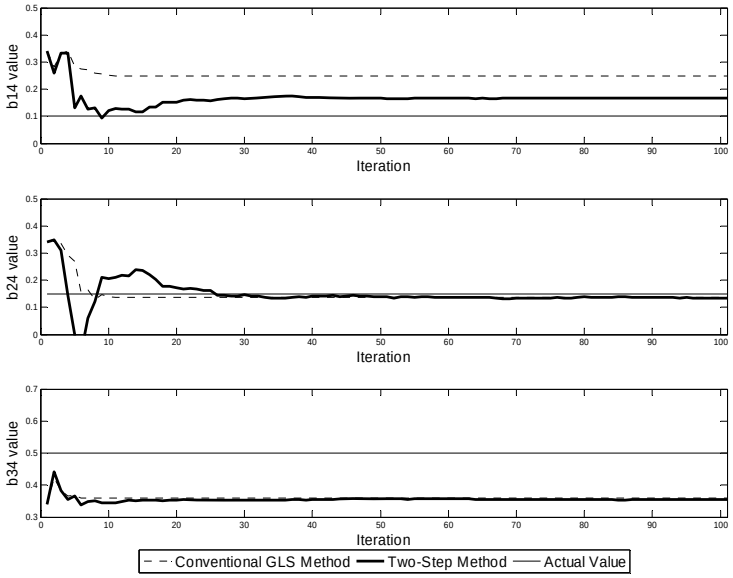


Fig. 3.7. Convergence of O-D parameters b_{i4} for Experiment 5

3.2.6 Tacking Time-Varying O-D Flows

Above we presented a two-step formulation for real-time estimation of O-D matrix of isolated actuation-controlled intersections with complete entering counts and incomplete exiting counts. Though our discussion focuses on the constant O-D matrix estimation problem, the proposed framework can be easily extended to track time-varying O-D matrices by modeling the O-D splits as a random walk process. Accordingly, the covariance matrix of the random deviation should be introduced to the Kalman filtering in order to correct the P matrices at each step.

We applied the extended two-step approach to the same intersection used in Section 3.2.5, and Figs. 3.8-3.11 compare the estimates of the new approach and conventional GLS with the time-dependent true values, and Fig. 3.12 reports the RMS errors of the two approaches. It can be observed that the two-step approach still outperforms the conventional GLS approach. Moreover, the former is more efficient as well. It should also pointed out that both approaches would benefit from accurate prior knowledge or partial O-D information.

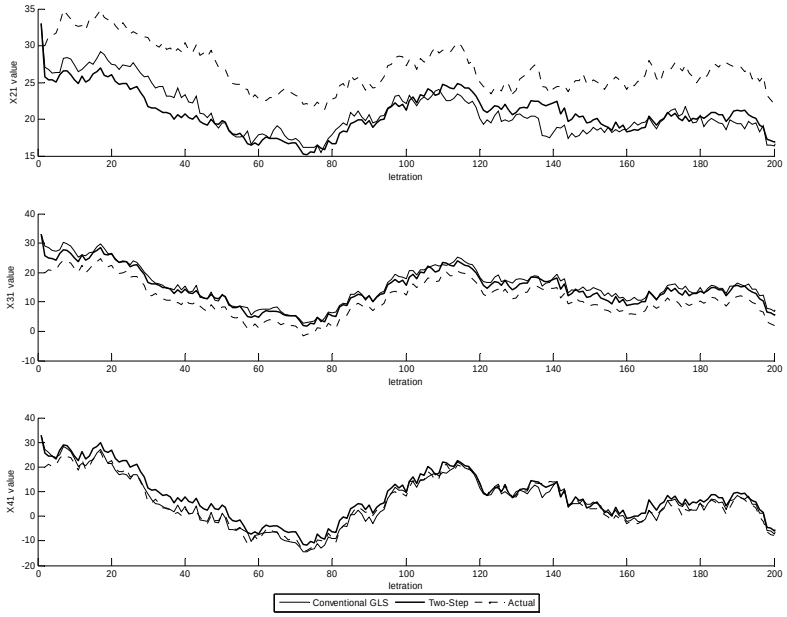


Fig. 3.8. Comparison of flows to Leg 1

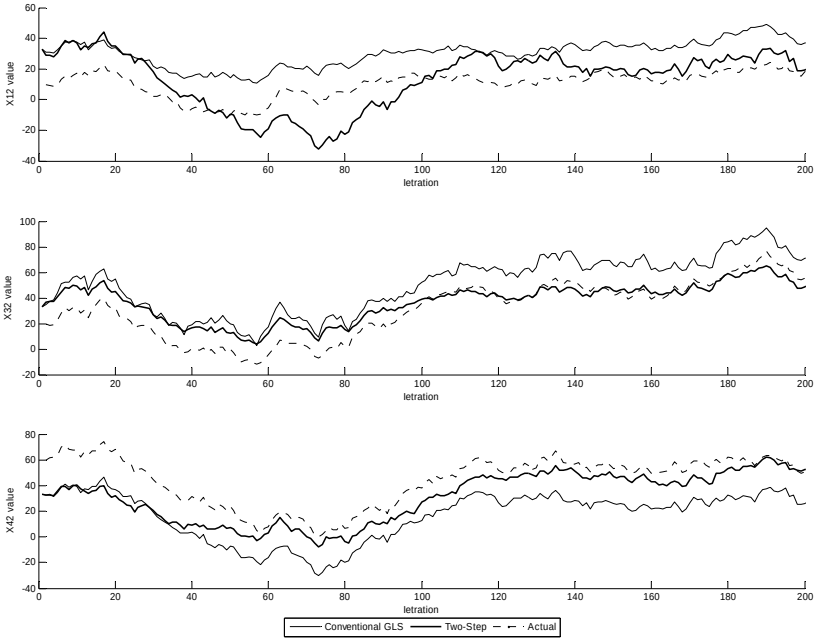


Fig. 3.9. Comparison of flows to Leg 2

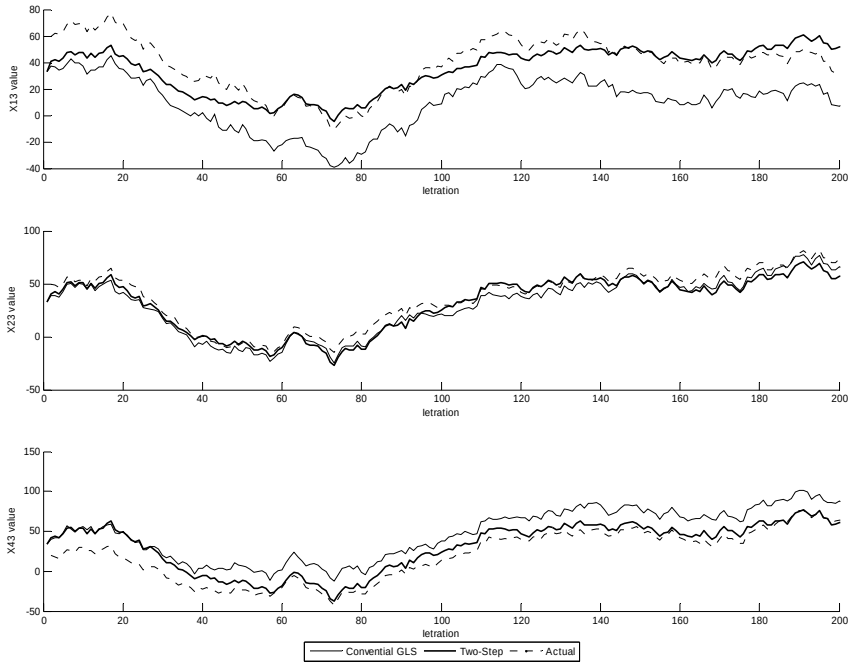


Fig. 3.10. Comparison of flows to Leg 3

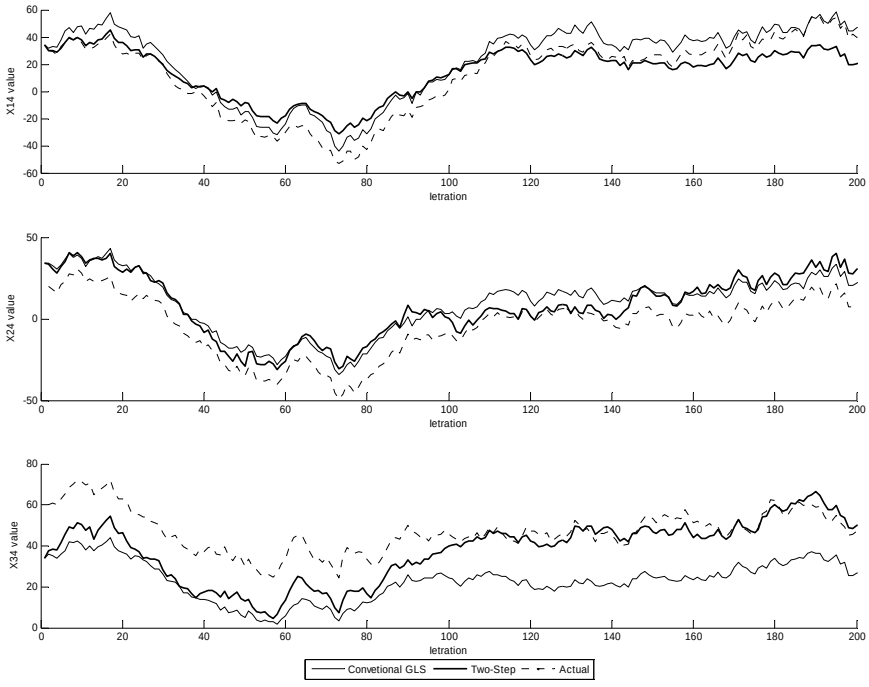


Fig. 3.11. Comparison of flows to Leg 4

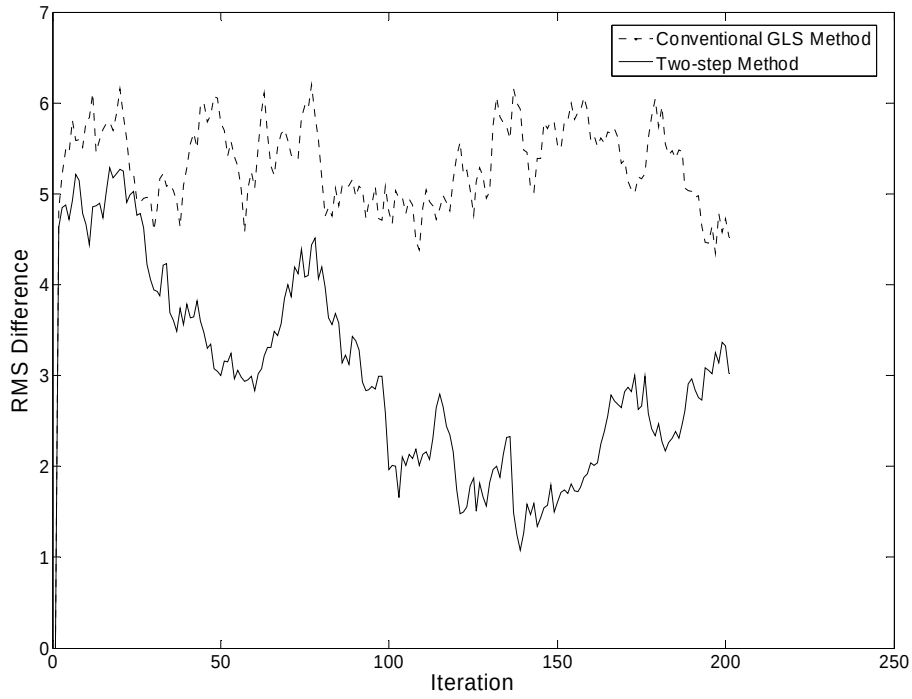


Fig. 3.12. Comparison of RMS errors

3.3 Estimation of Origin-Destination Flows for Actuation-Controlled Corridors

3.3.1 Introduction

The model presented in the previous section focuses on turning movement identification for isolated intersections where travel time can be safely assumed negligible. Extending the model to a corridor (more precisely, linear network) will inevitably involve determination of vehicle travel times and descriptions of flow propagation for constructing the casual relationship between count measurements and O-D flows to be estimated.

Existing closed-network models often make assumptions on the platoon dispersion or flow propagation and then endogenously estimate both O-D flows and the propagation parameters. In the literature, Bell (1991) proposed two methods to allow for distributions of travel times through intersection or network that span more than one interval. The first method assumes a geometrically distributed travel time for each exit. In this method, the parameters to be estimated are the platoon dispersion factor together with the O-D splits. The second method considers freely-distributed travel times, but assumes vehicles from any entrance should reach to a specified exit within three time intervals. Thus, the variables to be estimated at each time interval are exactly the O-D parameters for the three intervals before the current one. Chang and Wu (1994) used macroscopic traffic models to estimate travel times, and then assumed vehicles that reach one exit during an interval come from only two consecutive time intervals for each entrance. Moreover, the proposed method represents flow propagation by introducing link-use proportions, which are estimated simultaneously with the O-D parameters. Consequently, not only the dimension of the problem increases, but also the problem itself becomes nonlinear. Extended Kalman filtering was adopted to identify the nonlinear system.

In the open-network approach, the system equation used is as follows:

$$\vec{y}(t) = \sum_{k=t-p}^{t} A(k,t)\vec{f}(k) + \vec{e}(t)$$

where $\vec{y}$ is the vector of the measured link counts; $\vec{f}$ is the vector of O-D flows to be estimated; $\vec{e}$ is the random error vector and A is the assignment matrix, which encapsulates route choice and flow propagation. The element of the assignment matrix A is the link-use proportion, which is defined as the proportion of a particular O-D flow departing its origin during interval k, prior to the current interval t by at most p intervals, contributes to the flow on link l during interval t. The complexity of describing route choice and flow propagation is often avoided in many previous studies by simply assuming that the matrix is known and offering some general discussions that the matrix can be computed using simulation, or DTA models or the analytical equations, if the travel times are known (e.g., Okutani, 1987; Ashok and Ben-Akiva, 1993, 2000). In fact, even with the restrictive assumptions that travel times are known, and users are homogenous and there is no route choice, the assignment matrix cannot be exactly determined for a network with active bottlenecks and multiple O-D pairs because the

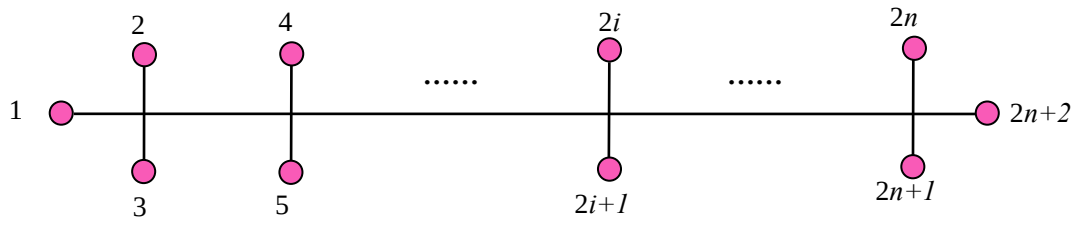
bottlenecks may prevent vehicles departing together arriving at the same time interval. Additional assumption has to be made on the platoon dispersion. For example, Cascetta et al. (1993) assumed that vehicles within a group departing at interval *k* using path *p* are uniformly comprised within the departure duration *T* and stay within the interval as they move across the network. Ashok and Ben-Akiva (2002) further relaxed the assumption to permit the effects of "stretching" and "squeezing" of packets as they traverse the network.

In this chapter, we estimate time-dependent O-D flows for actuation-controlled corridors where all of the entering flows and exiting flows on major streets are known while the exiting counts on minor streets are missing. We assume that travel times are known from exogenous source, e.g., a traffic surveillance system or are estimated using traffic simulation. For the latter, we may first decompose the network into a system of isolated intersections, and then apply the model proposed in the previous paper to estimate the turning movements for each intersection. The estimated movements can be further fed into a traffic simulation package, such as VISSIM, to estimate travel times. Based on the estimated travel times, flow propagation can be described for the linear network with an assumption of platoon dispersion, and the O-D flows can then be inferred accordingly. This approach may eliminate the potential inconsistency between the estimated O-D matrix and the travel times, which is one of the major problems in many previous dynamic network O-D estimation models.

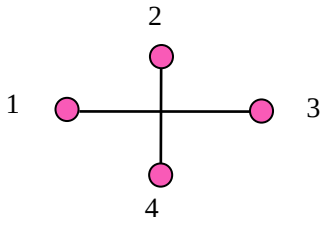
3.3.2 Model Formulation

3.3.2.1 Model Preparation

Consider a corridor consisting of *n* intersections, then the dimension of the O-D matrix is $2(n+1)\times 2(n+1)$, and the node numbering convention is presented in Fig. 3.13.



a. Corridor



b. Intersection

Fig. 3.13. Numbering convention for model formulation

The following notation is used for the model formulation:

- f_{ij}^{t}: number of vehicles entering the corridor from origin i to destination j, $i, j \in \{1,2,...,2n+2\}$, during time interval t;
- θ_{ijl}^{tk}: fraction of f_{ij}^{t} that arrives at intersection l during time interval k, $i, j \in \{1,2,...,2n+2\}, l \in \{1,2,...,n\}$ [4];
- $\eta_{ij,l}^{t}$: turning movements from leg i to leg j during time interval t at intersection l, $i, j \in \{1,2,3,4\}, l \in \{1,2,...,n\}$;
- T_j^1: maximum travel time from any intersection i ($i < j$) to intersection j;
- T_j^2: maximum travel time from any intersection i ($i > j$) to intersection j;

3.3.2.2 Decomposition Scheme

We propose a two-step decomposition scheme to estimate time-varying O-D matrices for corridors with incomplete information. The essential idea is to tackle the problem in two steps or levels: intersection and corridor levels. At the intersection level, the decomposition scheme presented in the previous chapter is applied to infer turning movements for each individual intersection; at the corridor level, the original problem is further decomposed into a series of sub-problems with respect to each destination or origin, using the turning movements estimated at the intersection level. As stated in the

[4] In the proposed method, we assume this factor is known from platoon dispersion models so that the problem remains linear. Otherwise, extended Kalman filtering may be adopted to infer both O-D flows f and the dispersion factors θ.

previous chapter, the decomposition scheme makes full use of available information and hence reduces the dependency on the quality of the prior O-D information at the intersection level. At the corridor level, it does not even require any prior O-D information. The two-step approach improves significantly the computational efficiency by decomposing the original high-dimensional problem into much smaller problems.

Specifically, at the intersection level, for each time interval *k*, the turning movement $\eta_{ij,l}^k$, $\forall i, j \in \{1,2,3,4\}, l \in \{1,2,...,n\}$ can be estimated at all intersections. As a consequence, 2*n*+8 O-D flows can be explicitly estimated, including minor-to-minor O-D flows $f_{2l,2l+1}^k, f_{2l+1,2l}^k$ for each intersection l , $f_{12}^k, f_{13}^k, f_{21}^k, f_{31}^k$ for intersection 1 and $f_{2n+2,2n+1}^k, f_{2n+2,2n}^k, f_{2n,2n+2}^k, f_{2n,2n+1}^k$ from intersection *n.* Among those estimates, only four are biased, namely $f_{12}^k, f_{13}^k, f_{2n+2,2n+1}^k$ and $f_{2n+2,2n}^k$, as shown in the previous chapter.

At the corridor level, to improve the computational efficiency, we decompose the original high-dimensional problem into a series of sub-problems with respect to each destination or origin. In other words, we estimate the O-D matrix column by column (with respect to each destination) or row by row (for each origin). These two estimates can be further combined through a weighted average to improve the quality of the final estimate. The weighing factors may be chosen based on the variances of the estimation errors, which are updated iteratively during the Kalman filtering process. Let $\hat{f}_o$ and $\hat{f}_d$ be two estimates obtained using the row (by origin) and column (by destination) decomposition respectively and $\hat{e}_o$ and $\hat{e}_d$ the corresponding estimation errors respectively. The final estimate may take the format as

$$\hat{f} = \frac{Var(\hat{e}_d)\hat{f}_o + Var(\hat{e}_o)\hat{f}_d}{Var(\hat{e}_o) + Var(\hat{e}_d)}$$

Assume $\hat{f}_o$ and $\hat{f}_d$ are independent estimates, it can be shown that the variance of the final estimation error is $\frac{Var(\hat{e}_d)Var(\hat{e}_o)}{Var(\hat{e}_o) + Var(\hat{e}_d)}$, less than that of either original estimates.

For row estimation, either the total inflow at each origin or the turning movements obtained from the intersection level problem can serve as observations. However, to avoid the accumulation of estimation errors, it is more favorable to estimate the O-D flows based on the first-hand total inflow information. Therefore, the causal relationships are:

$$q_o^k = \sum_{d=1}^{2n+2} f_{o,d}^k$$

where q_0^k is the inflow from origin *o* at time interval *k*.

The column estimation is more complicated since there is no direct observation of the total departures. For each destination $2j$, $j \in \{1,2,...,n\}$, two estimates of turning movements resulted from the intersection level problem, $\eta_{12,j}$ and $\eta_{32,j}$, may be selected and the causal relationships between turning movements and O-D flows are:

$$\eta_{12,j}^{k} = \sum_{t=k-T_j^1}^{k} \sum_{o=1}^{2(j-1)+1} \theta_{o,2j,j}^{t,k} f_{o,2j}^{t}$$

$$\eta_{32,j}^{k} = \sum_{t=k-T_j^2}^{k} \sum_{o=2(j+1)}^{2(n+1)} \theta_{o,2j,j}^{t,k} f_{o,2j}^{t}$$

For individual destination $2j+1$, $j \in \{1,2,...,n\}$, $\eta_{14,j}$ and $\eta_{34,j}$ may be selected to infer the O-D flows destined to the node:

$$\eta_{14,j}^{k} = \sum_{t=k-T_j^1}^{k} \sum_{o=1}^{2(j-1)+1} \theta_{o,2j+1,j}^{t,k} f_{o,2j+1}^{t}$$

$$\eta_{34,j}^{k} = \sum_{t=k-T_j^2}^{k} \sum_{o=2(j+1)}^{2(n+1)} \theta_{o,2j+1,j}^{t,k} f_{o,2j+1}^{t}$$

In addition, $\eta_{31,1}$ and $\eta_{13,n}$ may be used to estimate for the first and last column of the O-D matrix:

$$\eta_{31,1}^{k} = \sum_{t=k-T_1^2}^{k} \sum_{o=2}^{2(n+1)} \theta_{o,1,1}^{t,k} f_{o,1}^{t}$$

$$\eta_{13,n}^{k} = \sum_{t=k-T_n^1}^{k} \sum_{o=1}^{2n+1} \theta_{o,2n+2,n}^{t,k} f_{o,2n+2}^{t}$$

Note that $\eta_{12,1} = f_{12}$, $\eta_{14,1} = f_{13}$, $\eta_{32,n} = f_{2n+2,2n}$ and $\eta_{34,n} = f_{2n+2,2n+1}$. Therefore, the original problem is decomposed into $4n-4+2 = 4n-2$ sub system identification problems in total.

3.3.2.3 State-Space Representation

For the formulation, we assume that U-turn is prohibited at each intersection and $f_{ii}^{t} = 0$. Moreover, the O-D flows and turning movements are assumed to be first-order auto-regress processes, although more general structures can be accommodated if historical O-D data are available to determine the structure.

Due to the flow propagation, turning movements at particular interval can be attributed to O-D flows at previous multiple intervals. For example, the turning movement $\eta_{12,j}^{k} = \sum_{t=k-T_j^1}^{k} \sum_{o=1}^{2(j-1)+1} \theta_{o,2j,j}^{t,k} f_{o,2j}^{t}$, which suggests that each O-D flow be estimated multiple times. As in Ashok and Ben-Akiva (1993, 2000), we use state augmentation to achieve this, and the augmented space vector for this sub-problem would be:

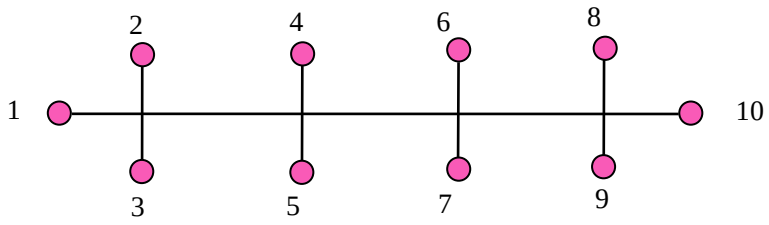
$$\left(\begin{array}{c} f_{1,2j}^{k-T_j^1} \\ \vdots \\ f_{2j-1,2j}^{k-T_j^1} \\ \hline f_{1,2j}^{k-T_j^1+1} \\ \vdots \\ f_{2j-1,2j}^{k-T_j^1+1} \\ \hline \vdots \\ \hline f_{1,2j}^{k} \\ \vdots \\ f_{2j-1,2j}^{k} \end{array}\right) \quad \begin{array}{l} \left.\begin{array}{l} \\ \\ \end{array}\right\} k-T_j^1 \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} k-T_j^1+1 \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \left.\begin{array}{l} \\ \\ \end{array}\right\} k \end{array}$$

As in the previous section, the Kalman Filtering is adopted to infer the time-varying O-D matrices. Note that the state vector is different for each time interval. After *k*-1 time intervals, we have the (*k*-1)$^{\text{th}}$ estimate for time dependent O-D flows f_{ij}^t, $t \in \{1,2,...,k-1\}$, denoted as $f_{ij}^t(k-1)$. At time interval *k*, since f_{ij}^k is now included in the state vector (meanwhile $f_{ij}^{k-1-T_j^1}$ is excluded), an initial value of this time *k* O-D flow is needed for the k^{th} estimation. We let $f_{ij}^k(k-1) = f_{ij}^{k-1}(k-1)$ be the initial value of f_{ij}^k. Consequently, $f_{ij}^t(k)$, $t \in \{1,2,...,k\}$, is equal to $f_{ij}^t(k-1)$ plus the correction item.

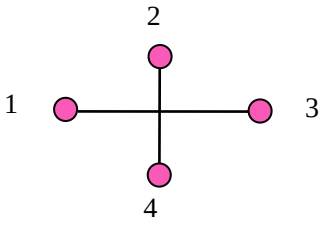
3.3.3 Numerical Experiment

3.3.3.1 Experiment Settings

We demonstrate and verify the proposed two-step approach on a hypothetical corridor with four intersections and 90 O-D pairs in total, as shown in Fig. 3.14. Two scenarios were tested, one for constant O-D flows and the other for time-varying O-D flows, generated based on the first-order auto-regressive assumption. The maximum travel times from any intersection to intersection j were set as $T_j^1 = j-1+2 = j+1$ and $T_j^2 = n-j+2 = 4-j+2 = 6-j$. Assuming no platoon dispersion (i.e, θ_{ijl}^{tk} is either zero or one, indicating whether O-D flow f_{ij}^t passes intersection *l* at time *k* or not), we used a traffic loading procedure to determine the entering, exiting flows and turning movements at each intersection. Finally, measurement errors were randomly added to those flows, which were then used by the proposed approach to infer the O-D flows.



a. Corridor



b. Intersection

Fig. 3.14. Numbering convention for the hypothetical corridor

3.3.3.2 Experiment Results

We first applied the proposed approach to infer constant O-D flows. The initial values of the O-D flows were randomly generated. Figures 3.15-3.18 compare the actual and estimated O-D flows for selected O-D pairs, including end to end (10-1), long-distance minor to minor (9-2), medium-distance minor to minor (2-6) and turning movement in the same intersection (8-10). It can be found that the estimates converge quickly within 10 intervals and there is very good agreement between the actual and estimated values for all cases (within 5% deviation).

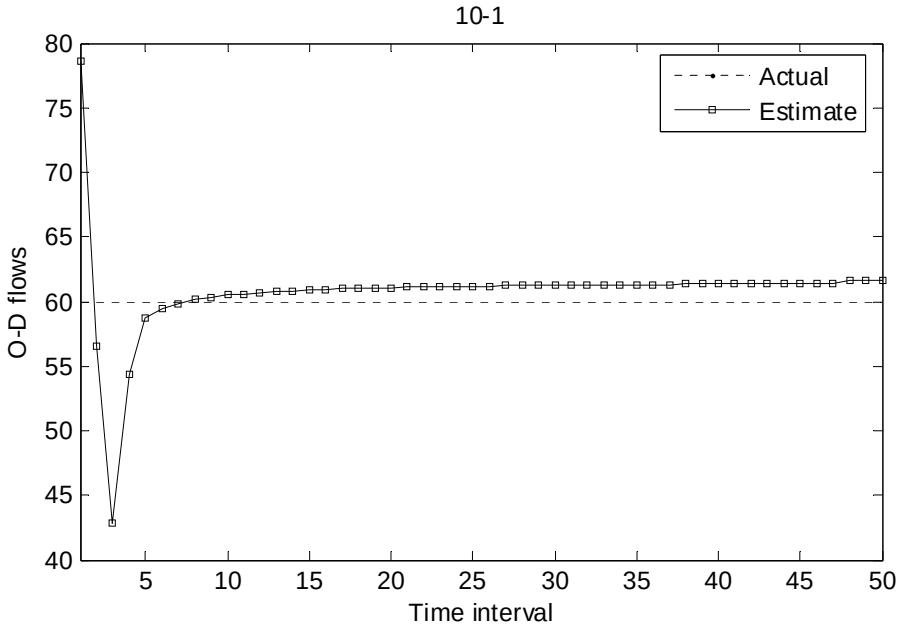


Fig. 3.15. Actual vs Estimated O-D flow for O-D pair 10-1

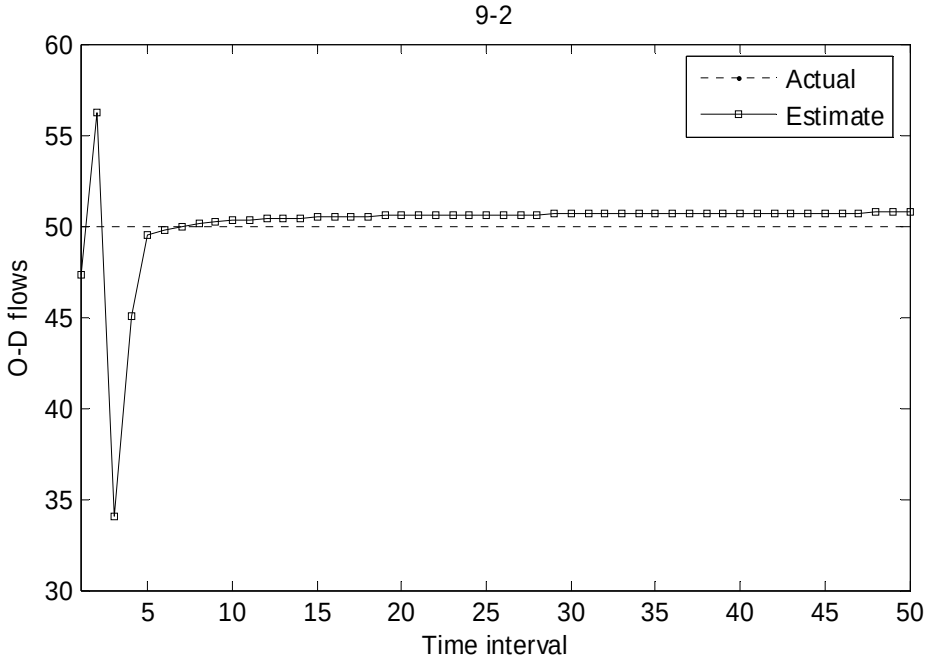


Fig. 3.16. Actual vs Estimated O-D flow for O-D pair 9-2

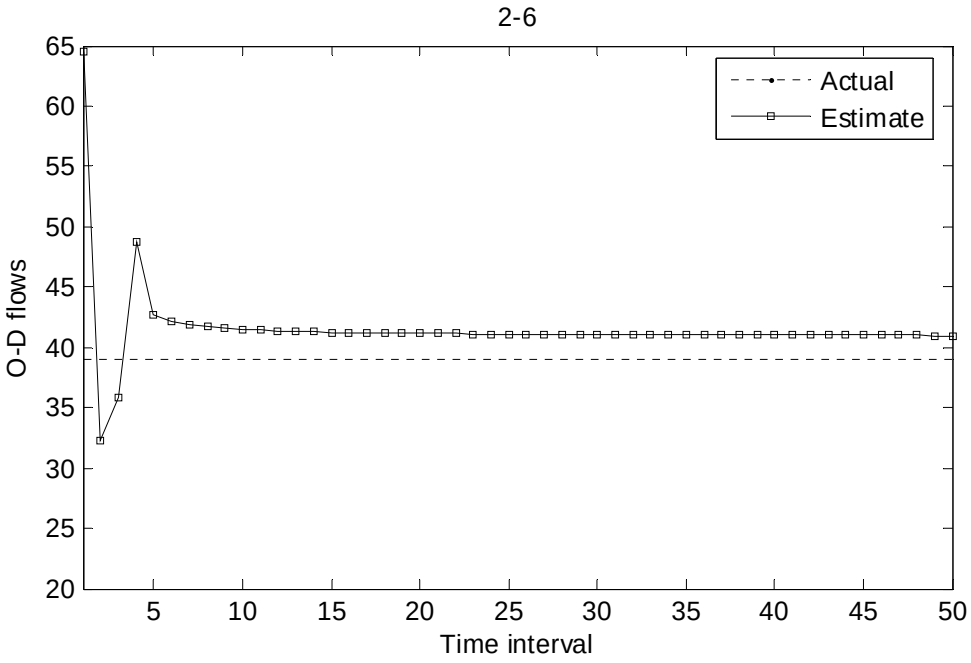


Fig. 3.17. Actual vs Estimated O-D flow for O-D pair 2-6

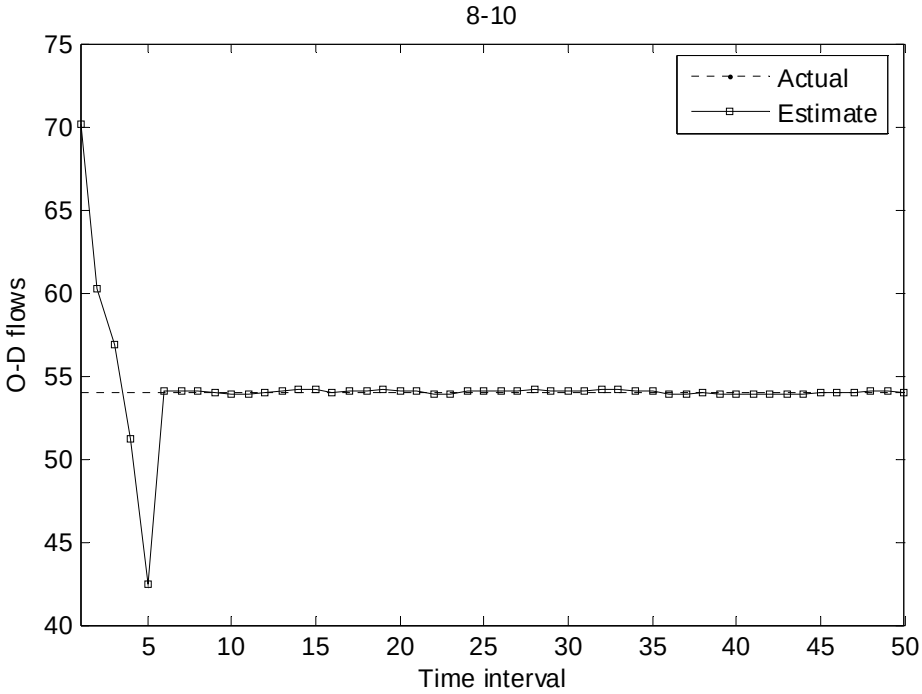


Fig. 3.18. Actual vs Estimated O-D flow for O-D pair 8-10

Table 3.2 presents the root mean square error normalized (RMSN) for each O-D pair. It shows that most of the O-D pairs experience rather small errors while certain O-D pairs, such as 4-1, 8-1 and 3-5, suffer higher errors, likely due to the poor quality of the prior O-D information.

Table 3.2 Root Mean Square Error Normalized (RMSN)

	1	2	3	4	5	6	7	8	9	10
1		0.090	0.030	0.259	0.171	0.070	0.241	0.099	0.201	0.102
2	0.051		0.050	0.375	0.218	0.122	0.232	0.052	0.223	0.105
3	0.040	0.061		0.166	0.522	0.127	0.237	0.205	0.091	0.109
4	0.452	0.106	0.047		0.107	0.141	0.049	0.146	0.325	0.129
5	0.140	0.231	0.271	0.054		0.143	0.160	0.117	0.208	0.113
6	0.062	0.268	0.237	0.185	0.114		0.071	0.305	0.186	0.100
7	0.068	0.395	0.168	0.059	0.233	0.060		0.201	0.208	0.040
8	0.414	0.267	0.100	0.064	0.053	0.065	0.176		0.058	0.058
9	0.071	0.055	0.224	0.072	0.075	0.141	0.211	0.066		0.055
10	0.067	0.278	0.142	0.146	0.074	0.050	0.391	0.085	0.065	

We then inferred time-varying O-D flows. The initial values of the O-D flows were randomly generated as well, and the comparisons between estimates and true values for selected O-D pairs are presented in Figs. 3.19-3.22. The RMSN of the O-D flows estimated by destination is 0.223; the values are 0.230 and 0.215 for the origin-based and the weighted-average estimates. Table 3.3 further presents RMSN of the weighted-average estimate for each O-D pair.

We have the following two observations:

- Both the column and row decompositions are able to track the trend of time-varying O-D flows and produce estimates pretty close to the actual values in an average sense. The row decomposition is more computationally efficient but the column decomposition provides better estimates in terms of the RMSN. The reason is that the row decomposition only makes use of the arrival information and infers each time-dependent O-D flow once it leaves while the column decomposition essentially uses both the arrival and departure information. However, the proposed column decomposition approach at the corridor level only focuses on the localized information (i.e., the turning movements at the intersection containing destination node) and discards other information available upstream (as the vehicles shall be observed at all these intermediate intersections).
- The weighted average of the two estimates replicates actual O-D flows better. The RMSN drops by 6.5% and 3.4%, compared with the row and column decomposition respectively. Figures 7-10 also suggest that the weighted-average estimate follows the actual O-D pattern well. Consequently, the RMSN for each O-D pair presented in Table 3.3 is comparable with its counterpart in Table 3.2 for the constant O-D case.

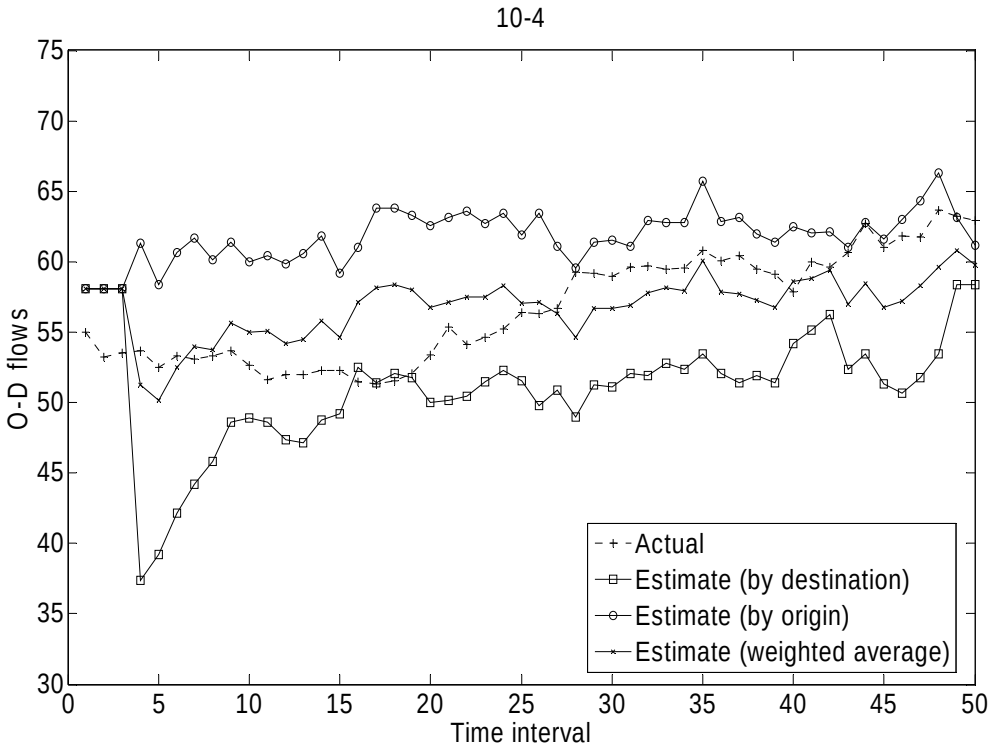


Fig. 3.19. Actual vs Estimated O-D flow for O-D pair 10-4

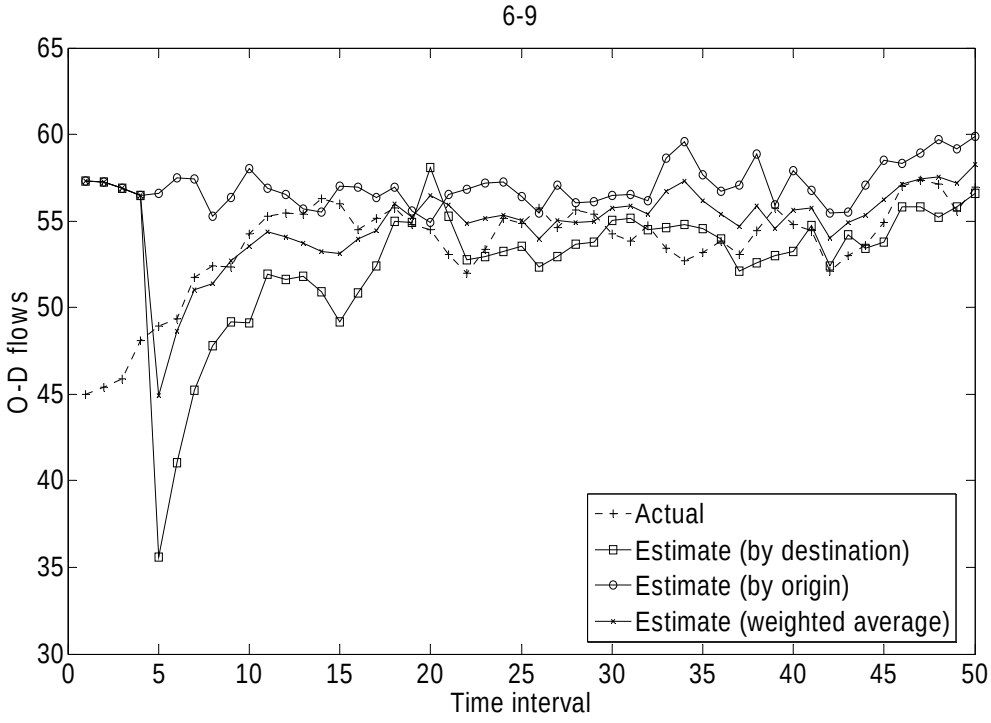


Fig. 3.20. Actual vs Estimated O-D flow for O-D pair 6-9

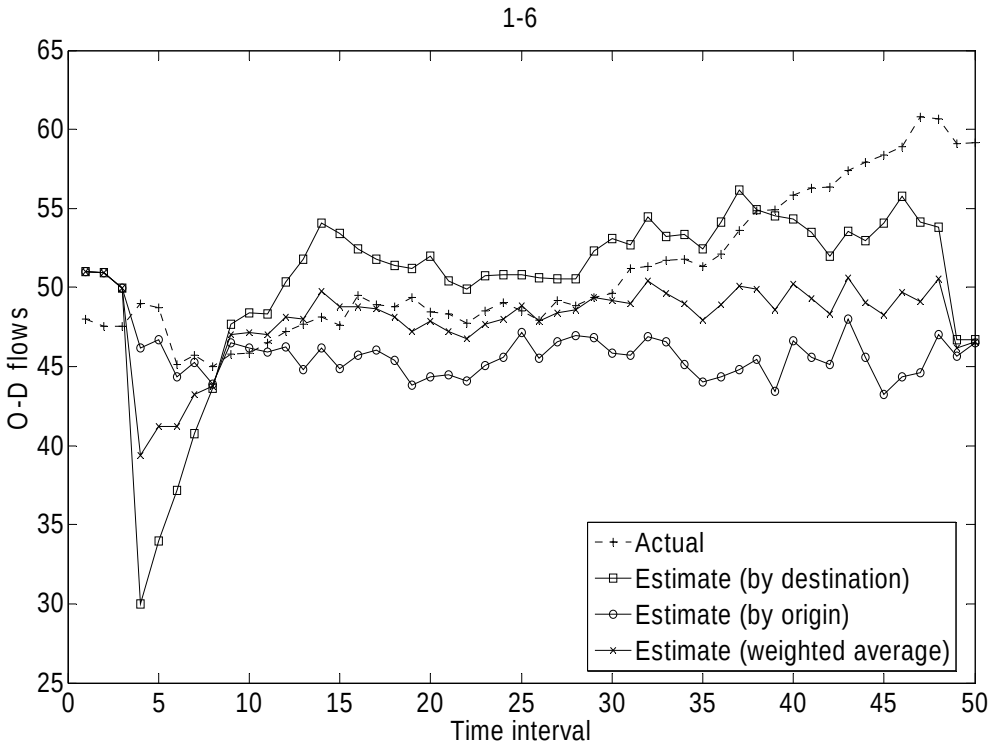


Fig. 3.21. Actual vs Estimated O-D flow for O-D pair 1-6

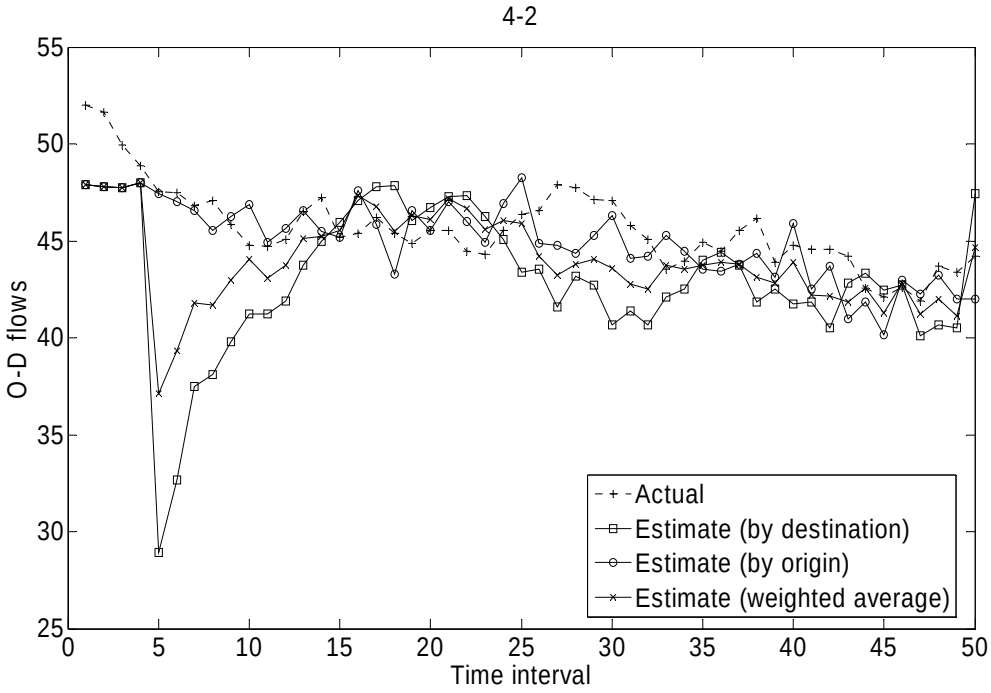


Fig. 3.22. Actual vs Estimated O-D flow for O-D pair 4-2

Table 3.3 RMSN of Estimates of Time-Varying O-D Flows

	1	2	3	4	5	6	7	8	9	10
1		0.147	0.184	0.290	0.127	0.079	0.251	0.077	0.204	0.145
2	0.311		0.163	0.107	0.076	0.114	0.385	0.273	0.087	0.414
3	0.174	0.213		0.317	0.216	0.519	0.080	0.103	0.108	0.104
4	0.070	0.066	0.146		0.302	0.377	0.087	0.073	0.292	0.067
5	0.077	0.117	0.461	0.090		0.095	0.187	0.080	0.253	0.137
6	0.295	0.052	0.114	0.325	0.369		0.067	0.073	0.070	0.193
7	0.059	0.053	0.297	0.121	0.137	0.291		0.143	0.086	0.210
8	0.184	0.241	0.150	0.109	0.457	0.253	0.141		0.195	0.059
9	0.170	0.091	0.393	0.090	0.270	0.169	0.195	0.290		0.095
10	0.089	0.228	0.160	0.056	0.291	0.113	0.121	0.069	0.049	

3.3.4 Real-World Application

We applied the proposed O-D estimator to a segment of El Camino Real, San Mateo, CA using the data collected on February 1st, 2007. Since we do not have real O-D observations from the corridor, the accuracy of the estimates cannot be verified. Therefore, the purpose of the application is to demonstrate that the estimator is able to readily work with actual field loop data.

The testing site consists of eight signalized intersections as shown in Figure 3.23. The distances between intersections (in meters) are presented in the illustration. There are four on/off ramps between the 17th Ave. and the 20th Ave. accessing J. Arthur Younger Freeway. However, since the vehicle counts onto/off the freeway are not available, it is assumed in this application that the ramp flows are zero.

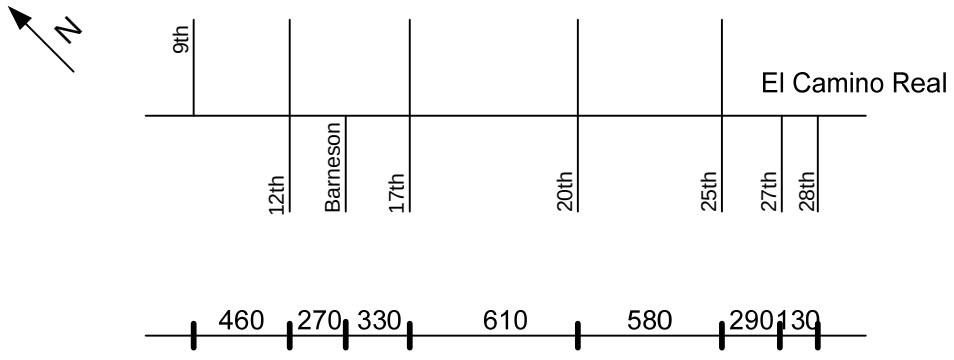


Fig. 3.23. Illustration of the Testing Site

In order to avoid possible pattern shifts between peak/off-peak hours, only off-peak loop counts from 9:49-14:04 were used. The raw data were aggregated into a five-minute resolution. Among these eight intersections, there is no loop information

available for 17th Ave., and the counts for the east and the north bounds of 28th Ave. intersection are missing as well. Therefore, with the assumption that there are no turning movements at the 17th intersection, the corridor is modeled as a four-intersection system as shown in Figure 3.24. Note that 9th Ave. and 27th Ave. intersections are excluded from the system as their main line departure counts (obtained from adjacent intersection advanced loop counts) are not available.

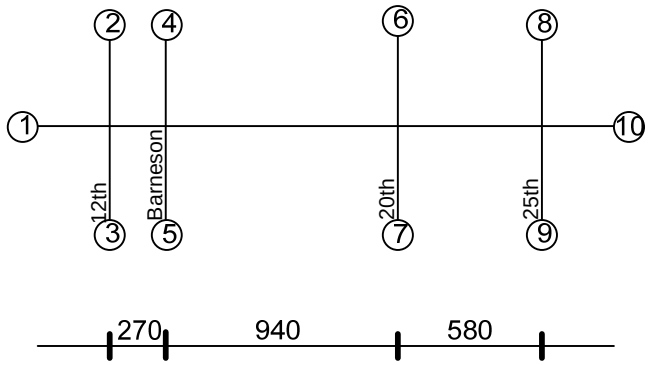


Fig. 3.24 Modeling of the Corridor

The initial values of the O-D flows were generated manually based on the counts of the first time interval. Since the travel time (estimated based on the speed limit of 35mph) between any of the two intersections is less than the duration of a single time interval, we assume here the flow propagation is neglectable and all the vehicles arrive at their destinations within the same time interval as they depart. Selected resulting estimates are displayed in Figures 3.25 – 3.27.

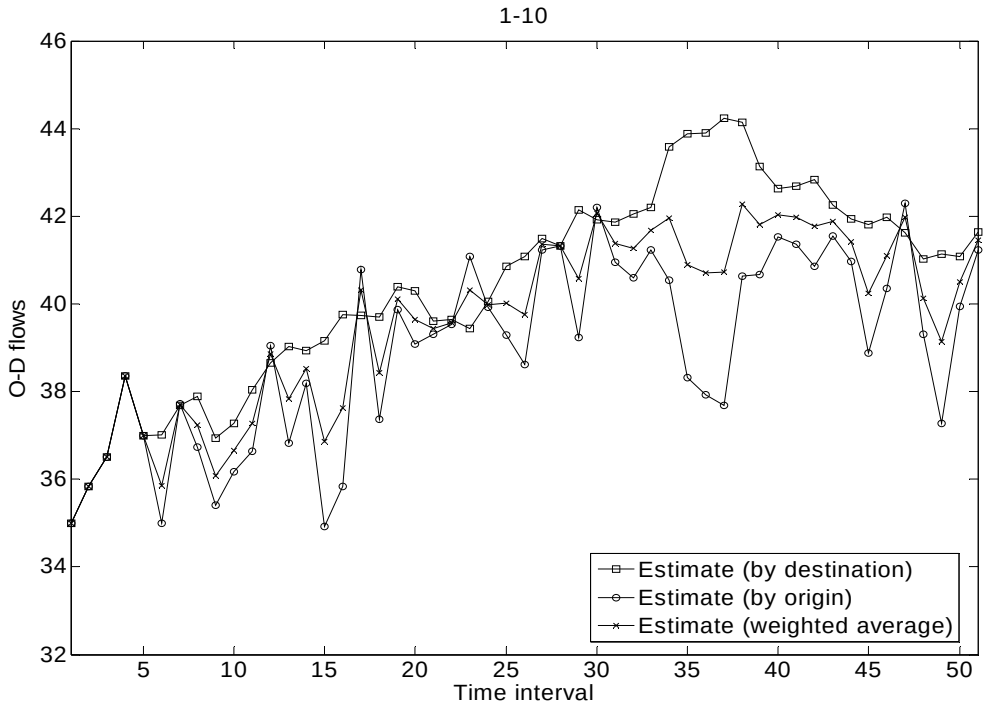


Fig. 3.25 Estimated O-D flow for O-D pair 1-10

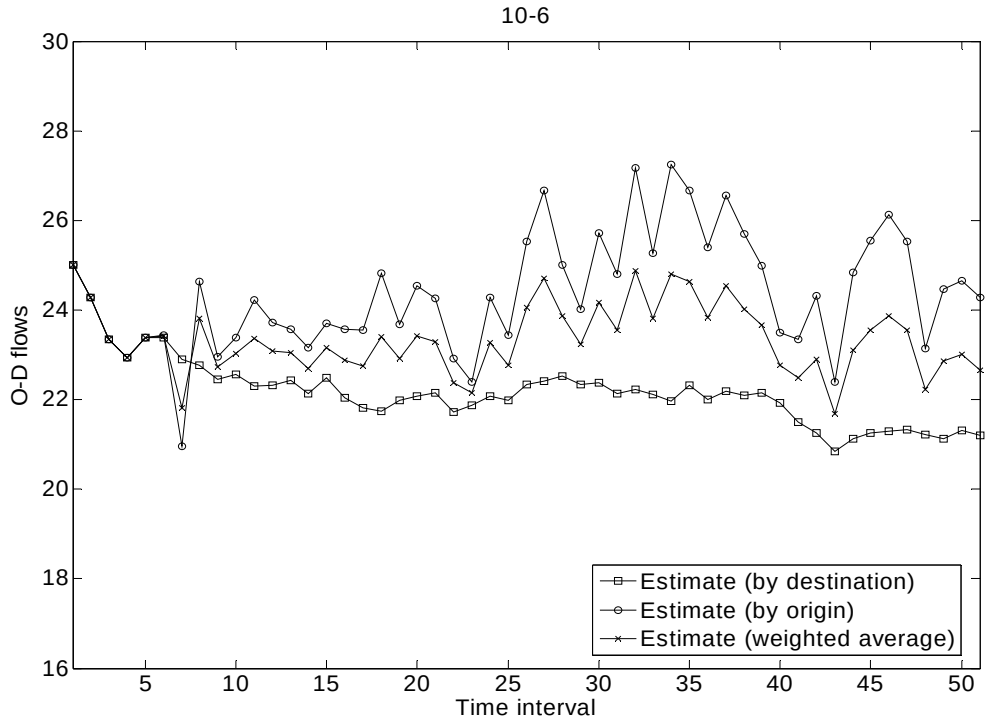


Fig. 3.26 Estimated O-D flow for O-D pair 10-6

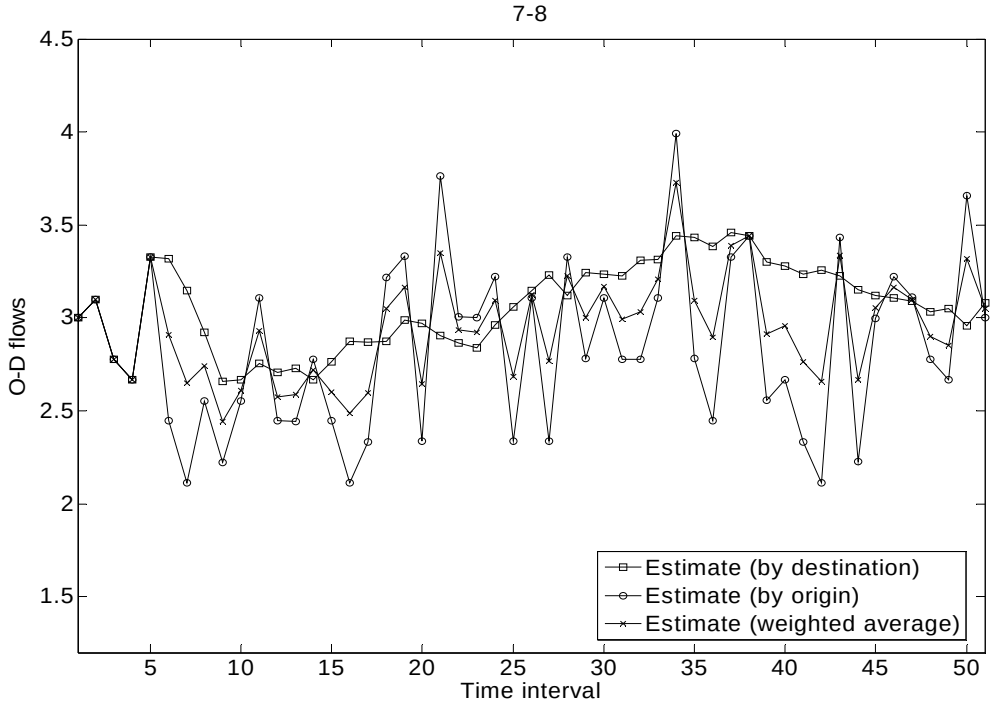


Fig. 3.27 Estimated O-D flow for O-D pair 7-8

3.3.5 Concluding Remarks

We have presented a two-step approach for estimating time-varying O-D flows for actuation-controlled corridors with incomplete information about entering and exiting flows. At the first step, intersection turning movements for each intersection are

estimated, and then used at the second step to construct the measurement equations to infer the corridor O-D flows. The proposed approach has been demonstrated and verified on a hypothetic corridor. The results suggest that the estimator is able to track the patterns of the O-D flows and provide estimates close to the actual values. Moreover, we have applied the O-D estimator to a segment of El Camino Real, San Mateo, CA. The application demonstrates that the estimator is able to readily work with actual field loop data.

The approach can be further enhanced by addressing the following issues:

- Incorporating platoon dispersion in determining the assignment matrix or simultaneously estimating the propagation parameters together with O-D flows using the nonlinear Kalman filtering.
- Developing an efficient method to incorporate mainline through movement information available at upstream intersections.

3.4 Investigation of Dynamic Structure of O-D Demand

3.4.1 Introduction

Previous two sections have presented approaches for real-time estimation of dynamic O-D flows for actuation-controlled intersections and corridors. The technique adopted is Kalman filtering and the fundamental idea is to formulate state-space equations where the state vectors to be estimated are assumed to be a dynamic auto-regressive process. In this manner, information of the structure of the O-D demand is obtained in addition to the measurement equations to estimate the demand without using further *a priori* information (Cremer and Keller, 1987).

This section addresses a fundamental issue in applying the above technique. In Kalman filtering, the state vectors can be (and have been) specified as O-D splits, defined as the percentage of trips generated from an origin to specific destinations (e.g., Cremer and Keller, 1987; Nihan and Davis, 1987; Bell, 1991 and Chang and Wu, 1994), O-D flows (e.g., Okutani, 1987), the deviations of O-D flows from historical data (e.g., Ashok and Ben-Akiva, 1993 and Hu et al., 2001) or the deviation of O-D splits (e.g., Ashok and Ben-Akiva, 2000). The selection of state vector is critical because it results in different model structure that requires different amount of computation effort. More importantly, it determines the performance of the estimator. If the underlying auto-regressive assumption does not replicate reality well, the estimator may not be able to provide reasonable estimates. Interestingly, many of these previous studies make the choice primarily to facilitate the model formulation without enough justification. At the same time, few studies have been done to offer insights how to make the choice.

This section is an empirical investigation on the selection of state vectors in the filtering process, using the traffic data collected from a single intersection in

Gainesville, Florida. The purpose of this investigation is not to determine which state variable should be selected (which seems impossible), but to offer some observations of the O-D demand structures and the estimator performances, and hopefully shed lights on how to choose the most appropriate state vector and the corresponding estimation model. To this aim, we conduct statistic time series analysis to examine the dynamic property of the O-D demand, and compare the estimation results from using different state vectors.

3.4.2 Historical Perspectives

As previously stated, existing dynamic O-D estimators with filtering techniques can be casually categorized into two classes: the closed-network and the open-network oriented approaches, based on the kernel system equations adopted.

For closed networks, all the entry and exit counts are known during all measurement intervals. Naturally, the O-D splits may be considered as state variables, estimated directly using the flow conservations representing the relationship between the real-time exiting counts of a certain destination j and real-time entering counts of all the related origins i:

$$y_j(t) = \sum_i b_{ij}(t) q_i(t) + e_j(t)$$

Here $y_j(t)$ and $q_i(t)$ are the exiting and the entering traffic counts, $b_{ij}(t)$ is the O-D split, the proportion of traffic flows entering at origin i and exiting at destination j, and $e_j(t)$ is the random measurement error. The O-D splits have been assumed to be auto regressive in the pioneering studies, e.g., Cremer and Keller (1987) and Nihan and Davis (1987). Both studies model the time-dependent O-D splits as a first-order auto-regressive (AR1) (ARq denotes the qth-order auto-regressive model) process:

$$b_{ij}(t) = b_{ij}(t-1) + w_{ij}(t)$$

where $w_{ij}(t)$ is a series of white noise terms with zero means and known covariance.

Under additional assumption that the O-D splits are independent among all O-D pairs, i.e. there is no correlation between all the error terms $w_{ij}(t)$, the O-D estimation problem can be decomposed into smaller identification problems, each concerning only the splits related to one specific exit j. Such decomposition increases the computation efficiency by avoiding matrix inversions in updating the Kalman gain.

Attention should be paid to satisfying constraints associated with O-D splits. The split $b_{ij}(t)$ is the proportion of the traffic entering at entry i at time interval t that leaves at exit j. Therefore, the parameters $b_{ij}(t)$ are obviously bounded between zero and one. Moreover, all the O-D splits from a specific entry i in a specific time interval should add up to one. These inequality and equality constraints cause some difficulties when applying Kalman filtering because the basic structure of this algorithm does not allow for

constraints.

How to deal with the constraints is one of the biggest issues for the closed-network approach. Although previous studies have proposed approaches to satisfy those equality and inequality constraints, e.g., Nihan and Davis (1987), Bell (1991) and Li and De Moor (1999), there is no effective algorithm to address both sets of constraints simultaneously. One may think of adjusting the preliminary results according to both sets of constraints iteratively, but this heuristic method will be computationally demanding. Moreover, the existence of the convergence and whether the results are unbiased remain unproved.

For open networks, not all entry and exit counts are known, and the information available is normally link flows recorded by detectors. Under this information pattern, the estimation of O-D flows is regarded as a reverse problem of the dynamic traffic assignment (DTA) problems. The system equation is as follows:

$$\vec{x}(t) = \sum_{r=t-m}^{t} A_r^t \vec{f}(r) + \vec{e}(t)$$

where $\vec{x}(t)$ is the measured link volume vector; m is the maximum number of time intervals needed to travel through the system; $\vec{f}(r)$ is the vector of O-D flows; $\vec{e}(t)$ is the corresponding measured error vector and A_r^t is the assignment matrix models mapping the O-D flows of time interval r to the current link volume. The assignment matrices incorporate the flow propagation and route choice information but are difficult to obtain and have been assumed to be known from exogenous resources such as direct observations or DTA models.

Consequently, the O-D flows rather than the O-D splits are the state vectors to be estimated. Define G_t^r as the transition matrix describing the effect of previous O-D flow $\vec{f}(r)$ on the current O-D flow at t, some researchers assumed the O-D flows follow the more general auto-regressive structure:

$$\vec{f}(t) = \sum_{r=t-p}^{t-1} G_t^r \vec{f}(r) + \vec{w}(t)$$

where p indicates the maximum number of lags and $\vec{w}(t)$ is a series of white noise vectors with zero mean (e.g., Okutani, 1987). This assumption is more realistic than the AR1 model for the O-D splits in the sense that it allows O-D flows from more than one previous period to have correlations with the current O-D flow. And the estimation is only constrained by the non-negativity conditions which, because the magnitude of O-D flows is much greater than the O-D splits, are rarely activated. However, historical data are required to estimate the transition matrices.

Distinctive from the above assumption, Ashok and Ben-Akiva (1993) assumed that rather than O-D flows themselves, the deviations of current O-D flows from historical O-D data are auto-regressive. They argued that the traditional assumption only captures

the temporal interdependencies while the more complicated structure of O-D patterns has been ignored. By incorporating historical data into the estimation, the spatial and other properties of the specific demand patterns may be represented. One more advantage is that the deviations may be better approximated by the auto-regressive model since they can take both negative and positive values. Accordingly, the non-negativity inequality constraints are not necessary. The only situation where the deviations need to be adjusted is when the absolute value of the negative deviation is greater than the historical flow, but this happens very occasionally in the experiments. Therefore, no equality and inequality constraints are imposed on the flow deviation variables. Let the superscription *H* denote the historical counterparts, the state-space equations should be reformulated as follows:

$$\begin{cases} \vec{f}(t) - \vec{f}^{H}(t) = \sum_{r=t-p}^{t-1} G_t^r \left(\vec{f}(r) - \vec{f}^{H}(r) \right) + \vec{w}(t) \\ \vec{x}(t) - \vec{x}^{H}(t) = \sum_{r=t-m}^{t} A_r^t \left(\vec{f}(r) - \vec{f}^{H}(r) \right) + \vec{e}(t) \end{cases}$$

Decomposition of this system is impossible unless the network structure is simple and some certain links are only related to limited O-D flows. Therefore, such approach is more computationally demanding. Besides, the quality of historical data may affect the efficiency of the estimator. Poor historical basis is expected to degrade the estimator's performance.

Any O-D flow can be expressed by the product of trip production at the origin and the O-D split. Ashok and Ben-Akiva (2000) further observed that these two components exhibited different variability with time. The trip production may be highly variable while the O-D splits are relatively stable. Allowing for this differential variability in the estimation process could arguably increase the performance of the estimator. Consequently, two sets of transition equations can be specified as follows:

$$\begin{cases} \vec{d}(t) - \vec{d}^{H}(t) = \sum_{r=t-p}^{t-1} \Phi_t^r \left(\vec{d}(r) - \vec{d}^{H}(r) \right) + \vec{w}_1(t) \\ \vec{b}(t) - \vec{b}^{H}(t) = \sum_{r=t-m}^{t-1} \Psi_t^r \left(\vec{b}(r) - \vec{b}^{H}(r) \right) + \vec{w}_2(t) \end{cases}$$

where $\vec{d}(t)$ is the trip production vector; Φ_t^r is the transition matrix describing the effect of previous trip production on the current production; $\vec{b}(t)$ is the split vector and Ψ_t^r is the corresponding transition matrix; $\vec{w}_1(t)$ and $\vec{w}_2(t)$ are the error vectors and *p* and *m* are the order of the auto-regressive processes.

In summary, the state vectors can be specified as O-D splits, O-D flows, O-D flow deviations and O-D split deviations. Although the choice of state vectors is very critical to the structure, computation complexity and performance of the estimator, there is no previous study to reveal intrinsic structures of the O-D demands and offer insights on

how to determine state vectors.

3.4.3 Empirical Investigation

In the following we examine the traffic data collected from an intersection of 34th Street and University Avenue in Gainesville, Florida (See Fig. 3.28 for the intersection layout and the numbering of the four approaches) to reveal the true structure of the O-D demand at the intersection. To provide more pragmatic comparison, we estimate the O-D demand using each state variable under the simple first-order auto-regressive assumption and then compare the estimate with the true value.

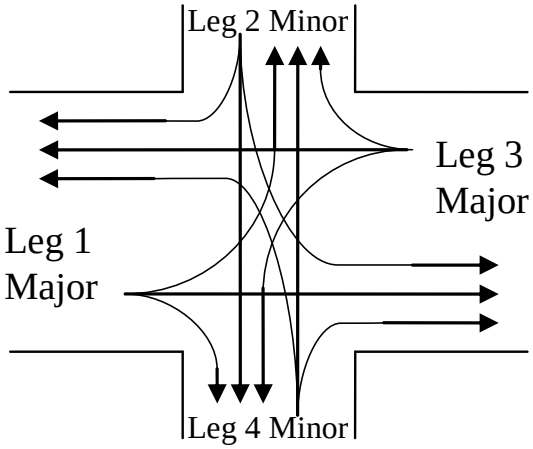


Fig. 3.28. 34th Street-University Avenue intersection
(Major = University Avenue, Minor = 34th Street)

3.4.3.1 Statistic Time Series Analysis

Each turning movement is considered as an independent stochastic process. Since U-turn is prohibited, there are 12 time series available for each state variable, and they are modeled separately. We use the ARIMA module in SPSS version 13.0 to do the time series analysis.

3.4.3.2 O-D Estimators

We estimate the O-D demand using each state variable under the simple first-order auto-regressive assumption. For the state variable of O-D splits, the estimator can be represented as:

$$\begin{cases} \vec{b}(t) = \vec{b}(t-1) + \vec{w}(t) \\ \vec{y}(t) = Q(t)\vec{b}(t) + \vec{e}(t) \end{cases}$$

where $Q(t) = diag(\vec{q}^T(t), \vec{q}^T(t), \vec{q}^T(t), \vec{q}^T(t))$ and other elements denote the same variable in vector form. This system consists of four observation equations and 12 unknowns. Denote $V(t)$ as the covariance matrix of error vector $\vec{w}(t)$ and $R(t)$ the covariance matrix of $\vec{e}(t)$, the unconstrained Kalman filter solution is:

$$\begin{cases} K(t) = P(t)Q^T(t)\left[Q(t)P(k)Q^T(t) + R(t)\right]^{-1} \\ \vec{b}(t) = \vec{b}(t-1) + K(t)\left[\vec{y}(t) - Q(t)\vec{b}(t-1)\right] \\ P(t+1) = \left[I - K(t)Q(t)\right]P(t) + V(t) \end{cases}$$

In actual application, $V(t)$ can be regarded as a parameter of the estimator if unknown. Taking account of the equality constraints in vector form, $D\vec{b}(t) = 1_{4\times 4}$, the original solution should be corrected by:

$$\hat{\vec{b}}(t) = \vec{b}(t) - P(t)D^T\left[DP(t)D^T\right]^{-1}\left[D\vec{b}(k) - 1\right]$$

With the O-D split deviations as the state variable, the estimator can be formulated as:

$$\begin{cases} \vec{b}(t) - \vec{b}^H(t) = \vec{b}(t-1) - \vec{b}^H(t-1) + \vec{w}(t) \\ \vec{y}(t) - \vec{y}^H(t) = Q(t)\left(\vec{b}(t) - \vec{b}^H(t)\right) + \vec{e}(t) \end{cases}$$

The solution can be straightforwardly derived by replacing $\vec{b}(t)$ and $\vec{y}(t)$ in the O-D splits solution by $\vec{b}(t) - \vec{b}^H(t)$ and $\vec{y}(t) - \vec{y}^H(t)$.

To estimate O-D flows directly, the four entries and exits of the intersection are considered as eight virtual links. Consequently, the entering and exiting counts are treated as the link volumes. Therefore, the number of observation equations increases from four to eight in this approach. The flow propagation can be ignored since a cycle is considered one time interval. Arranging these 12 O-D flows in the same order as the O-D splits, the state-space equation becomes:

$$\begin{cases} \vec{f}(t) = \vec{f}(t-1) + \vec{w}(t) \\ \begin{bmatrix} \vec{y}(t) \\ \vec{q}(t) \end{bmatrix} = \begin{bmatrix} H \\ D \end{bmatrix}\vec{f}(t) + \vec{e}(t) \end{cases} \quad \text{where} \quad H = diag\left(\begin{bmatrix} 1_{1\times 3} & 1_{1\times 3} & 1_{1\times 3} & 1_{1\times 3} \end{bmatrix}\right)$$

Similar formulation for the deviations of O-D flows is as follows:

$$\begin{cases} \vec{f}(t) - \vec{f}^H(t) = \vec{f}(t-1) - \vec{f}^H(t-1) + \vec{w}(t) \\ \begin{bmatrix} \vec{y}(t) \\ \vec{q}(t) \end{bmatrix} - \begin{bmatrix} \vec{y}^H(t) \\ \vec{q}^H(t) \end{bmatrix} = \begin{bmatrix} H \\ D \end{bmatrix}\left(\vec{f}(t) - \vec{f}^H(t)\right) + \vec{e}(t) \end{cases}$$

3.4.3.3 Data Description

Actual O-D flows (turning movements) of the 34th Street - University Avenue intersection in Gainesville, Florida were collected cycle by cycle from the videos

recorded in 2001. Since totally different auto-regressive patterns may exist for different times of day, only the off-peak turning movements were counted in order to more likely obtain flows with homogenous O-D patterns. Without loss of generality, one cycle is regarded as a time interval, in which the entering and exiting counts of each approach were aggregated from the turning movements. The O-D split b_{ij} was calculated as the corresponding O-D flow f_{ij} divided by the entering flow of approach i. Limited by the videos available, 46 data points were obtained from 9:00AM to 11:00AM for the first day (July 17, 2001), 27 data points from 9:00AM to 10:00AM and 28 data points from 10:00AM to 11:00AM were collected for the second day (July 18, 2001) and third day (July 20, 2001) respectively.

Several assumptions were made in the following statistical analysis. To examine the O-D flows and splits, it was assumed the auto-regressive pattern remains constant over these three periods so that one can assemble three short time-series data into one longer series to have enough observations. To evaluate the deviations of O-D flows or splits, the first-day data were chosen as the basis while the second- and third-day data were combined together. It was also assumed that all the lane groups were independent. In the estimation part, experiments without the independence assumption were also conducted to show the impact of the covariance matrix of noise vector, $V(t)$, on the estimates.

3.4.4 Empirical Results

3.4.4.1 Time Series Model Specification

Although most of the statistic software has built-in time series analysis modules, the preliminary test of the data and the transformation to stationarity cannot be done automatically. Sequence graphs and histograms were plotted as a reference to check whether the original data are stationary. The sequence graphs indicated no strong trend and periodicity in the series of O-D flows, splits, flow deviations and split deviations. But the histograms for these variables were not symmetrically distributed and were far away from the typical Gaussian marginal distribution. To take account of this effect, first-order differencing was applied to eliminate possible unrevealed complicated trends. The new series were created as: New differenced O-D data (t) = Original O-D data (t) – Original O-D data (t-1), t=2, 3…

Table 3.4 Model Specifications for Differenced O-D Flows

Lane Groups	AR1	AR2	AR3	AR4
Approach 1 Left Turn	-1.03	-0.73	-0.55	-0.30
Approach 1 Right Turn	-0.73	-0.35	--	--
Approach 1 Thru	-0.82	-0.55	-0.28	--
Approach 2 Left Turn	-0.89	-0.67	-0.41	-0.31
Approach 2 Right Turn	-0.59	-0.30	--	--
Approach 2 Thru	-0.71	-0.57	--	--
Approach 3 Left Turn	-0.56	-0.31	--	--
Approach 3 Right Turn	-0.63	-0.44	--	--
Approach 3 Thru	-0.74	-0.32	--	--
Approach 4 Left Turn	-0.87	-0.42	--	--
Approach 4 Right Turn	-0.63	-0.33	--	-0.21
Approach 4 Thru	-0.95	-0.66	-0.27	--

For O-D splits and flows, the new differenced series consist of 100 data points while the deviation series include only 45 data points. Sequence graphs and histograms suggested these differenced variables were more likely to be stationary. Therefore, the statistical analyses were based on these new series. Note that the fact that the difference series is *k*th-order auto-regressive model does not necessarily mean the original data is (*k*+1)th-order model. The interruption item plays a very important role in statistic model estimations.

The coefficients of the best fit models for all the 12 O-D movements for these four state variables are listed in Tables 3.4 to 3.7. Inconsistency between different lane groups is observed. Moreover, the results indicate that for this particular data set, the simple first-order auto-regressive model cannot fit the original data well. More sophisticated auto-regressive models up to the fourth order are needed to describe the O-D structures. Though complicated, it is still feasible to incorporate these transition relationships into the Kalman filtering estimator.

Table 3.5 Model Specifications for Differenced O-D Flow Deviations

Lane Groups	AR1	AR2	AR3
Approach 1 Left Turn	-1.11	-0.82	-0.35
Approach 1 Right Turn	-0.67	-0.41	--
Approach 1 Thru	-0.55	--	--
Approach 2 Left Turn	-0.57	-0.55	-0.36
Approach 2 Right Turn	-0.54	-0.31	--
Approach 2 Thru	-0.73	-0.59	--
Approach 3 Left Turn	-0.54	-0.34	--
Approach 3 Right Turn	-0.62	-0.36	--
Approach 3 Thru	-0.71	--	--
Approach 4 Left Turn	-0.93	-0.39	--
Approach 4 Right Turn	-0.61	-0.32	--
Approach 4 Thru	-0.91	-0.55	--

3.4.4.2 Estimation Results

The four estimators described in Section 3.4.3 were applied to estimate the O-D demand, and the root mean square (RMS) of the error between the estimates and actual O-D flows was calculated as a performance measurement for each estimator. For estimators with state variable of O-D split or deviations, the O-D flows were computed as the production of estimated O-D splits and the entering counts or summation of the estimated deviation and the historical data.

Table 3.6 Model Specifications for Differenced O-D Splits

Lane Groups	AR1	AR2	AR3	AR4
Approach 1 Left Turn	-0.97	-0.63	-0.42	-0.25
Approach 1 Right Turn	-0.71	-0.31	--	--
Approach 1 Thru	-0.86	-0.51	-0.38	-0.20
Approach 2 Left Turn	-0.52	--	--	--
Approach 2 Right Turn	-0.44	--	--	--
Approach 2 Thru	-0.48	--	--	--
Approach 3 Left Turn	-0.77	-0.55	-0.50	-0.28
Approach 3 Right Turn	-0.72	-0.72	-0.39	-0.24
Approach 3 Thru	-0.73	-0.49	-0.29	--
Approach 4 Left Turn	-0.98	-0.45	--	--
Approach 4 Right Turn	-0.58	-0.38	--	--
Approach 4 Thru	-0.75	-0.33	--	--

Different combinations of initial values and the estimator parameter V representing the covariance of the state transition equation were tested. For the O-D splits, the initial value was set as

$$\begin{matrix} 0 & .33 & .33 & .34 \\ .33 & 0 & .33 & .34 \\ .33 & .33 & 0 & .34 \\ .33 & .33 & .34 & 0 \end{matrix}$$

The true O-D splits at the first cycle and the average values over all the time intervals were selected as two other initial scenarios. They are denoted respectively as Basic, First Cycle and Mean initial conditions. Basic initial condition for O-D flows arbitrarily set all the components equal to one. V matrix was set as either the estimated covariance of the sample O-D data (Sample Covariance) or the identity matrix (Identity). Note that under the Sample Covariance condition, the O-D data of different lane groups cannot be regarded as independent since the estimated sample covariance would rarely be a diagonal matrix.

Table 3.7 Model Specifications for Differenced O-D Split Deviations

Lane Groups	AR1	AR2
Approach 1 Left Turn	-0.96	-0.45
Approach 1 Right Turn	-0.51	--
Approach 1 Thru	-0.55	--
Approach 2 Left Turn	-0.23	-0.54
Approach 2 Right Turn	-0.51	-0.32
Approach 2 Thru	-0.53	-0.66
Approach 3 Left Turn	-0.57	--
Approach 3 Right Turn	-0.62	-0.49
Approach 3 Thru	-0.55	--
Approach 4 Left Turn	-0.98	-0.43
Approach 4 Right Turn	-0.53	-0.54
Approach 4 Thru	-0.53	--

The estimation results are presented in Tables 3.8 and 3.9. Note that the second- and third-day O-D demands were estimated separately if deviations are state variables. Reported RMS of the corresponding two estimators (the O-D split deviation estimator and the O-D flow deviation estimator) is the average RMS of these two sets of estimation. It can be observed from these tables that in this particular case study, assuming O-D flows to be first-order auto-regressive leads to the best estimation under the same initial conditions and estimator parameters. The estimators focusing on the O-D data themselves outperformed their counterparts of deviations. See Figs 3.29 and 3.30 for a visualized comparison of O-D flow estimator and flow deviation estimator. The latter corresponds to the last 27 intervals in the former figure.

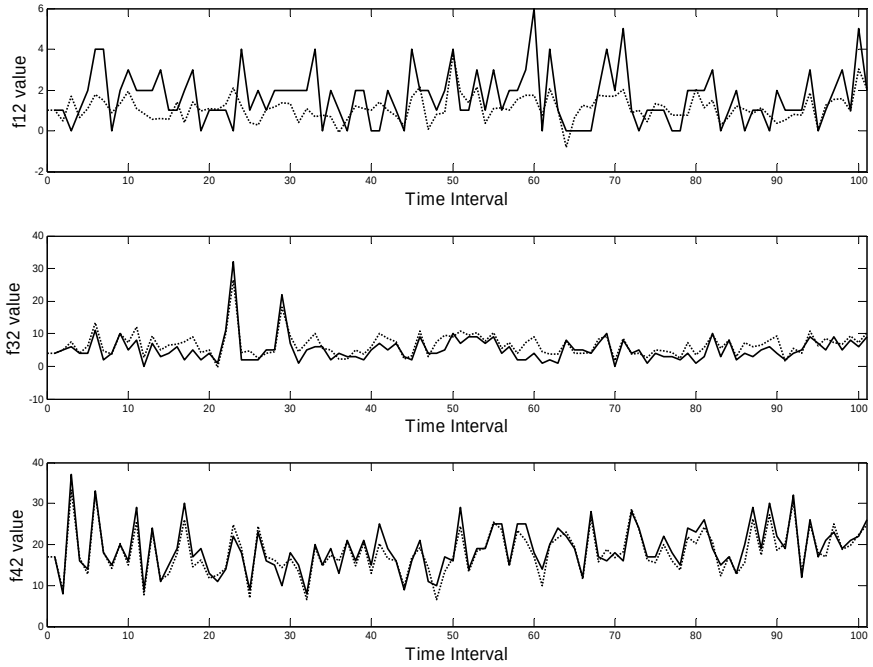


Fig. 3.29. Results of O-D flows destined to Leg 2 estimated by O-D flow estimator

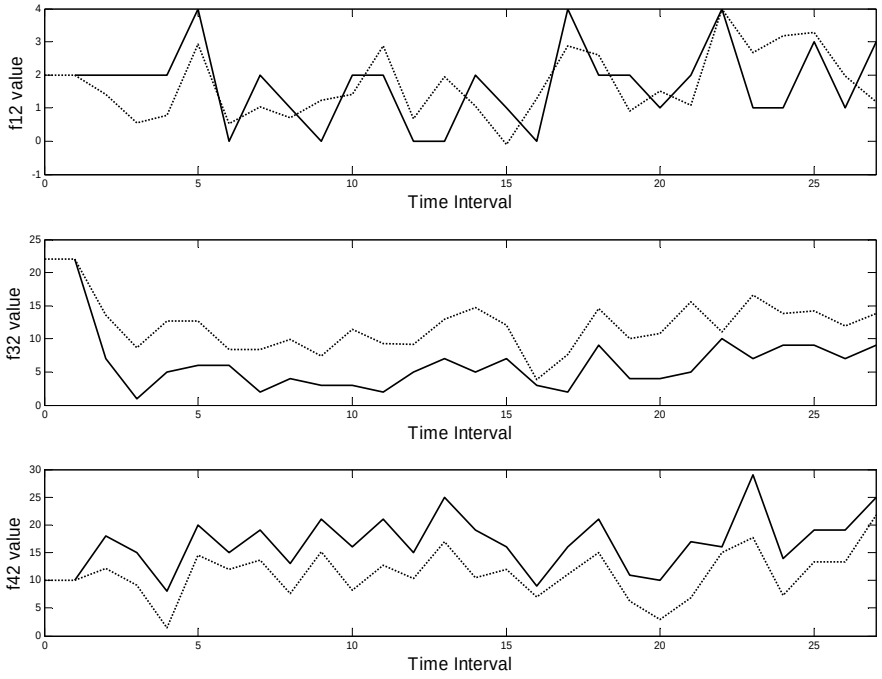


Fig. 3.30. Estimates of the third-day O-D flows destined to Leg 2 with O-D flow deviation estimator

Impact of the transition function was further explored by estimating the O-D flows using the estimated AR1 of the differenced series. The transition matrix *G* shall be

introduced to the Kalman filtering. The sate-space equations are:

$$\begin{cases} \vec{f}(t)-\vec{f}(t-1)=G\left(\vec{f}(t-1)-\vec{f}(t-2)\right)+\vec{w}(t) \\ \begin{bmatrix}\vec{y}(t)\\ \vec{q}(t)\end{bmatrix}-\begin{bmatrix}\vec{y}(t-1)\\ \vec{q}(t-1)\end{bmatrix}=\begin{bmatrix}H\\ D\end{bmatrix}\left(\vec{f}(t-1)-\vec{f}(t-2)\right)+\vec{e}(t) \end{cases}$$

Solution to this model is:

$$\begin{cases} K(t)=P(t)\begin{bmatrix}H\\ D\end{bmatrix}^T\left[\begin{bmatrix}H\\ D\end{bmatrix}P(k)\begin{bmatrix}H\\ D\end{bmatrix}^T+R(t)\right]^{-1} \\ \vec{f}(t)-\vec{f}(t-1)=G\left(\vec{f}(t-1)-\vec{f}(t-2)\right)+K(t)\left[\begin{bmatrix}\vec{y}(t)\\ \vec{q}(t)\end{bmatrix}-\begin{bmatrix}\vec{y}(t-1)\\ \vec{q}(t-1)\end{bmatrix}-\begin{bmatrix}H\\ D\end{bmatrix}G\left(\vec{f}(t-1)-\vec{f}(t-2)\right)\right] \\ P(t+1)=G\left[I-K(t)\begin{bmatrix}H\\ D\end{bmatrix}\right]P(t)G^T+V(t) \end{cases}$$

The RMS under Mean initial and Sample Covariance condition is 1.35, which is 211% more than that of the original O-D flow estimator under the same conditions. This result seems not consistent with that of the time series analysis. A plausible explanation is that the time series analysis was conducted after the subjective transformation process to stationarity and does not necessarily reveal the true structure. On the other hand, it may suggest that reasonable approximation of the O-D patterns is able to lead to acceptable results.

The experiments also verified that the initial value does not have a primary impact on the performance of the estimator and the effect of covariance matrix V seems even more negligible. On the other hand, performance of the O-D split and split deviation estimators may improve when they were constrained.

3.4.5 Conclusion

We have conducted an empirical analysis of the O-D demand structures and examined their impacts on the O-D estimation using the data from the 34th Street-University Avenue intersection in Gainesville, Florida in 2001. Auto-regressive models of the differenced O-D flows, flow deviations, split and split deviations were estimated, and the best fit models range from AR1 to AR4. The comparison of estimators with different state variables suggested that the estimator with state variable of O-D flow outperforms the others in this particular case. We fully recognize that O-D patterns would be site-dependent, and the results of this case study should not be generalized. However, our empirical investigation does offer the following observations, which may be of use for future studies and practices:

- The idea of deviations, more specifically, the deviations of O-D flows or splits do not necessarily better represent the O-D demand patterns. If historical data are available, statistical analysis may be conducted to reveal the intrinsic structure of the O-D demand;
- Consequently, the estimators with state variables of deviations do not necessarily produce better O-D estimates;
- Demands or flows at different O-D pairs may possess different structures, which are very often not first-order auto regressive;
- Incorporating all of these “true” structures into the Kalman filtering makes the model formulation very complicated. On the other hand, the simple first-order auto regressive assumption produces acceptable results in our empirical experiments and previous studies. Therefore, unless there are sufficient O-D data that suggest otherwise, it may be sensible to simply use the state variable of O-D flows or splits and assume that they are first-order auto regressive.

Table 3.8 RMS of O-D Estimator of Splits and Split Deviations

Estimator	Initial Values			Parameter V		Constraints	RMS
	Basic	First Cycle	Mean	Identity	Sample Covariance		
O-D Splits	X			X		No	2.43
	X			X		Equality	2.17
	X			X		Inequality	2.12
	X				X	No	2.27
	X				X	Equality	2.01
Deviation of O-D Splits	X			X		No	2.85
	X			X		Equality	2.55
	X				X	No	2.77
	X				X	Equality	2.32
		X		X		No	2.74
		X		X		Equality	2.41
		X			X	No	2.72
		X			X	Equality	2.31
			X	X		No	2.73
			X	X		Equality	2.38
			X		X	No	2.74
			X		X	Equality	2.36

Table 3.9 RMS of O-D Estimator of Splits and Split Deviations

Estimator	Initial Values			Parameter V		RMS
	Basic	First Cycle	Mean	Identity	Sample Covariance	
O-D Flows	X			X		2.30
	X				X	1.80
		X		X		1.13
		X			X	0.58
			X	X		0.78
			X		X	0.64
Deviation of O-D Flows	X			X		2.40
	X				X	1.92
		X		X		1.60
		X			X	1.05
			X	X		1.42
			X		X	0.88

References

Antoniou, C., Ben-Akiva, M.E. and Koutsopoulos, H.N. (2004) Incorporating automated vehicle identification data into origin-destination estimation. Transportation Research Record No. 1882, 37-44

Asakura, Y., Hato, E. and Kashiwadani, M. (2000) Origin-destination matrices estimation model using automatic vehicle identification data and its application to the Han-Shin expressway network. Transportation, 27, 419-438.

Ashok, K. and Ben-Akiva, M.E. (1993) Dynamic origin-destination matrix estimation and prediction for real-time traffic management system. In Proceedings of the 12th International Symposium on Transportation and Traffic Theory, ed. C. F. Daganzo, 465-484, Berkeley. Elsevier, NY.

Ashok, K. and Ben-Akiva, M.E. (2000) Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. Transportation Science Vol. 34, 21-36

Ashok, K; Ben-Akiva, ME (2002) Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation Science, 36(2), 184-198.

Bell, M.G.H. (1991) The real time estimation of origin-destination flows in the presence of platoon dispersion. Transportation Research, 25B, 115-125.

Bell, M.G.H., Lam, W.H.K and Iida, Y. (1996) A time-dependent multi-class path flow estimator. In Proceedings of the 13th International Symposium on Transportation and Traffic Theory, 173-193, Elsevier Science.

Bierlaire, M. and Crittin, F. (2004) An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Operations Research 52, 116-127

Cascetta, E., Inaudi, D., and Marquis, G. (1993) Dynamic estimators of origin-destination matrices using traffic accounts. Transportation Science, 27, 363-373.

Chang, G.L. and Tao, X. (1996) Estimation of dynamic network O-D distributions for urban networks. In Proceedings of the 13th International Symposium on Transportation and Traffic Theory, 1-20, Elsevier Science.

Chang, G.L. and Tao, X. (1999) An integrated model for estimating time-varying network origin-destination distributions. Transportation Research, 33A, 381-399.

Chang, G.L. and Wu, J. (1994) Recursive estimation of time-varying O-D flows from traffic counts in freeway corridors. Transportation Research, 28B, 437-455.

Chen, A., Chootinan, P., Recker, W. and Zhang, H.M. (2004) Development of a path flow estimator for deriving steady-state and time-dependent origin-destination trip tables. California PATH research report, UCB-ITS-PRR-2004-29, September 2004.

Crème, M. and Keller, H. (1981) Dynamic identification of flows from traffic counts at complex intersection. Proceedings of the 8th International Symposium on Transportation and Traffic Theory, V.F. Hurdle et al., (eds.), University of Toronto Press, Toronto.

Crème, M. and Keller, H. (1987) A new class of dynamic methods for the identification of origin-destination flows. Transportation Research, 21B, 117-132.

Dixon, M. P. and Rilett, L. R. (2005) Population origin-destination estimation using automatic vehicle identification and volume data. Journal of Transportation Engineering, 131(2), 75-82

Eisenman, S. M. and List, G. F. (2004) Using Probe Data to Estimate OD Matrices. IEEE Intelligent Transportation Systems Conference, Washington, D.C., USA, 2004.

FHWA (2004) http://www.dynamictrafficassignment.org, Federal Highway Administration Dynamic Traffic Assignment Research Program.

Hu, S.R., Madanat, S., Krogmeier, J.V. and Peeta, S. (2001) Estimation of dynamic assignment matrices and OD demands using adaptive Kalman filtering. ITS Journal, Vol.6, 281-300.

Kwon, J. and Varaiya, P. (2005) Real-time estimation of O-D matrices with partial trajectories from electronic toll collection tag data. 84th Annual Meeting of Transportation Research Board, Washington, D.C., USA, 2005

Li, B., Moor, B.D. (1999) Recursive estimation based on the equality-constrained optimization for intersection origin-destination matrices. Transportation Research 33B, 203-214

Li, B., Moor, B.D. (2002) Dynamic identification of origin-destination matrices in the presence of incomplete observations. Transportation Research, 36B, 37-57.

Madanat, S.M., Hu, S.R., and Krogmeier, J (1996) Dynamic estimation and prediction of freeway O-D matrices with route switching considerations and time-dependent model parameters. Transportation Research Record 1537, 98-105.

Nihan, N.L. and Davis, G.A. (1987) Recursive estimation of orgin-destination matrices from input/output counts. Transportation Research, 21B, 149-163.

Okutani, I. (1987) The Kalman filtering approaches in some transportation and traffic problems. Proceedings of the 10th International Symposium on Transportation and Traffic Theory, N.H. Gartner and N.H.M. Wilson (eds.), Elsevier, New York, 397-416.

Paige, C.C and Saunders, M. A. (1982) LSQR: An algorithm for sparse linear equations and sparse least squares, ACM TOMS 8(1), 43-71.

Sherali, H.D. and Park, T. (2001) Estimation of dynamic origin-destination trip tables for a general network. Transportation Research, 35B, 217-235.

Simon, D. and Chia, T.L. (2002) Kalman Filtering with State Equality Constraints. IEEE Transactions on Aerospace and Electronic Systems, Vol. 38, 128-136

Sorenson, H.W. (1970) Least-squares estimation: from Gauss to Kalman. IEEE Spectrum, Vol. 7, 63-68

Willumsen, L.G. (1984) Estimating time-dependent trip matrices from traffic counts. Proceedings of the 9th International Symposium on Transportation and Traffic Theory, the Netherlands, Delft University.

Van Der Zijpp, N. (1996) Dynamic origin-destination matrix estimation on motorway networks. Ph.D. Dissertation. Delft University of Technology, Delft, Netherlands.

Wu, J. and Chang, G.L. (1996) Estimation of dynamic network O-D with screenline flows. Transportation Research, 30B, 277-290.

Zhou, X. (2004) Dynamic origin-destination demand estimation and prediction for off-line and on-line dynamic traffic assignment application. Ph.D. Dissertation. University of Maryland, College Park, MD.

CHAPTER 4

DEVELOPMENT OF A PRACTICAL COMPUTER TOOL FOR DYNAMIC ORIGIN-DESTINATION MATRICES ESTIMATION

Prepared by:

Meng Li

California PATH

University of California, Berkeley

The objective of this sub-task is to implement the models and algorithms developed in the previous task and develop a tool with a user-friendly interface to allow practitioners to apply these developed models and algorithms.

The developed tool is designed to estimate the dynamic OD matrices for a period of interest when traffic flow is in a steady state and for a linear network where loop detector count information is available. In this study, the loop detector count information is assumed to come from the Freeway Performance Measurement System (PeMS). The general framework is shown in Figure 4.1. It is composed of four subsystems (Interface, Software, Network ID, and PeMS).

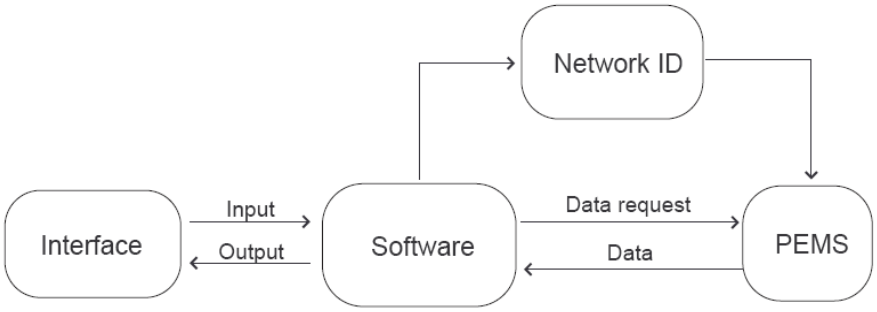


Figure 4.1 General Framework

The simple user interface is shown in Figure 4.2. The first step after loading the software is to define the network. Practitioners can choose from “Interstate-80 Westbound at Berkeley” which is the testbed selected by this study as shown in Figure 4.3, or “Customized network using PeMS system”. For the customized network, two historical OD matrices or historical counts for at least two days are required as the input.

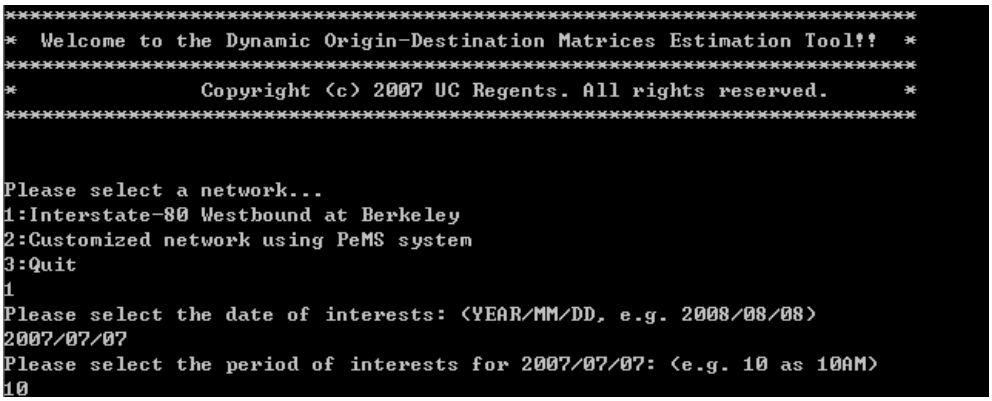


Figure 4.2 Simple User Interface

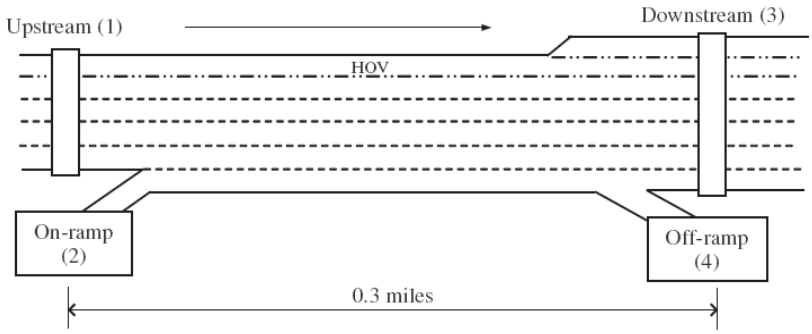


Figure 4.3 Testbed of I-80W at Berkeley

The next step for users is to select the date and period of interests for the estimation. With the given selections, the software will read the data files, which contains 30-second loop detector count information from PeMS.

After obtaining all the required inputs with deviations (i.e. A, V, F, W, and y_k), the software will run the developed methodology to estimate the OD matrix.

It is important to note that the software has to store the estimated OD matrix, because it will be used as the second historical OD matrix to compute the matrix F for the next time. That is, once the period ends, the software will use the difference (or deviation) between the just-estimated OD matrix and the OD matrix estimated the day before to compute a new matrix F. In the future, the software could be continuously running so the historical OD matrix could be self-learned and accordingly updated.

The outputs of the software are presented by the bottom part of Figure 4.4. With the given period of interests, the software calculates a new OD matrix for each 30-second interval. Users can choose "1:Display estimated OD" or "2:Display model performance". If "1" is selected, users can further select an interval for display. The estimated OD flow for each OD pair will be displayed. If "2" is selected, the estimated OD will be compared with the given real OD matrix. Then the root mean square error (RMS) and the root mean square error normalized (RMSN) will be calculated and displayed. It is note that the function to display the model performance is only available for the given period of time at the I-80W testbed, because the special data has only been collected using Berkeley Highway Lab (BHL)'s cameras installed on the roof of a building located next to the testbed. The real OD matrix was then extracted from the video recordings. This process is time consuming.

Figure 4.4 Software Outputs

The current user interface is still very simple and the functions are very limited. Therefore, the user interface, some of the data collection and processing procedures, and more convenient functions will be further developed and debugged.