A NON-CONTINUUM APPROACH TO OBTAIN A MACROSCOPIC MODEL FOR THE FLOW OF TRAFFIC

A Dissertation

by

VIPIN TYAGI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2007

Major Subject: Mechanical Engineering

A NON-CONTINUUM APPROACH TO OBTAIN A MACROSCOPIC MODEL FOR THE FLOW OF TRAFFIC

A Dissertation

by

VIPIN TYAGI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Darbha Swaroop
	K. R. Rajagopal
Committee Members,	Kevin Balke
	Gene Hawkins
	Reza Langari
Head of Department,	Dennis L. O' Neal

May 2007

Major Subject: Mechanical Engineering

ABSTRACT

A Non-Continuum Approach to Obtain a Macroscopic Model for the Flow of Traffic. (May 2007)

Vipin Tyagi, B. Tech., IIT Bombay, India;
M. Tech., IIT Bombay, India

Co–Chairs of Advisory Committee: Dr. Darbha Swaroop
Dr. K. R. Rajagopal

Existing macroscopic models for the flow of traffic treat traffic as a continuum or employ techniques similar to those used in the kinetic theory of gases. Spurious two-way propagation of disturbances that are physically unacceptable are predicted by continuum models for the flow of traffic. The number of vehicles in a typical section of a freeway does not justify traffic being treated as a continuum. It is also important to recognize that the basic premises of kinetic theory are not appropriate for the flow of traffic. A model for the flow of traffic that does not treat traffic as a continuum or use notions from kinetic theory is developed in this dissertation and corroborated with traffic data collected from the sensors deployed on US 183 freeway in Austin, Texas, USA.

The flow of traffic exhibits distinct characteristics under different conditions and reflects the congestion during peak hours and relatively free motion during off-peak hours. This requires one to use different governing equations to describe the diverse traffic characteristics, namely the different traffic flow regimes of response. Such an approach has been followed in this dissertation. An observer based on extended Kalman filtering technique has been utilized for the purpose of estimating the traffic state. Historical traffic data has been used for model calibration. The estimated model parameters have consistent values for different traffic conditions. These estimated model parameters are then subsequently used for estimation of the state of traffic in real-time.

A short-term traffic state forecasting approach, based on the non-continuum traffic model, which incorporates weighted historical and real-time traffic information has been developed. A methodology for predicting trip travel time based on this approach has also been developed. Ten and fifteen minute predictions for traffic state and trip travel time seem to agree well with the traffic data collected on US 183.

To my Parents, Brother and Gurus

ACKNOWLEDGMENTS

I would like to thank the following individuals, who gave me the spirit, knowledge and insight at various stages of this research.

My greatest appreciation and thanks goes to both my academic advisors, Professor Swaroop Darbha and Professor K. R. Rajagopal for being a constant source of inspiration, support and motivation. I consider myself very fortunate for being able to work under their supervision. I have learned much from them, both academic and otherwise, and for this I am indebted.

With great pleasure I extend my gratitude to Professor Gene Hawkins, Dr. Kevin Balke and Professor Reza Langari for kindly agreeing to be members of my dissertation advisory committee. I would also like to thank Professor Bryan Rasmussen for kindly substituting during both my proposal and final defense. I thank them all for their many suggestions that helped enhance the quality of this work.

I would like to thank my research group mates, Sai Krishna Yadlapalli, Waqar Malik, Sin Cheon Kang and Sandeep Dhar for the quality time I spent with them while discussing research and other mundane interesting things in life. They are very much appreciated.

I would like to extend my deepest gratitude and appreciation to my parents and my younger brother for their relentless support and unselfish love. A very special thanks to my very dear friend Sarika Tyagi for her patience, constant encouragement and motivation during the last three years.

I would also like to thank all my friends and fellow students, too numerous to list, who made my Texas A&M experience pleasant. Thank you to the administrative staff in the Department of Mechanical Engineering at Texas A&M University for their help and assistance.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

The interest in the subject of road traffic is very old and has attracted a large group of investigators and theoreticians from diverse backgrounds who have proposed various mathematical models to explain the observed phenomena in road traffic. As early as in 1934, Kinzer [1] drew an analogy between road traffic and electrical networks. In 1935, Greenshields [2] made an experimental study of traffic by measuring actual flows (vehicles per hour) and speeds of observed vehicles. He plotted the traffic speed against traffic density (number of vehicles per mile) for one-lane traffic and fitted the points by a straight line. In the 1950s and the 1960s, there were considerable publications, especially in the journal of operations research. These papers introduced the relationship between the traffic flow and traffic density, formally known as the "Fundamental diagram of road traffic" (a term coined by Haight [3] in 1963). The motivation for such studies was put into perspective by Greenberg in 1959 [4]: "The volume of vehicular traffic in the past several years has rapidly outstripped the capacities of the nation's highways. It has become increasingly necessary to understand the dynamics of traffic flow and obtain a mathematical description of the process. This is especially true for extremely high traffic when the roadway must perform at its peak."

More than forty five years later, due to increased demands for mobility, the volume of vehicular traffic has increased to such an extent that cities like New York, Los Angeles and San Francisco (to name a few) suffer from heavy traffic congestion around the clock. It has been reported [5], [6] that in USA in 2003, the eighty-five

The journal model is *IEEE Transactions on Automatic Control.*

largest metropolitan areas experienced 3.7 billion vehicle-hours of delay, resulting in 2.3 billion gallons in wasted fuel and $63 billion in lost productivity due to congestion. It is estimated that roughly half of these losses are due to recurring congestion, caused by recurring demands that exist virtually every day, where *road use* exceeds existing capacity.

Because it is not always publicly acceptable and financially affordable to expand the existing freeway infrastructure, more efficient means of managing freeway operations are required to tackle the problem of increasing roadway traffic congestion [7]. Increasing attention is being paid to Intelligent Transportation Systems (ITS) as a means of alleviating urban and suburban congestion. ITS is an umbrella term that encompasses a variety of advanced technologies in the areas of communication, computers, information display, road infrastructure, and traffic control systems. Important applications of ITS include management of freeway traffic and traveller information systems[1]. Freeway management and traveller management systems attempt to improve capacity utilization of the freeways in real-time. Some typical measures include ramp control (includes ramp metering, closure, and priority access), lane management (includes high occupancy vehicle (HOV) lanes, lane control, variable speed limits etc.), information dissemination (example, dynamic message signs), providing both pre-trip (example, provided via internet web-sites and specialized telephone services) and en-route (example, provided via wireless services or radio) information to travellers. A desirable and important feature of such systems is the ability to *predict* future traffic conditions, the rationale being that without projection of traffic conditions into the future, control or route guidance strategies are very likely to be irrelevant and outdated by the time they take effect.

[1] http://www.itsoverview.its.dot.gov

In the above mentioned context, thus it becomes imperative to have a good understanding of freeway traffic dynamics. This in turn requires that we should be able to abstract from a complex system a simplified mathematical representation - a *model* - which is able to describe a specific phenomenon[2], observed in the system. In the context of *traffic*, we shall call these mathematical models as *traffic flow models.* Thus, there is a pronounced need for construction of traffic flow models which are able to describe the various dynamic traffic flow phenomena and can be computer coded for various traffic engineering tasks (example: simulation, planning and traffic flow control strategy design).

We can broadly classify the traffic flow models into two categories according to the level of detail with which they describe the flow of traffic. Some models, known as *microscopic traffic flow models*, offer high-level of detail in their description for traffic by considering the space-time behavior of individual vehicles (or drivers) under the influence of other vehicles in their proximity. *Macroscopic* models on the other hand, attempt to describe the traffic at a low-level of detail by describing it from the viewpoint of collective traffic behavior.

For different freeway management and traveller information management system applications, estimation and prediction of traffic state is of utmost importance. Macroscopic traffic flow models are very useful in this context since they are able to offer a collective traffic behavior description. In this disseration, we focus on a macroscopic traffic flow modeling approach, proposed by Darbha and Rajagopal [8], and its utilization in estimation and prediction of traffic state. We now define the problem and present the dissertation objectives.

[2] Or various observed phenomena.

A. Problem definition and dissertation objectives

The problem addressed in this dissertation focuses on the corroboration of the macroscopic traffic flow modeling approach proposed by Darbha and Rajagopal [8]. The main focus in the corroboration procedure is to estimate (and predict) the traffic state for real traffic flow case studies. We classify the available traffic data as *historical* and *real-time*[3]. Corresponding to these two classes of traffic data, we distinguish between two types of estimation problems. The first, which is essentially an *offline* problem, involves estimation of the traffic state from the historical data and its utilization in calibrating the traffic flow model parameters. The *real-time* estimation problem on the other hand, involves the estimation of the traffic state in real-time. An additional issue of relevance and great importance in the real-time problem is the *prediction* of the traffic state for freeway management and traveller information management applications. More rigorously, we define the problem as follows:

1. *Offline estimation:* Given a time series of historical traffic data, estimate the traffic state and use it for calibrating the proposed traffic model.

2. *Real-time estimation and prediction:* Given the traffic data in the time interval k and the historical data time series,

   (a) estimate the traffic state for the time interval k, and

   (b) predict the traffic state for future time intervals.

The main objective of this dissertation is to corroborate the macroscopic traffic flow modeling approach proposed by Darbha and Rajagopal [8]. To that end, the objective is to develop a methodology to estimate and predict the traffic state in real-time. The

[3] We discuss about the available traffic data in Chapter III.

methodology should be robust enough to handle traffic data collected from the sensors deployed on the freeways and be capable of estimating and predicting the traffic state for relatively long lengths of the freeway. The second objective of the dissertation is to develop an algorithm for predicting trip travel times for ITS applications.

B. Literature review of traffic flow modeling approaches

Research in the field of traffic flow modeling has been very active since the 1950s. Investigators and researchers from diverse backgrounds in mathematics, physics, engineering and operations research have contributed to the literature of traffic flow modeling. This has resulted in a broad scope of models which attempt to describe different aspects of traffic flow operations[4]. However, it is has been argued that the traffic flow theory is unlikely to reach the descriptive accuracy attained in other domains of science like thermodynamics for example [9], [10]. This is because, the only accurate physical law in traffic flow is the balance of number of vehicles. There are other difficulties which hamper in attaining a detailed *accurate* description of traffic. For example, it is far from clear how to model human factors such as the driver's psychological disposition, the complex consequences of weather conditions, disturbances to the flow of traffic due to road work, a major accident on the roadway, sudden influx of traffic due to a sporting event ending, etc. Although human and other natural elements such as weather involved in driving makes the problem of modeling complex, everyday experience suggests that the aggregate driving behavior is predictable within reasonable bounds and enables us to estimate the travel times with reasonable consistency so that we can plan our driving to and from work and at other instances with some degree of certainty. For this reason, development of traffic flow models

[4] Probably more than hundred different traffic flow models have been suggested [9].

continue to be pursued.

1. Traffic flow modeling approaches

The general modeling approaches which aim to describe the observed behavior of real systems are either physics-based (example, Newtonian physics applied to develop a model of a mechanical system) or try to fit available input/output real data to generic mathematical models (example, autoregressive integrated moving average models (ARIMA)[5], neural networks). There are also some approaches which follow an intermediate path, in a sense that these approaches attempt to formulate a basic mathematical model first (via physical reasoning and/or adequate assumptions and idealizations) and then try to fit a specific structure to the real data. Most of the traffic flow models suggested so far in the literature try to fit the available data to generic mathematical models or follow the intermediate approach. We now discuss the various traffic flow modeling approaches based on the level of detail by which they attempt to describe the behavior of traffic.

a. Microscopic models

Microscopic traffic flow models typically consider a string of vehicles following each other in a single lane. They attempt to describe both the space-time behavior as well the interactions of individual vehicles (and drivers). The interactions between the vehicles are described through a proposed or an assumed *vehicle-following* rule[6].

[5] We discuss ARIMA models in more detail in Appendix B.

[6] There also has been some research done on microscopic traffic flow models which attempt to describe the *lane-changing* behavior of drivers. This typically is done by modeling the gap acceptance behavior of the drivers [11], [12]. For the purposes of this dissertation, we will restrict our focus to this vehicle following behavior for single lane traffic when we discuss both microscopic as well as macroscopic models.

Most of the microscopic models attempt to describe the vehicle following behavior by assuming a form of the response (either braking or acceleration) of the following vehicle. This can be done, for example, by prescribing a *following distance* (or distance headway)[7]. By following distance we refer to the distance between the rear and front bumpers of the preceding and the following vehicles respectively. Fig. 1 shows a schematic of a string of vehicles following each other. We index the vehicles in an increasing order as we traverse the string upstream. Thus, a vehicle indexed as n follows a vehicle indexed as $n-1$. We now briefly discuss the various microscopic traffic flow models that have been proposed.

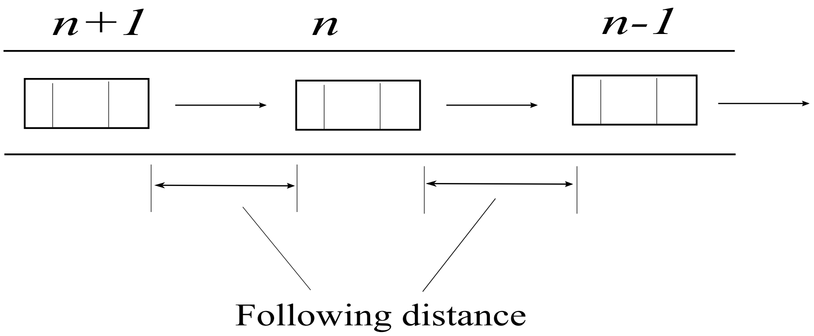

Fig. 1. Schematic of the vehicle-following: vehicle n following vehicle $n-1$

The models proposed by Reuschel[8] and Pipes [13] seem to be the earliest microscopic models for the flow of traffic. They hypothesized that each driver maintains a following distance proportional to the speed of their vehicle plus a constant distance headway at standstill. In the words of Pipes [13], "A good rule for following another vehicle at a safe distance is to allow yourself at least the length of a car between you and the vehicle ahead for every ten miles an hour of speed at which you are traveling".

[7] We have used "following distance" and "distance headway' interchangeably in this dissertation.

[8] As mentioned in [9].

Using this rule, we can define the following distance D_n of the vehicle n, with respect to its preceding vehicle $n-1$ as follows:

$$D_n(v) = L_n(1 + v_n/10) \tag{1.1}$$

where L_n and v_n denote the length and speed of the n^{th} vehicle. A similar approach was also proposed by Forbes et al. [14]. Leutzbach [15] proposed a more refined model by considering the contribution of *driver reaction time* and the *braking distance* in the distance headway as

$$D_n(v) = L_n + Tv_n + \frac{v_n^2}{2\mu} \tag{1.2}$$

where T is the total reaction time and μ is the maximum deceleration possible. Leutzbach assumed that drivers consider braking distances large enough to permit them to brake to a stop without causing a rear-end collision with the preceding vehicle if the latter comes to a stop instantaneously. Jepsen [16] proposed a model in which he also considers a *speed risk* factor in prescribing the following distance as follows

$$D_n(v) = (L_n + d_{min}) + v_n(T + v_nF) \tag{1.3}$$

where d_{min} is some constant minimal distance between vehicles at standstill, F is the speed risk factor and T is over all reaction time. According to Jepsen, the speed risk factor stems from the observation that drivers aim not only to prevent rear end collisions but also to minimize the potential damage or injuries due to a collision, and are aware that in this respect their velocity is an important factor. He thus proposed that drivers increase their time headway by some factor - the speed risk factor - linear to v_n[9]. Gipps [17] has also proposed a model in which limits were imposed

[9] The time headway is defined by the difference in passage times of two successive vehicles.

on the performances of the vehicle and the driver. He then used these limits to calculate a safe distance with respect to the preceding vehicle. Dijker et al. [18] have discussed some empirical findings on vehicle-following behavior in congested traffic flow conditions.

There are also some microscopic models which attempt to model the vehicle following on the assumption that the drivers try to conform to the behavior of their preceding vehicle. In general, they assume that the response is a function of the *sensitivity* of the driver. They also assume that the *stimulus* driving this response is the speed difference between the preceding and the following vehicle. In general, they model the response of the following vehicle, delayed by some overall reaction time T. Let $x_n(t)$ denote the position of the vehicle n at some time instant t. Chandler et al. [19] proposed the following form of the response with a constant *driver sensitivity* (γ)

$$a_n(t+T) = \gamma[v_{n-1}(t) - v_n(t)] \tag{1.4}$$

where $v_n(t)$ and $a_n(t)$ denote the speed and acceleration respectively of a vehicle n at time t. Gazis et al. [20] proposed a more general driver sensitivity expression as follows:

$$\gamma = c\frac{[v_n(t+T)]^m}{[x_{n-1}(t) - x_n(t)]^l} \tag{1.5}$$

The response can then be modeled as in Equation 1.4 with the driver sensitivity (γ) as in Equation 1.5. Thus, the following vehicle adjusts its speed proportionally to both speed difference and the following distance. The extent to which the vehicle adjusts its speed depends on the values of c, l and m.

For the sake of completeness, we also mention about the *Cellular Automaton* models. These are microscopic models, which distinguish and trace the individual vehicles but do not describe any vehicle following behavior. The Cellular Automaton model describes, in a discrete way, the movement of vehicles from one *cell* to another[10] [21], [22]. There have been some other microscopic models proposed as well (the following papers consider some, but by no means a complete list of such models that have been published in the literature [23] - [27]).

Because of the human and other natural elements involved in the modeling of microscopic models, experimental corroboration of such models is an arduous task as there can be considerable variance in the driving behavior of individuals as well as the variance in the driving behavior of the same individual in different states of mind. Nevertheless, these models have been employed in a variety of simulation software for describing the flow of traffic and can be used to understand the dynamic interaction between the traffic management system and the drivers on the roadway network [28], [29].

b. Macroscopic traffic flow models

Macroscopic traffic flow models deal with the aggregate behavior of a collection of vehicles. In contrast to microscopic traffic flow models, they are restricted to the description of the collective vehicle dynamics in terms of the evolution of appropriate "macroscopic" variables which can express the aggregate behavior of a collection of vehicles at any location and instant of time. The macroscopic variables that are usually considered are: spatial vehicle density $\rho(x,t)$, speed $V(x,t)$ and the flow

[10] The freeway is assumed to be divided into cells. The size of the cells is chosen in such a way that a vehicle driving with a unit speed (in any chosen unit system) moves into the immediate downstream cell during one time step. The vehicles are assumed to attain only a limited number of speed values, ranging from zero to some maximum preassigned value.

$Q(x,t)$ of traffic. In the literature we can distinguish between macroscopic models which describe traffic using a continuum approach as that of a compressible fluid or a statistical approach via the kinetic theory of gases. We now briefly discuss the macroscopic models proposed in the literature based on these two approaches.

Models based on continuum approach as that of a compressible fluid

One of the most popular macroscopic traffic flow models was proposed by Lighthill and Whitham in 1955 [30]. It appears that Richards developed the same model independent of them in 1956 [31]. Their theory, called the 'hydrodynamic theory of traffic flow' underlies many of the existing macroscopic traffic flow models[11]. They employ a balance of mass equation (Equation 1.6) along with a constitutive assumption that the velocity of traffic changes instantaneously with the density[12]. It is also assumed that the relation in Equation 1.7 holds exactly.

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(x,t)}{\partial x} = 0 \tag{1.6}$$

$$Q(x,t) = \rho(x,t)V(x,t) \tag{1.7}$$

Equations 1.6 and 1.7 constitute a system of two independent equations and three unknown variables. Consequently, to get a complete description of traffic flow, a third independent model equation is required. To that end, Lighthill and Whitham used a nonlinear constitutive relationship between speed and density (or equivalently, between flow and density). Equation 1.8 shows speed as being described as a function

[11] This theory is also sometimes known as L-W-R theory of traffic flow.

[12] It is assumed that the macroscopic variables ($\rho(x,t)$, $V(x,t)$ and $Q(x,t)$) are differentiable functions of time and space.

of the traffic density[13].

$$V(x,t) = V^e(\rho(x,t)) \tag{1.8}$$

By using the above mentioned relationship between the traffic speed and density and Equation 1.7, Lighhill and Whitham obtained the partial differential equation to describe the traffic flow as shown in Equation 1.9

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial [\rho(x,t) V^e(\rho(x,t))]}{\partial x} = 0 \tag{1.9}$$

The relationship between the traffic flow and density is what is usually referred to as the Fundamental Diagram of Traffic or Fundamental Traffic Characteristic. It is important to note that since the dynamics (or inertia) of traffic is neglected, such a model is also referred to as a kinematic wave model. Subsequently, Payne [32] attempted to include the dynamics of traffic through the consideration of the dynamics of traffic speed [14].

Models based on kinetic theory of gases

The models based on kinetic theory of gases describe the dynamics of traffic in terms of speed distribution functions of vehicles at specific locations and time instants. These distributions are generally governed by the dynamics of various processes such as acceleration, interaction between vehicles and lane changing.

[13] Greenshields in 1935 [2] had suggested an empirical linear relation between speed and density as

$$V^e(\rho) = V_0(1 - \rho/\rho_{jam})$$

where ρ_{jam} is the traffic density for *bumper-to-bumper* traffic.

[14] An an overview of Payne's model and other recent developments of Payne-type models can be found in [33]. A detailed discussion on the macroscopic models based on the hydrodynamic theory can be found in a review article of Papageorgiou [10].

Newell [34] seems to be have been the first to develop a traffic flow model based on kinetic theory of gases. He treated the flow of traffic as analogous to the flow of rarified gases. However, most of the work in gas-kinetic continuum models is based on the work of Prigogine and coworkers [35] - [37].

Gas-kinetic continuum models are based on the equation for the phase-space density,

$$\tilde{\rho}(x, v, t) = \rho(x, t)\tilde{P}(v; x, t) \tag{1.10}$$

which is the product of the vehicle density $\rho(x, t)$ and the distribution $\tilde{P}(v; x, t)$ of vehicle speeds v at location x at time t. The phase-space density can be interpreted as follows: at an instant t the expected number of vehicles present at an infinitesimal region $[x, x + dx)$ driving with a speed $[v, v + dv)$ equals $\tilde{\rho}(x, v, t)dxdv$. Prigogine and coworkers assumed that changes of the phase-space density are caused by the following processes:

1. Vehicles moving with speed v which enter or exit the roadway segment $[x, x+dx)$ causes changes in the $\tilde{\rho}(x, v, t)$.

2. Vehicles not driving at their desired speeds will accelerate if possible.

3. A vehicle that *interacts* with a slower vehicle will need to reduce its speed when it cannot immediately *pass* the slower vehicle. They described this interaction by the word “collision” and assumed that this interaction was via the senses of the drivers rather than via the bumpers of the cars.

These assumptions lead to a partial differential equation for the total temporal change in the phase-space density with the crucial assumption that traffic can be treated as a continuum. Prigogine and coworkers distinguished between contributions caused by acceleration towards the desired speed and interactions between vehicles. They then suggested specific expressions for both these contributions.

There have been several modifications to the model proposed by Prigogine and coworkers. A review of models based on kinetic theory of gases can be found in [38].

Discrete Lighthill-Whitham-Richards model

These macroscopic models are governed by partial differential equations, solutions to which are sensitive to boundary conditions and can be computationally intensive, but this is not the main deficiency of these models[15]. Moreover these models are applied in practice via the application of finite difference schemes to the continuous equations described in the models. Most of these solution approaches involve numerical approximations (discretizations) in the spatial domain, temporal domain or both.

Daganzo [39] proposed the cell transmission model based on an earlier model of Newell [40] borrowing from the models of Lighthill and Whitham, and Richards to predict the traffic conditions for a stretch of a freeway by evaluating the flow at a finite number of intermediate points, selected *a priori*, including the entrances and exits.

The cell transmission model of Daganzo is a discrete flow model that uses carefully selected *cell* sizes (the freeway stretch is assumed to be divided into cells) and a piecewise linear relationship between traffic flow Q and traffic density ρ[16]. In the

[15] There are far more serious philosophical problems concerning the appropriateness of using such continuum models to describe the flow of traffic, which we discuss later in the chapter.

[16] The piecewise linear relationship assumed was:

$$Q = min[v\rho, Q_{max}, v(\rho_{jam} - \rho)], \quad \text{for } 0 \leq \rho \leq \rho_{jam} \tag{1.11}$$

where v is the free-flow speed, ρ_{jam} is the jam density, and $Q_{max} \leq \rho_{jam} v/2$ is the maximum flow possible.

model formulation, given a time step, the length of the cells are chosen such that under free-flow conditions all vehicles in a cell will flow into the immediate downstream cell. The number of vehicles flowing out of a cell is upper bounded by the space left in the immediate downstream cell.

For the purposes of this dissertation, we have briefly discussed the various traffic flow modeling approaches. More comprehensive reviews of traffic flow models can be found in [9], [41]. We will now focus mainly on macroscopic traffic flow modeling approaches. We next discuss several important issues regarding the inappropriateness of using the continuum approach to model the macroscopic flow of traffic and then we propose an alternative *non-continuum* approach to describe the macroscopic behavior of traffic.

C. Problems with using the continuum approach to model the macroscopic flow of traffic

The definition of the above mentioned “macroscopic” variables has been and continues to be a source of problems in developing macroscopic models of traffic. In all these works, the macroscopic variable “density” is defined analogous to the variable in mechanics; however, it is this definition of “density” that proves to be most problematic from the point of view of the appropriateness of the model. For the notion of density to hold for traffic and hence, for the governing partial differential equations to hold for traffic, the representative section under consideration should have a “sufficiently large” number of vehicles (just a change of few vehicles can change such a quantity significantly). Moreover the following three ratios are very important in determining whether traffic can be modeled as a continuum:

1. Following distance/Length of section

2. Length of vehicle/Length of section

3. Length of vehicle/Following Distance

Of course the third ratio above is not an independent ratio in the sense that it is a composite of the first two. The equivalent of the "length of a vehicle", "following distance" and the "length of a section" in a *continuum setting* are the "diameter of a molecule", "mean free path" and the "characteristic length of a flow domain" respectively. While the ratio (*1*), which is essentially the Knudsen number is sufficiently small for traffic (of the order of 10^{-3}), there are not sufficient number of vehicles in the section for the flow to be regarded as a continuum. Also, for the approximation of a continuum to be meaningful, the ratio (*2*) should be of the order of 10^{-8} or less as hundreds of millions of molecules might occupy the cross-section; the corresponding ratio for traffic at best can be of the order of 10^{-3}. To have a comparable number of vehicles in a section as molecules in a representative volume, the section lengths to be considered must be at least millions of miles long. It would be unreasonable to model a handful of molecules in a flow domain as a continuum and it is equally unreasonable to model a handful of cars in a "section" of a freeway as a continuum.

Another important concern with the use of the flow of traffic as a continuum is that a decrease in the speed of traffic in an upstream section also results in a decrease in the speed of traffic in the downstream section even if there is little change in the rate of vehicles entering the downstream section from the upstream section. This phenomenon is not observed in freeway traffic. *This is a fatal flaw concerning the model and this fact cannot be overemphasized.* Also, there is no transparency in incorporating the heterogeneity of traffic in existing macroscopic traffic flow models. This concern can be redressed by considering the continuum as an inhomogeneous body; but the principal objection that remains is that of treating the traffic flow as flow of a continuum.

Traffic flow models based on kinetic theory treat traffic analogous to a rarified gas. But this again associates a notion of "density". This we believe is once again inappropriate, as the kinetic theory presumes sufficient number of molecules that are constantly colliding in the flow domain of interest, many more so than the number of vehicles in a section of a highway. While one can reason away the need for collisions by replacing it by a hypothetical interaction when the vehicles get appropriately close, the main idea behind the model is ill-conceived, we do not have a sufficient number of vehicles.

Also, more importantly, the time scales associated with gases and vehicles on freeway are of distinctly different orders. The time gap, defined as the ratio of the following distance to the speed is of the order of one second in traffic. An equivalent "time gap" in gases is of the order of 10^{-10} seconds. It is the very large number of molecules at length scales of interest, and the rapidity of fluctuation of molecules, which when integrated over a characteristic time scale, allows for a meaningful definition of field variables such as the probability density; in a unidirectional movement of vehicles which constitutes the flow of traffic, the speed of vehicles as well as the number of vehicles in a representative section do not allow for such a concept to be defined in a meaningful way for a theory *a la* the kinetic theory of gases.

An additional concern with the existing approaches is that there is no transparency in modeling how the information at the vehicular level affects the macroscopic dynamics.

D. A non-continuum approach to model the macroscopic flow of traffic

As an alternative to the continuum approach, we present a discrete dynamical system approach which seems well suited to describe the dynamics of such large scale systems as the flow of traffic [8]. The main features of the proposed non-continuum approach to model the flow of traffic are:

1. Unlike most previous approaches, where first a continuum approximation of the large collection is obtained, and for practical control applications, a spatial discretization of the continuum model is sought, in the proposed approach, we directly obtain a spatially discrete, but lower dimensional model of the large collection of vehicles.

2. The proposed model based on this approach overcomes the limitation associated with existing models - namely, that the speed in the downstream section decreases in response to a slow down in the traffic in the upstream section.

3. The proposed approach integrates aggregate vehicle-following behavior into developing a macroscopic model. This is in keeping with our intuition that what we observe in traffic is a consequence of the aggregate behavior of drivers. In this respect, it sharply differs from hydrodynamic models or models based on kinetic theory of gases where one resorts to analogies with a fluid. While there are many types of fluids [42], existing methodologies do not provide a rationale as to why an analogy is drawn to a particular type of fluid (e.g., Navier-Stokes fluid).

4. The proposed methodology also allows for refining the vehicle following part of the model to account for heterogeneity in vehicle following. In this respect,

it overcomes the limitation of opacity in incorporating such effects in existing models.

5. The evolution equation for the speed of traffic is described through the vehicle following behavior, which changes with the regime of traffic. For example, in the un-congested regime, there is little change in speed in response to the following distance, while it is not so in the congested regime of traffic. While the definitions of regimes themselves are empirical, the models proposed here are hybrid in nature reflecting the different phases of congestion in traffic. Correspondingly, the model of the flow of traffic may be considered as a model of a hybrid system.

E. Dissertation contributions

This dissertation is an attempt to advance the state-of-the-art in macroscopic traffic flow modeling. The main contribution of the dissertation is the modification and corroboration of the non-continuum traffic model with traffic data collected from the sensors deployed on the freeways. Specifically,

- A comprehensive methodology for traffic state estimation and prediction for real-time ITS applications is developed.
  - Aggregate vehicle following behavior is identified based on traffic data collected from the sensors deployed on the freeways.
  - The performance of the developed methodology is rigorously evaluated by validating it against the real traffic data.
- A methodology for predicting trip travel times is developed.

F. Organization of this dissertation

This thesis is organized as follows. In Chapter II, we introduce the non-continuum traffic flow model. Chapter III consists a discussion of the available traffic data from US 183 freeway in Austin, TX. We then present the methodology developed to corroborate the non-continuum traffic flow model in Chapter IV. In Chapter V we present the developed methodology for short term traffic state forecasts and prediction of trip travel time. We conclude the thesis with possible directions for future research in Chapter VI.

CHAPTER II

A NON-CONTINUUM TRAFFIC FLOW MODEL

A. Introduction

The traffic flow model must have two components - one that reflects the balance of vehicles in a section and the other which reflects how vehicles react to the following distance in a section. The second component concerns aggregate vehicle following behavior [8]. In addition to these components, a traffic flow model should have the following desired attributes:

1. it should be based on sound physical principles,

2. it should be expressible in terms of physical, and if possible, measurable variables. If this is not possible, then such variables should be quantifiably inferrable,

3. it should *preferably* be governed by an ordinary differential equation of a low order for traffic flow control and estimation purposes, and[1]

4. it should be easy to take care of the translation from *Lagrangian* (vehicle following) description of the traffic flow to *Eulerian* (traffic dynamics at a fixed point on the freeway) description.

The translation from a Lagrangian to an Eulerian description is very important since traffic operations such as ramp metering are performed at fixed points on the freeway, while vehicles occupy different points on the freeway at different times. In this chapter, we present a simple and a spatially discrete model for the flow of traffic. This is done

[1] This is because the systems described by ordinary differential equations are easier to deal with considering the state-of-the-art in control and estimation.

through the introduction of the concept of a limit of a collection of dynamical systems. We model the traffic as a countably infinite collection of *homogenous*, interconnected dynamical systems. To this end, each vehicle in the traffic is considered as a dynamical system which has its speed and following distance as state variables describing its dynamics. The traffic can then be thought as a collection of these vehicles interacting on the freeway. The interaction between the vehicles is defined through a *vehicle following* behavior.

The central hypothesis of a non-continuum approach to traffic modeling is the existence of a few "representative" vehicles on every section of a freeway. A "representative" vehicle (equivalently a dynamical system) can be thought of as a limit of a collection of dynamical systems (vehicles) of finite state space dimension. This is analogous to a limit of a sequence. The choice of a limit of a collection as an aggregate is motivated by the need to translate from a Lagrangian description to an Eulerian description of traffic flow. The vehicle following behavior of the representative vehicles reflects the aggregate vehicle following behavior of traffic. Together with a balance of the number of vehicles in any section of the freeway, the limit of a collection of dynamical systems, which is the governing equation for the evolution of the speed of traffic, describes the dynamics of the flow of traffic at a point on a section of a freeway.

Before we discuss the model further, we must define a section. Consider a freeway equipped with detector stations. We consider the detector stations indexed in an increasing manner from the downstream end of the highway to the upstream end of the highway. For pragmatic reasons, we consider a section to be the stretch of highway between two consecutive odd numbered detector stations. If the number of detector stations is even, we consider that a fictitious detector station exists at the upstream detector station so that the total number of detector stations is odd.

We index the sections in an increasing order as we traverse the freeway from the downstream end towards the upstream end. Thus if any section is indexed as i, the immediate upstream section is indexed as $i+1$. Fig. 2 shows a schematic for the division of freeway into sections. We next present the issues and assumptions in treating traffic as a large collection of dynamical systems and precisely define the limit of a collection of dynamical systems.

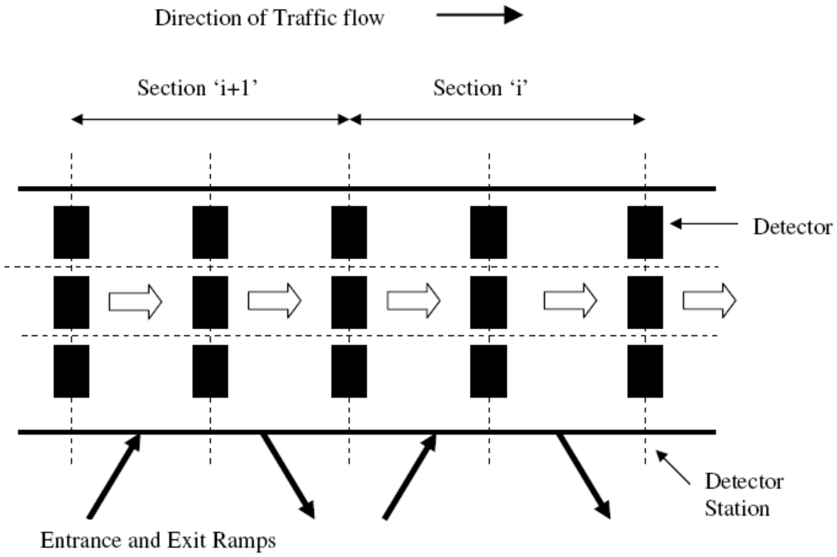

Fig. 2. Division of freeway into sections

B. Limit of a collection of dynamical systems

Traffic is considered as a large collection of dynamical systems with each vehicle in the traffic treated as a dynamical system. We are interested in a systematic methodology for obtaining a macroscopic description of this large scale system from an awareness of its microscopic description (vehicle following behavior). Thus we are interested in formulating the problem to model the movement of traffic that can be verified with the measurements that can be obtained from the roadside sensors. Since the sensors are placed at discrete points on the freeway, the focus is to model the traffic movement where the sensors are placed. We assume that the following three variables are either measurable or inferrable at the location where sensors are placed: speed and following distance of vehicles as they cross the point where the sensors are placed, and the rates at which vehicles enter or leave a section of a freeway. In this section, we use the assumptions and definitions of a limit of a collection of dynamical systems from [8]. They have been provided here for completeness sake.

1. Problem formulation

We can define a general problem in modeling the dynamics of traffic as follows: Suppose that we are given a fixed location on the freeway and we ask the question whether it is possible to approximate the evolution of measured/inferrable variables (example: speed and following distance) by constructing a dynamical model. To put the same question mathematically, we are given a strictly monotonically increasing and unbounded sequence of time instances, $\{t_i\}$ (instances of crossing a fixed point on the freeway) and the outputs of different vehicles (discrete dynamical systems) at those time instances, $\{y_i(t_i)\}$. The question is whether there exists a model of the form:

$$\dot{\omega} = g(\omega, u), \tag{2.1}$$

such that for some smooth h, we have:

$$h(\omega(t_i)) \rightarrow y_i(t_i) \quad as \quad i \rightarrow \infty \tag{2.2}$$

The specific requirement is that the above system be finite dimensional. The answer depends on the initial conditions for the vehicles in the traffic and whether disturbances in traffic get attenuated i.e., whether the string of vehicles in the traffic is stable. We assume the following before going on to define the limit of a collection of dynamical systems.

- Collection of vehicles is homogenous, i.e., all vehicles in the traffic have identical governing equations. This assumption can be easily relaxed to have a finite number of representative vehicles. For example, if there are significant percentage of trucks in a lane on a freeway, one can have two representative vehicles - a car and a truck.
- Currently we assume that no vehicles enter or leave the collection and that vehicles always maintain the same ordering throughout.
- Actions of a vehicle are only dependent on the state of the vehicle immediately preceding it and that the state space dimension of all the vehicles are identical.

We model the traffic movement via an infinite system (string) of coupled ordinary differential equations of the form:

$$\begin{aligned}\dot{x}_i =& f(x_i, y_{i-1}, u), \quad i = 1, 2, ..., \\ y_i =& h(x_i) \\ x_j =& x_1, \quad \forall j \leq 0;\end{aligned}$$

Here, u is the reference information, if any, available to each vehicle and $x_k(t) \in \Re^n$ and $y_k(t) \in \Re^m$ for all k. The output of a system (vehicle), y_k may be a subset of the states of the system. The *interactions* in this model are lower-triangular or “look ahead”. This description of traffic via the use of an infinite system of ordinary differential equations is reasonable because the response of the drivers is based on the state of the vehicle in front of him/her. As a result, the introduction of vehicles at the tail of a finite vehicle string does not alter its behavior.

Now suppose at an instant of time we consider the sequence of state of systems (vehicles) in the collection. The following questions are of importance:

1. Does this sequence have a limit?

2. If it does has a limit, then is the convergence to the limit uniform in time?

In physical terms, the question can be thought of as whether disturbances originating from the lead vehicle attenuate spatially. If so, vehicles at the tail of the string do not get affected at all and behave as though they are almost identical and hence spatially shift invariant. In fact, a similar, simplified car-following model was recently suggested by Newell [43]. In such a scenario, what the sensors detect asymptotically in time is the limit of the string of vehicles, when disturbances attenuate. In this sense, a limit of a collection takes care of the conversion from Lagrangian description (vehicle following) to the Eulerian description (dynamics at a fixed point). It is specifically for this purpose that we are interested in using the limit as an aggregate or a macroscopic quantity to describe the traffic dynamics.

2. Definition of a limit of a collection of dynamical systems

The class of large scale systems considered are described by an infinite system of differential equations of the form:

$$\dot{x}_k = f(x_k, x_{k-1}, ..., x_{k-r+1}, u), \quad x_k(t) \in \Re^n \tag{2.3}$$

$$x_j(t) \equiv x_1, \quad \forall j \leq 1 \tag{2.4}$$

Definition 1. A system of the form:

$$\dot{\omega} = g(\omega, u), \quad \omega(t) \in \Re^n, \tag{2.5}$$

is a limit of the collection if

1. for every $\omega(0)$ and every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_j \parallel x_j(0) - \omega(0) \parallel < \delta \Rightarrow \sup_j \sup_t \parallel x_j(t) - \omega(t) \parallel < \epsilon \tag{2.6}$$

   and

2.

$$\lim_{j\to\infty} x_j(0) = \omega(0) \Rightarrow \lim_{j\to\infty} \sup_t \parallel x_j(t) - \omega(t) \parallel \to 0 \tag{2.7}$$

We will refer to $\{x_k(t), k \geq 1\}$ as the solution of the string corresponding to initial conditions, $\{x_{k0}\}$ if

$$\dot{x}_k(t) = f(x_k, x_{k-1}, ..., x_{k-r+1}, u), \quad k = 1, 2, ..., \tag{2.8}$$

$$with\ \ x_j(t) := x_1(t), \quad \forall j \leq 1, \ \ and \tag{2.9}$$

$$x_k(0) = x_{k0}, \quad k = 1, 2, ... \tag{2.10}$$

An important ingredient for the determination of a limit of a collection of systems is that of stability of the solutions. We use the following notion of stability of a solution, also referred to as string stability [44].

L_∞ stability: Let $\{x_i(0)\}$ and $\{\hat{x}_i(0)\}$ be two sets of initial conditions and let $\{x_i(t)\}$ and $\{\hat{x}_i(t)\}$ be the corresponding solutions. The solution $\{x_i(t)\}$ is Ł$_\infty$ stable if there exists $\epsilon > 0$ such that

$$\sup_i \parallel x_i(0) - \hat{x}_i(0) \parallel < \delta \Rightarrow \sup_i \sup_{t\geq 0} \parallel x_i(t) - \hat{x}_i(t) \parallel < \epsilon.$$

Spatial Asymptotic Ⱡ$_\infty$ stability: Let $\{x_i(0)\}$ and $\{\hat{x}_i(0)\}$ be two sets of initial conditions and let $\{x_i(t)\}$ and $\{\hat{x}_i(t)\}$ be the corresponding solutions. The solution $\{x_i(t)\}$ is spatially asymptotically Ⱡ$_\infty$ stable if

$$\lim_{k\to\infty} \parallel x_k(0) - \hat{x}_k(0) \parallel = 0 \Rightarrow \lim_{k\to\infty} \sup_{t\geq 0} \parallel x_k(t) - \hat{x}_k(t) \parallel = 0.$$

We now state the main result given in [8].

Proposition 1. *For the collection of dynamical systems considered above, suppose that every solution $\{x_k(t)\}$ is spatially asymptotically stable; then the limit of the collection is given by:*

$$\dot{\omega} = f(\omega, \omega, ..., \omega, u).$$

Proof:

It is easy to verify that when the initial condition of all the systems is ω_0, $x_k(t) \equiv \omega(t)$, where[2]

$$\dot{\omega} = f(\omega, \omega, ..., \omega, u), \quad \omega(0) = \omega_0$$

Since the solution $\omega(t)$ is spatially asymptotically stable, it implies that, from the definition of stability and that of the limit of the collection of dynamical systems, the system described by

$$\dot{\omega} = f(\omega, \omega, ..., \omega, u),$$

is the limit. This limit of the collection of dynamical systems is unique.

[2] The underlying assumption is that a solution to this differential exists and is unique

C. Variables that are used to describe the flow of traffic

We use the number of vehicles, N, in the section at any given time, aggregate following distance $\bar{\Delta}$, and the aggregated speed of traffic, $\bar{v}$ as the variables that can describe the traffic dynamics in a section of a freeway. The aggregate following distance and number of vehicles are considered as different state variables for the following reasons:

1. We believe that the psychology/inconsistency of drivers can render $\bar{\Delta}$ and N as independent variables while satisfying an inequality constraint that their product be at most the length of the section. However, we relax such a constraint in this dissertation.

2. Should a traffic model of higher fidelity, involving distinct classes of vehicles be developed, there will be two variables representing the aggregated following distance for each class of vehicles. Hence, to be consistent with possible future refinements, it makes sense to consider the two as independent variables.

3. Data concerning the flow of traffic suggests the existence of different traffic regimes (for example congested and un-congested) which probably can be explained due to a switch in the driving behavior of drivers. This switch can be potentially made on the basis of the increase (or decrease) in the number of vehicles in a section (which leads to increased or decreased levels of occupancy). We have found that a switch of the regime based on the number of vehicles in traffic to be problematic and we expect that it will continue to be problematic with different classes of vehicles and driving behavior. In fact, it may be simpler to base it on occupancy as occupancy can be measured while the average following distance must be inferred from traffic data. Switching of a traffic regime can also potentially explain the hysteretic (here by hysteresis

we mean the lagging of the effect behind that of the cause and not to its use in mechanics to signify the conversion of working into thermal energy) behavior observed in traffic (eg. [7], [45], [46]).

It is reasonable to assume that vehicles on a freeway react to changes in the following distance and the relative velocity to maintain a safe distance from their preceding vehicles on a freeway. One may model the vehicle following behavior of an automated as well as a non-automated vehicle on the freeway as (for any function $x(k)$, we use $\dot{x}(k)$ in this dissertation to mean $\frac{x(k+h)-x(k)}{h}$ where h is the time step):

$$\dot{v} = f(v, \Delta, \dot{\Delta})$$

where, Δ and $\dot{\Delta}$ represent respectively the following distance and the rate of change with respect to time of the following distance of a vehicle (hereafter referred to as the *relative speed* of the vehicle) travelling on a freeway. This model of vehicle following neglects variations of the driving behavior of drivers; nevertheless this is a reasonable model for the following reasons:

1. We are interested in the aggregate behavior of vehicles on the freeways. Also, the observation of stable throughput on a number of freeway sections suggests that the aggregate behavior of vehicles is well defined, although the behavior of individual vehicles may not be.

2. Such a vehicle following behavior is reasonable for automatically controlled vehicles on automated freeways.

3. The granularity of the model required dictates the heterogeneity of vehicle following behavior that should be considered at the microscopic level.

We also make an important assumption that the vehicle following behavior is either string stable (in the case of non-automated traffic) or can be engineered to be string stable (in the case of automated traffic). By string stability, we mean that the following distance of any vehicle in the section remains close to its desired following distance at any time [44], [47]. This assumption enables us to approximate the vehicle following dynamics of each of the vehicles in a section of a freeway with that of a "representative" vehicle. In physical terms, it enables one to approximate the evolution of traffic speed from the dynamics of representative vehicles in the traffic. In other words, if there are l consecutive sections on a freeway, and if $\bar{\Delta}_i$, $\dot{\bar{\Delta}}_i$ and $\bar{v}_i$ represent the following distance, relative speed and speed of vehicles as they cross a generic location A_i in the i^{th} section of a freeway at any instant of time, the aggregated speed dynamics, as seen by an observer at the location A_i, may be approximated by:

$$\dot{\bar{v}}_i = f(\bar{v}_i, \bar{\Delta}_i, \dot{\bar{\Delta}}_i)$$

The exact structure of the function, f, requires further discussion and can be found later in Chapter III and Chapter IV.

D. The model

Let $\dot{N}_i$ denote the rate of change of the number of vehicles with respect to time of the number of vehicles in ith section. Then $\dot{N}_i$ is computed using the balance of vehicles on the freeway as follows:

$$\dot{N}_i = \dot{N}_i^{en} - \dot{N}_i^{ex} + \dot{\tilde{n}}_i.$$

In the above equation, $\dot{N}_i^{en}$ is the rate of vehicles entering the section from its upstream section, $\dot{N}_i^{ex}$ is the rate of vehicles exiting the given section into a downstream section (if there is one), and $\dot{\tilde{n}}_i$ is the net inflow into the section from the ramps.

If there are l sections under consideration and are indexed in an increasing order from the downstream end to the upstream end, the following must be true to ensure compatibility:

$$\dot{N}_{i+1}^{ex} = \dot{N}_i^{en}, \qquad\qquad i = 1, 2, ..., l-1.$$

The constitutive equation for $\dot{N}_i^{ex}$ is

$$\dot{N}_i^{ex} = \frac{\bar{v}_i N_i}{L_{s,i}},$$

where $L_{s,i}$ is the length of the i^{th} section.

To complete the model for the flow of traffic, we must provide an equation for the evolution for the aggregated following distance. We hypothesize that the evolution equation for the aggregate following distance in a section consists of two components: The first component is due to the *net* influx of vehicles from ramps and the mainline and the second component is due to the speed differential between the vehicles in the section and those in the section immediately downstream. The addition of the second component is critical to obtain a correct directional propagation of disturbances in traffic. Since the first section does not have any sections downstream from it, the evolution equation for aggregated following distance contains only the first component. One may hypothesize the first component for the i^{th} section as:

$$-\frac{(\bar{L}_{car} + \bar{\Delta}_i)^2}{L_{s,i}}\dot{N}_i.$$

The second component is

$$\beta_{i,i-1}\bar{v}_{i-1} - \bar{v}_i.$$

where, $\beta_{i,i-1}$ is called the *speed-correction* factor. The subscripts $(i, i-1)$ follow the notion of section indexing as defined earlier. In a sense, it reflects the speed of a representative vehicle from the downstream section as seen by the representative vehicle upstream. It is intuitive to understand that $\beta_{i,i-1}$ will be equal to one if the aggregate traffic speeds in the downstream and the upstream section are equal. This is basically due to our assumption that all the vehicles in a section have the same speed as that of the representative vehicle.

Putting everything together, the model for the flow of traffic in the i^{th} section of a freeway is:

$$\begin{aligned}
\dot{N}_i =& \dot{N}_i^{en} - \dot{N}_i^{ex} + \dot{\tilde{n}}_i \\
\dot{\bar{\Delta}}_1 =& -\frac{(\bar{L}_{car} + \bar{\Delta}_1)^2}{L_{s,1}}\dot{N}_1, \\
\dot{\bar{\Delta}}_i =& -\frac{(\bar{L}_{car} + \bar{\Delta}_i)^2}{L_{s,i}}\dot{N}_i + \beta_{i,i-1}\bar{v}_{i-1} - \bar{v}_i, \quad i > 1 \\
\dot{\bar{v}}_i =& f(\bar{v}_i, \bar{\Delta}_i, \dot{\bar{\Delta}}_i) \\
\dot{N}_i^{ex} =& \frac{\bar{v}_i N_i}{L_{s,i}} \\
\dot{N}_i^{en} =& \dot{N}_{i+1}^{ex}
\end{aligned} \tag{2.11}$$

E. Summary

In this chapter we have presented the non-continuum traffic flow model. We still need to discuss the mechanism for traffic congestion as depicted by this model. We discuss that later in Chapter IV, once the traffic regimes and their respective vehicle following structures have been described and identified.

CHAPTER III

FREEWAY TRAFFIC DATA AND ITS ANALYSIS

A. Introduction

We can classify the traffic information into three categories: historical, current or real-time and predictive. To these three categories of traffic information we can associate the respective traffic data. Historical traffic data describes the traffic conditions/states at some previous time periods. Historical traffic data can be used to observe aggregate traffic behavior which can be useful in developing heuristics for traffic forecasting. Interesting traffic flow patterns can be observed which are useful in classifying the traffic flow into different categories, example working day or a holiday. Real-time traffic data describes the current state of traffic. It is dynamic in nature and it can be used with historical data to perform on line analysis for various transportation management applications. Predictive traffic information associates with itself forecasted traffic state data. Predictive traffic information is anticipative in nature and involves the process of estimating the anticipated traffic conditions at some future point in time.

Freeway traffic data is gathered by the use of traffic surveillance systems. In general, there are two types of traffic surveillance systems: road-based and vehicle-based.

1. *Road-based*: Road-based detection systems are local in nature, as they are installed at specific locations on a freeway. Typically they can be classified as either “in-road” or “road side”. Most commonly used “in-road” detection systems are the inductive loop detectors and they have been a principal element of freeway surveillance and incident detection for many years. Other in-road detec-

tion technologies include the use of magnetometers and piezoelectric detectors. All of these technologies can be used to measure vehicle passage, vehicle count, and occupancy [1]. "Road side" detectors are generally mounted on overhead structures or to the side of the pavement. Typically they consist of video image detection systems (example, a closed circuit TV camera) or other technologies based on such as infrared, microwave, radar and ultrasonic.

2. *Vehicle-based*: Vehicle-based traffic surveillance systems involve probe vehicles equipped with tracking devices, such as transponders (electronic tags) and GPS, that allow the vehicles to be tracked by a central computer facility. There have been efforts to collect traffic data using aerial surveillance as well [48]. They try and capture photographs which indicate the positions of various vehicles on the freeway. Successive snapshots are taken at relatively short intervals of time to establish the trajectories of vehicles by processing consecutive picture frames. Vehicle-based surveillance systems are not yet widely used. But they show great promise in estimating travel times and identifying origin-destination patterns.

For the purposes of this study and corroboration of the non-continuum traffic flow model, we have used the data collected by an inductive loop detection system. In the next section we briefly describe the working principle of the inductive loop detector and the traffic data collected by the same.

[1] Occupancy of a point, to be more precise, an "appropriately small" neighborhood of the point on the freeway at a given time, t, is the percentage of time it is occupied by a vehicle in $[t - T, t]$, where $T > 0$ is sufficiently large. One can define the occupancy as the probability of finding a vehicle at the point over a suitably normed unit observational time scale of measurement.

B. US 183 dual inductive loop detector data

For this study, we use the dual inductive loop detector data collected by Texas Department of Transportation (Tx DOT) on US 183 freeway in Austin, Texas, USA. The data was made available by the Translink Research Laboratory[2] at the Texas Transportation Institute (TTI).

1. Inductive loop detector

Inductive loop detectors were introduced in the early 1960s. Since then, they have become the most common form of traffic detection system. The main or principal component of an inductive loop detector system is one or more turns of insulated loop wire wound in a shallow slot sawed in the pavement. This wire is connected to the detector electronics and controller unit. The detector electronics unit drives energy through the loop system. The loop system forms a tuned electrical circuit of which the loop wire is the inductive element. When a vehicle passes over the loop or is stopped within the loop, it decreases the inductance of the loop. This decrease in inductance then actuates the detector electronics output relay or solid state circuit which, in turn, sends an impulse to the controller unit signifying that it has detected the passage or presence of a vehicle. The most common loops used are square loops of edge equal to six feet. The traffic data measured by these single loop detectors are the vehicle count and occupancy [49].

A dual-loop detector is formed by two consecutive single-loop detectors few meters apart. With two loops, one can record the time taken by a vehicle to traverse from the first loop to the second loop. Since, the distance between the two loops is predetermined, a dual-loop detector can calculate traffic speed fairly accurately based

[2] http://translink.tamu.edu

on such information. By applying the calculated speed from the dual-loops and the single-loop measured lane occupancies, the length of a vehicle can be estimated and the vehicle can be assigned to a certain class based on its length. Thus, the dual-loop detectors distinguish themselves from single-loop detectors by giving speed and vehicle-classification data. Fig. 3 show a typical dual loop detector setup.

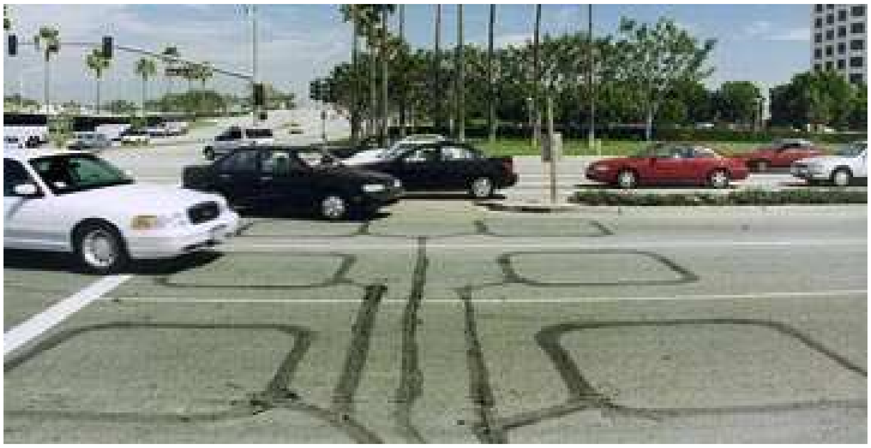

Fig. 3. Dual loop detector setup

2. US 183 data set

The provided data set by Texas Transportation Institute contains archived traffic data that were collected during the years 2003 and 2004 on select Austin freeways by Texas Department of Transportation [50], [51]. The provided data is preprocessed and aggregated into one minute intervals. Data is provided in the form of comma-separated value (csv) files. Each data file contains one hour data for all detectors (example, on select locations on US 183 in Austin) for which data are being collected. Files are named following a specific naming convention that contains the freeway name, year, month, date, and time at which the data was collected. For example, the following file contains detector data for US 183 for the "0900" hour (09 : 00 AM - 09 : 59 AM) on January 2, 2004:

$$\underbrace{\text{US 0183}}_{\text{Freeway}}\ \text{SCU\_}\underbrace{\text{0900}}_{\text{Time}}\text{\_}\underbrace{\text{20040102}}_{\text{Date}}.\ \underbrace{\text{DET}}_{\text{Extension}}$$

In each file, the first line contains data that defines the number of detectors for which data is being collected, detector identification tag[3], and the station name[4]. It has the following format:

nnn,xxxxxxx,yyyyyyyyyyyyy, xxxxxxx,yyyyyyyyyyyyy, · · ·

where,

nnn = number of detectors for which data is being collected,

xxxxxxx = detector ID, and

yyyyyyyyyyyyy =station name string.

The next sixty lines in the file provide minute by minute-by-minute traffic data for all the detectors in the following format:

hhmmss,xxxxxxx,vvv,ooo,sss,ttt,xxxxxxx,vvv,ooo,sss,ttt, · · ·

[3] Detector ID is a unique seven digit identifier assigned to each detector by Tx DOT

[4] Station name is a unique identifier used for grouping together all travel lanes in a given direction for a given roadbed. Ramps are grouped separately from the adjacent mainlines. Station names have been assigned by TTI

where,

hhmmss = time stamp (hh = hour, mm = minute, ss = second) of the data recorded,

xxxxxxx = detector ID,

vvv = number of vehicles passing through the detector location,

ooo = occupancy (0-100%),

sss = average speed of vehicles that passing through the detector location, and

ttt = average percentage of trucks (0-100%).

C. Historical traffic data and observed traffic flow patterns

With two years of dual loop detector data available, we treat 2003 and first half of 2004 data as historical traffic data. Because the traffic data is available for each individual lane in the main-line (there are either three or four lanes on US 183 in Austin), we aggregated this data to obtain traffic data for each station. We obtained aggregated vehicle counts and occupancies (in percentage) by summation of the vehicle counts and occupancies across all the lanes. We computed aggregate traffic speed by averaging the speeds across the individual lanes.

Upon careful observation of the data for many days at multiple locations on US 183 in Austin, we make a preliminary observation that the aggregate behavior of traffic is different for working days and non-working days. By non-working days we imply weekends (Saturdays and Sundays) and other special holidays (example Christmas day). Fig. 4 and Fig. 5 show typical aggregate speed, number of vehicles passing per minute, and the occupancy (in percentage) at a location in Austin (at Metric Blvd. on US 183 North Bound) across three lanes for a working and a non-working (Sunday in this case) day respectively. The difference between the traffic behavior on a working and a non-working day can be observed in terms of the number of vehicles passing per minute and the occupancy levels. Also, there is no drop in the aggregate traffic speed on the non-working days. This observation is similar to the observations of Zackor and of Chrobok et. al. on German freeways [52], [53].

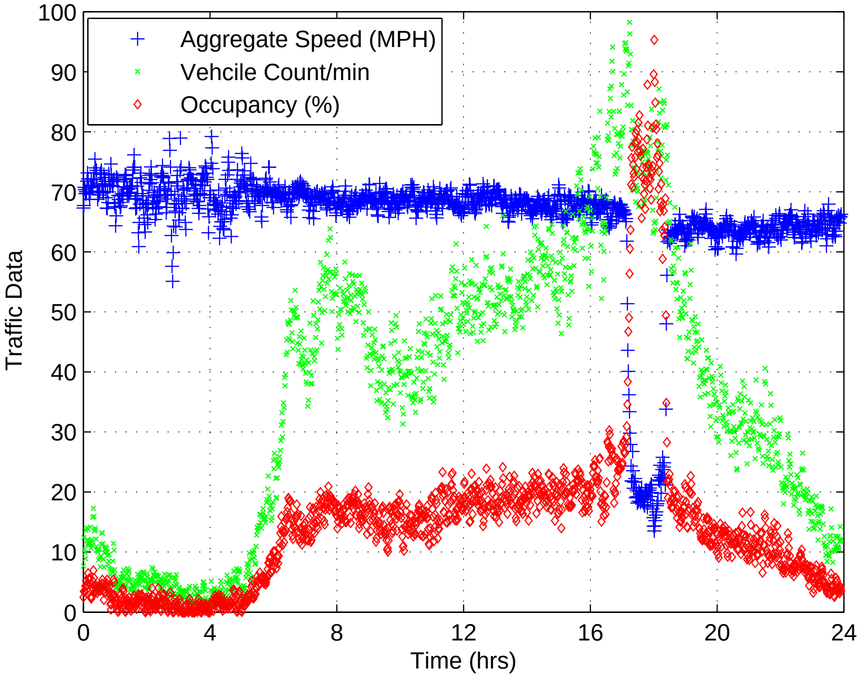

Fig. 4. Traffic data for May 17, 2004 (Monday) at a location on US 183 NB (Metric Blvd.)

For working days, the aggregate behavior of traffic does show a stable throughput. By stable throughput, we mean that consistent values of traffic flow are observed on a number of days and in a repeatable manner. Also the traffic behavior on all working days except the Fridays' exhibit sharp decrease in aggregate traffic speed during the evening hours. Fig. 6 shows the aggregate traffic speeds observed on typical working days.

Although there is no decrease in the aggregate traffic speed on a Friday, as shown

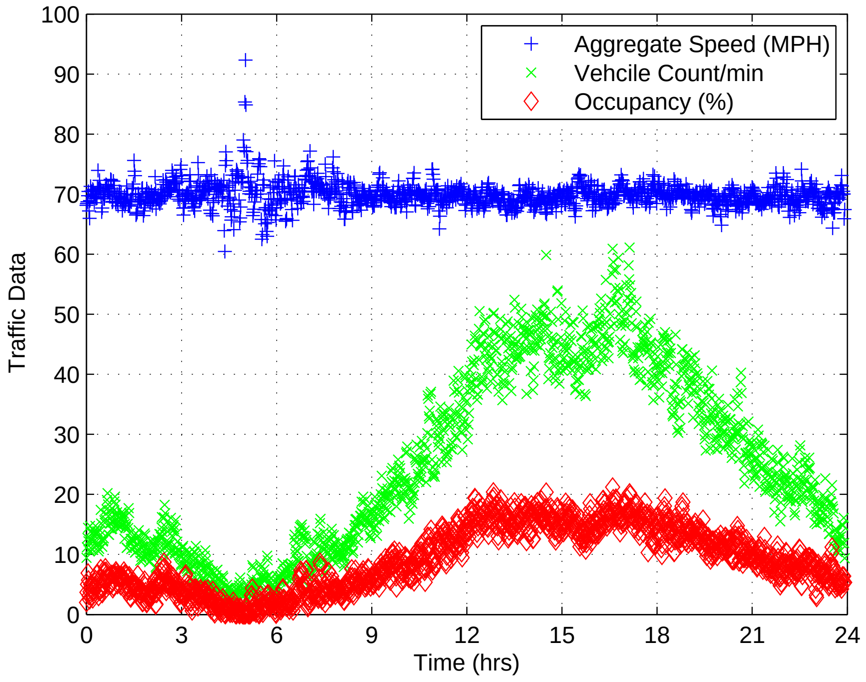

Fig. 5. Traffic data for May 16, 2004 (Sunday) at a location on US 183 NB (Metric Blvd.)

in Fig. 6, it is important to note that the traffic behavior on a Friday is different from that of a non-working day. This can be seen in terms of the quite different level of vehicle counts observed on Fridays' and non-working days as shown in Fig. 7. A possible explanation for this different traffic behavior on Fridays' is provided later in the dissertation.

For the working days (except Fridays'), the traffic data suggests the existence of different traffic flow regimes, viz:

1. A sharp decrease in the aggregate speed of traffic between 4:30 PM to 6:30 PM. This is accompanied by an increase in the number of vehicles passing per minute and the occupancy levels.

2. An almost constant aggregate traffic speed at all other times of the day.

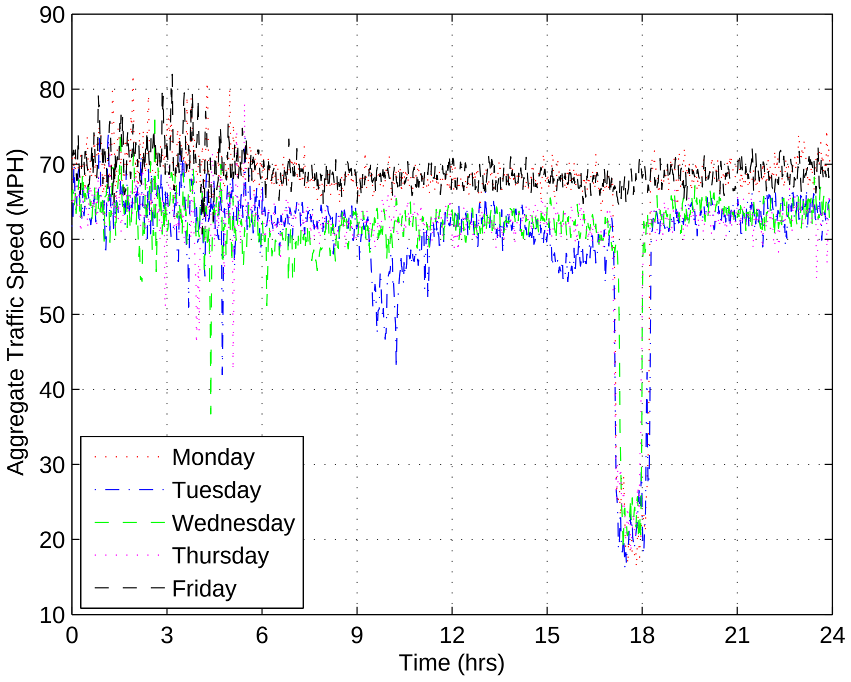

Fig. 6. Aggregate traffic speed for all five working days

There seems to be a threshold occupancy level after which sharp decrease in the aggregate traffic speed is observed. This threshold occupancy level is location specific and is observed to vary between 30 - 50 %. This suggests that period of low aggregate traffic speed begins when traffic switches from a high-speed, high vehicle count (free flow) state to a low-speed, low vehicle count glut of vehicles. This transition seems to occur when the number of vehicles in the section (or alternatively the occupancy) exceeds a critical level. Once the traffic enters a congested state, it takes a long time to return to a non congested state. Fig. 8 illustrates this congestion phenomenon. It plots the aggregate traffic speed against the five minute aggregated vehicle counts at Metric Blvd. on US 183 in Austin from 6 AM to 9 PM. From early in the morning till about some time after 5 PM in the evening, the traffic aggregate speed remains high (between 50 and 70 miles per hour) while the vehicle count consistently increases.

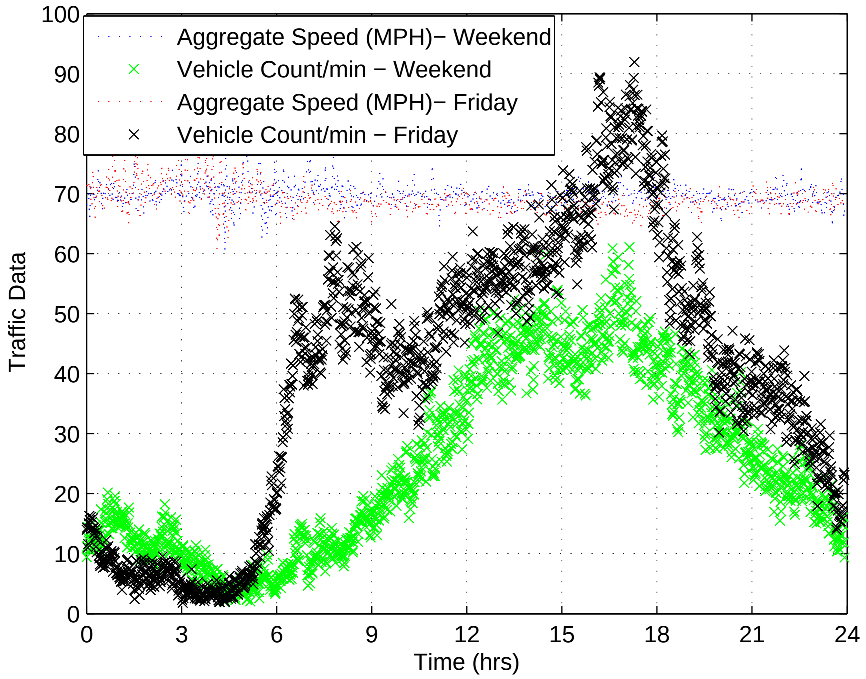

Fig. 7. Traffic behavior on a Friday and a non-working day

An influx of vehicles in the evening rush hour pushes the number of vehicles trying to utilize the freeway above some critical level (or equivalently the occupancy levels reaching some critical threshold level), forcing the traffic into a congested state as illustrated in Fig. 8. For example, at 5:45 PM the traffic has slowed down considerably (to about 20 miles per hour) and the vehicle count in a five minute segment has also decreased to about 369. Traffic does not return to a relatively un-congested state again till about 6:24 PM. Fig. 9 illustrates the above explanation with respect to percentage occupancy. It can be seen from the figure that there exists some threshold occupancy level, above which the aggregate traffic speed begins to decrease and the traffic goes into a congested state. These sharp drops in the speed of the vehicles and hysteretic behavior have also been reported by other investigators [7], [45], [46], [54].

Periods of almost constant aggregate speed and of sharp speed drops suggests

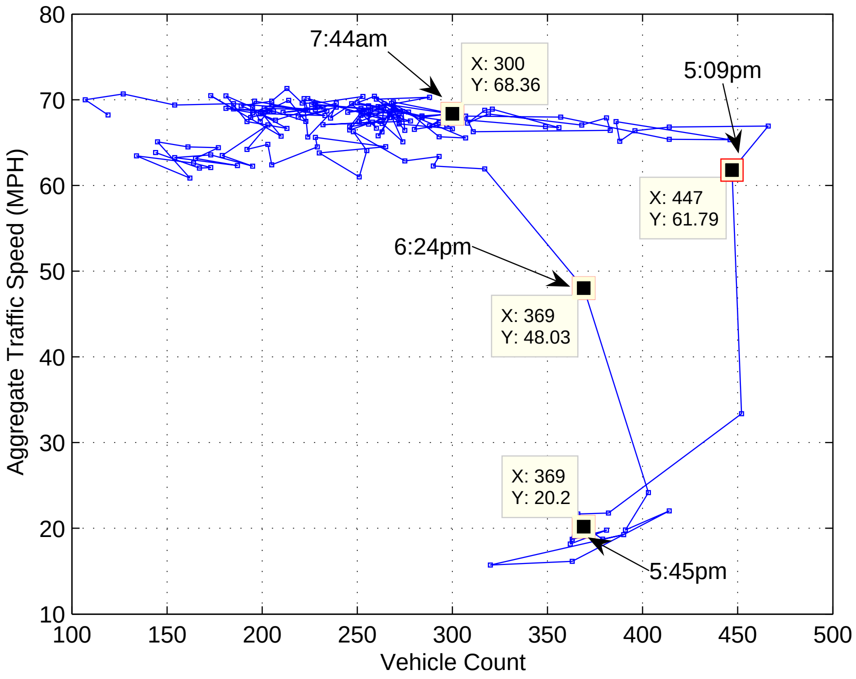

Fig. 8. Aggregate traffic speed versus vehicle count in 5 minute intervals at Metric Blvd. in Austin

different driving patterns as a reaction to the number of vehicles that are trying to use the same section of the freeway at the same time. Various other researchers have also reported the existence of different traffic flow regimes, which have been related to different behavioral characteristics of the drivers. Some vehicle following (microscopic traffic flow) models which subscribe to this line of thought can be found in [17], [55], [56], and more recently by Kerner [26]. The work of Kerner is a microscopic model in which the author hypothesizes the existence of three phases of driving. The distinction here is that we are making an assumption on the aggregate vehicle following behavior in a section and our vehicle following component is macroscopic in its context.

Typically for working days (except Fridays'), we can classify the flow of traffic

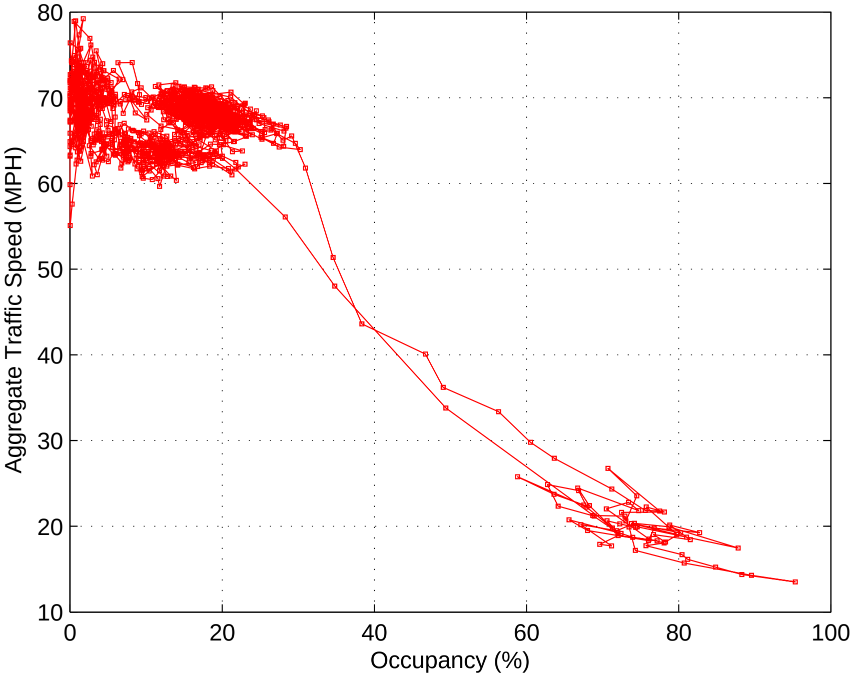

Fig. 9. Aggregate traffic speed versus percentage occupancy

into the following regimes:

1. **Free Regime:** Observed during early morning hours and late night hours. Typically from 10 PM to 7 AM.

2. **Regime 1:** Observed from 7 AM to 10 PM except the time duration when there is a sharp decrease in traffic aggregate speed.

3. **Regime 2:** Observed sometime between 4:30 PM to 6:30 PM. The duration of this regime is determined by observing the occupancy levels. The exact time for the onset of this regime is not known. Even on similar working days (example Mondays'), the time periods during which the traffic remains in *Regime 2* are quite different. Fig. 10 show this for four different Mondays' in 2004. It can be seen, that on all four days the onset and the time duration during which the

traffic remains in *Regime 2* are different.

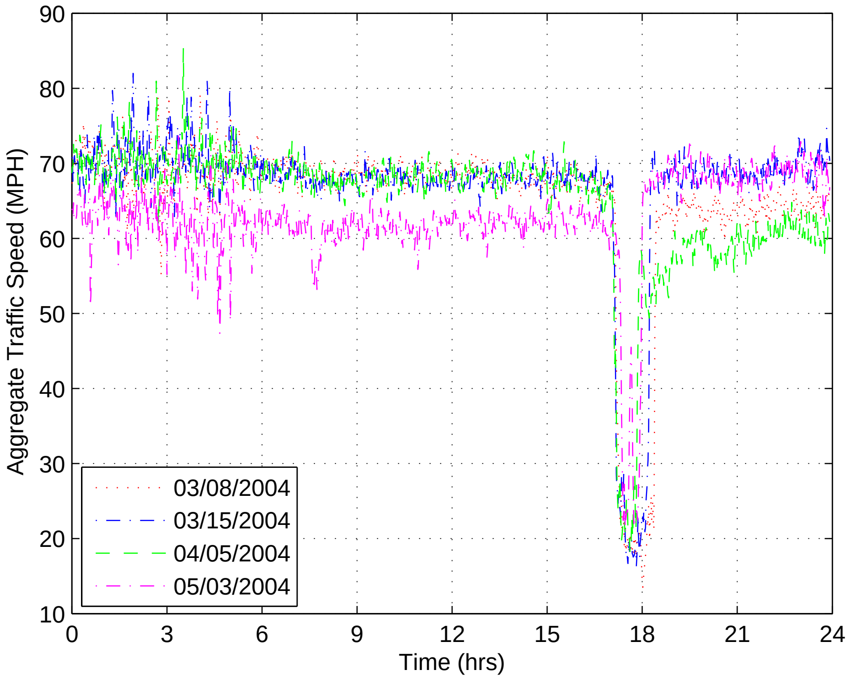

Fig. 10. Aggregate traffic speed illustration for four Mondays'

By different traffic regimes, we mean that different vehicle following behaviors can be hypothesized. During the *Free Regime*, high aggregate speeds and very few number of vehicles per minute passing thorough a location are measured on the freeway. This can be deduced both from the traffic data consisting of occupancy and number of vehicles per minute through a location. The vehicles which belong to this regime can be thought of as being driven with their desired speeds without any interaction with other vehicles. When considering vehicles in *Regime 1* and *Regime 2*, we need to take cognizance of the fact that there is an interaction between the vehicles as there is a far greater level of occupancy and a much larger number of vehicles pass through the location per minute. These observations are found to be repeatable on different working days and locations at the same time during the day.

It is important to note that the time of the day when these traffic regimes manifest themselves can be different for different locations and direction of travel. At the same location (i.e. Metric Blvd.) on US 183 south bound, sharp drops in traffic speed are observed during the morning hours (typically between 7:30 AM to 9:30 AM). The traffic speed is almost constant during other times.

During the *Free Regime*, there is little interaction between the vehicles and vehicles can be driven at the desired speeds for most of the time. An understanding of this regime is not critical to relieving congestion. In this dissertation, we have not tried to corroborate this regime with the US 183 traffic data.

The structure of vehicle following dynamics for traffic in the regimes 1 and 2 demand greater scrutiny and are discussed in the next chapter.

D. Data integrity issues

The inductive loop detector traffic data which have been utilized for the purposes of this dissertation has to be set up in right form to be useful for non-continuum model corroboration purposes. To that end, the traffic data should be meaningful and consistent. Upon careful analysis of the data archives many locations on US 183 were observed to have bad or inconsistent data consistently.

Table I shows the observed traffic data for three stations viz,: upstream, downstream and an entrance ramp between the upstream and the downstream stations on US 183 north bound. In the table, “Vol”, ”Occ.”, “Sp.”, “Tr.” denote the vehicle count, percentage occupancy, traffic speed in miles per hour and percentage of trucks respectively. The seven digit identifier (example “2000511”) denotes the loop detector. The following data integrity issues were observed:

1. Some stations consistently produced zero values for all the four available data

for all days and during all times.

2. In some instances (example for detector "2001015" in Table I) report "-1" for speed and percentage truck data while the vehicle count and percentage occupancy data are reported as "0". In such cases, it was not clear how to interpret these data given that the speed and truck percentage are flagged with a negative value. No information was available to determine if part of the data can be useful or if all the detector data should be considered as invalid.

3. Many detectors reported very low speeds even during early morning or late night hours during which very light traffic is expected.

4. In many instances inconsistent vehicle counts in terms of the total number of vehicles recorded at an upstream station does not match with the entrance/exit ramps and the vehicle count recorded at the downstream station. In almost all instances, there was a "loss" of vehicles between the upstream and the downstream stations. This "vehicle loss" was very common even for detector stations which otherwise seemed to report consistent and meaningful data.

E. Setting up the database for the corroboration of the proposed model

For the corroboration of the proposed traffic flow model, setting up the not-faulty and consistent traffic database is of utmost importance. To that end, the primary task was of selecting the site for the course of this study. After careful analysis of the traffic data, the stretch of US 183 from *Lamar Blvd.* to *Mopac* was selected. The main factors contributing to this selection were:

- It offered a stretch of US 183 freeway where at-least two sections (as per the section definition mentioned earlier) could be considered in tandem with "good"

Table I. Illustration of bad traffic data - Monday, May 3, 2004

Upstream Freeway Station - Chevy Chase Drive												
Time	2000511				2000512				2000513			
	Vol.	Occ.	Sp.	Tr.	Vol.	Occ.	Sp.	Tr.	Vol.	Occ.	Speed	Tr.
3:00pm	8	2	48	0	18	6	46	0	0	0	0	0
3:01pm	6	2	49	0	7	3	45	0	0	0	0	0
3:02om	9	3	48	0	16	3	46	0	0	0	0	0
3:03pm	9	9	51	0	16	5	47	0	0	0	0	0
3:04pm	20	6	49	0	18	6	49	0	0	0	0	0

Downstream Freeway Station - Carver Avenue												
Time	2001011				2001012				2001013			
	Vol.	Occ.	Sp.	Tr.	Vol.	Occ.	Sp.	Tr.	Vol.	Occ.	Speed	Tr.
3:00pm	0	0	0	0	0	0	0	0	0	0	0	0
3:01pm	0	0	0	0	0	0	0	0	0	0	0	0
3:02om	0	0	0	0	0	0	0	0	0	0	0	0
3:03pm	0	0	0	0	0	0	0	0	0	0	0	0
3:04pm	0	0	0	0	0	0	0	0	0	0	0	0

Entrance Ramp Station - Between Upstream and Downstream Station				
Time	2001015			
	Vol.	Occ.	Sp.	Tr.
3:00pm	0	0	-1	-1
3:01pm	0	0	-1	-1
3:02om	0	0	-1	-1
3:03pm	0	0	-1	-1
3:04pm	0	0	-1	-1

traffic available.

- There were both entrance and exit ramps in each of the two defined sections. This consideration was important in not selecting one other stretch of freeway which offered good traffic data.

- This stretch of freeway is reasonably long (about 6.6 lane miles or 2.2 miles). The length of the freeway to be considered is also an important factor in corroboration of the trip travel time estimation and prediction.

Fig. 11 and Fig. 12 show a map and the freeway configuration of the selected stretch of US 183 respectively.

Setting up the historical traffic data database:

Historical data has been used for identifying patterns in traffic behavior. For the purposes of traffic state forecasting (discussed later in the dissertation), a historical database was setup. Data corresponding to the loop detector stations on the selected stretch of the freeway was aggregated into a historic database. To that end, twenty five sets of data corresponding to each of working days (Monday - Friday) were selected from 2003 and the first five months of 2004. A simple mathematical mean of all the twenty five days data was taken for all the detector stations and was written into one file. Thus, we had five different files (one for each working day) which constituted the historical traffic data database.

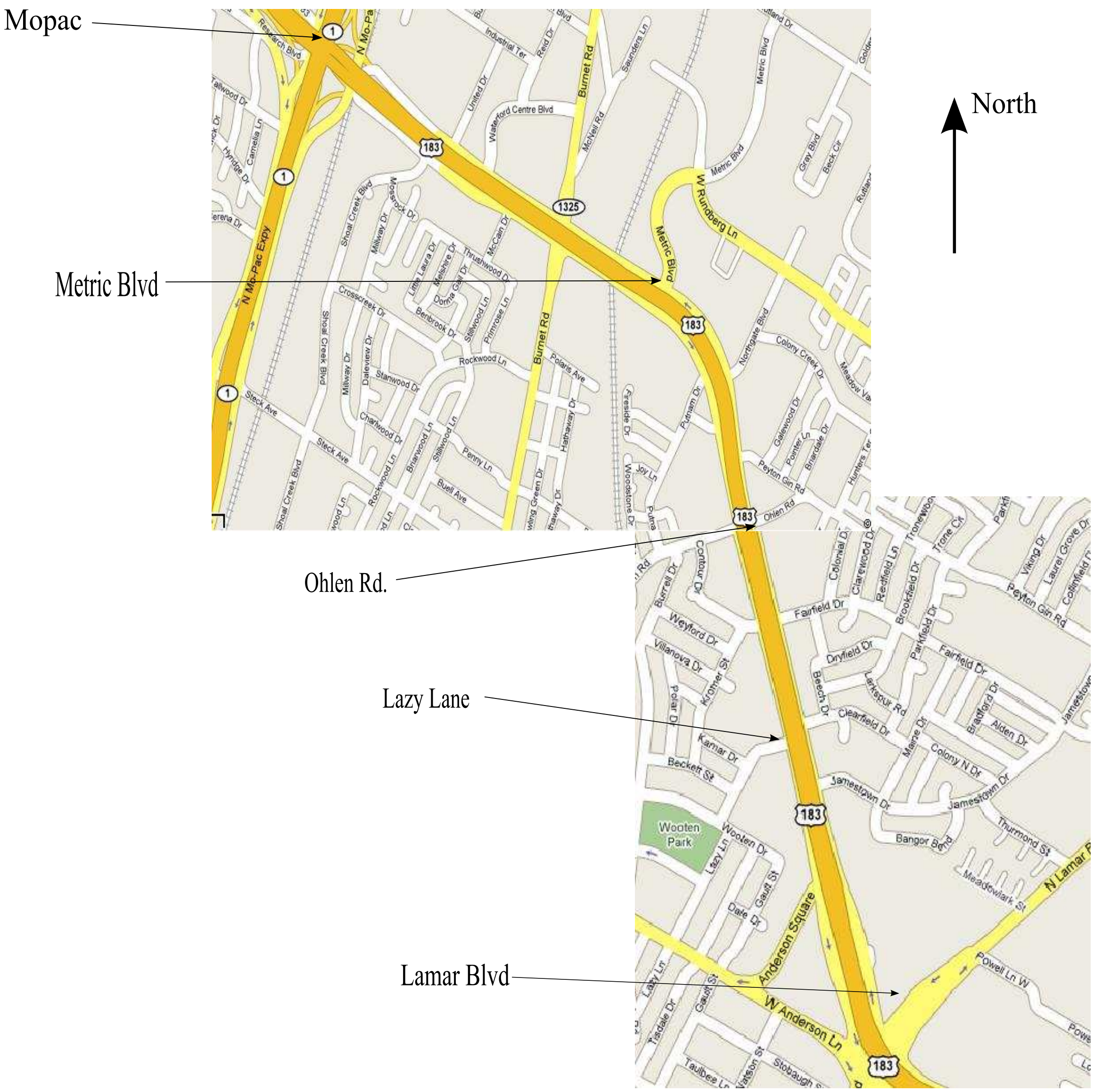

Fig. 11. Map of the selected stretch of US 183 freeway

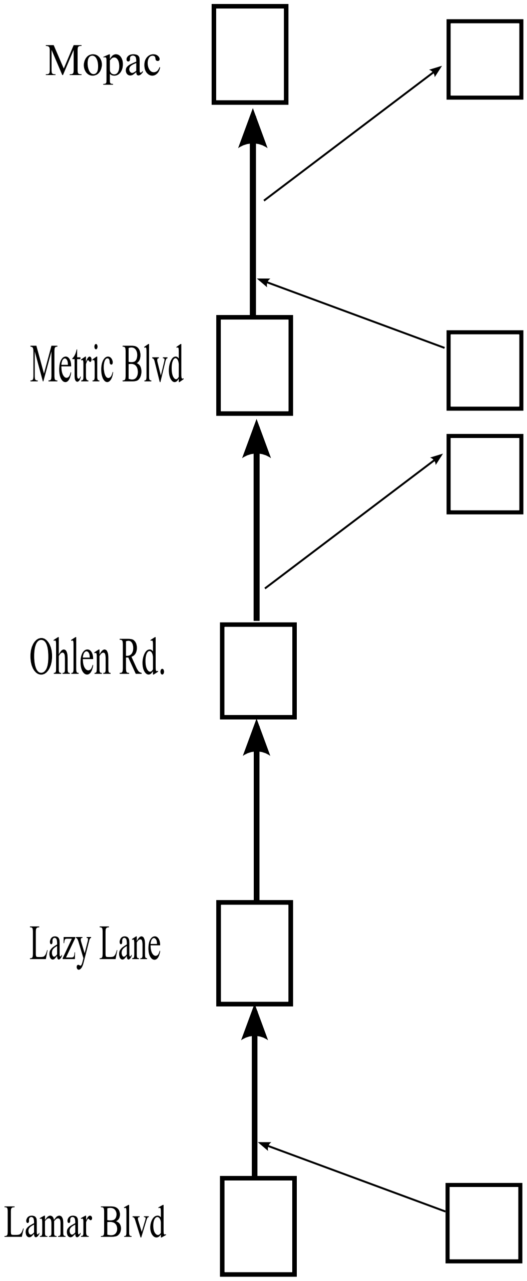

Fig. 12. Freeway configuration of the selected stretch of US 183 freeway

F. Summary

In this chapter we have described the available traffic data from US 183 in Austin, TX, USA. We have made some important and useful observations on macroscopic traffic flow patterns from this data. We have also discussed the various data quality issues and setting up the historical database which will be utilized in traffic state forecasting and is discussed more later in the dissertation.

CHAPTER IV

CORROBORATION OF THE NON-CONTINUUM TRAFFIC FLOW MODEL

A. Introduction

The traffic model that has been developed is primarily based on the idea of representing the aggregate behavior of the flow of traffic with a “representative” vehicle, whose dynamics is based on the interactions between the vehicles in the traffic. In this chapter we corroborate the non-continuum traffic model with the traffic data collected from the loop detectors deployed on US 183 freeway in Austin, TX. The corroboration procedure has been carried out in two steps, with the aim of corroborating the non-continuum model for a large stretch of the freeway. In the first step, we corroborate the non-continuum model for a single stand alone section and for the second step, we utilize the results and techniques from first step to corroborate the non-continuum model for two linked sections. To that end, the freeway is divided into consecutive sections as described earlier in the dissertation. Each section is defined to be the stretch of the freeway between three consecutive detector stations. The two extreme detector stations become the entrance and exit for the section of the freeway (depending on the direction of flow of traffic) that is considered and the data from the intermediate detector station is used to corroborate the model that is developed in this dissertation.

B. Single section corroboration

The process of corroboration consists of three steps described in the three following subsections. The first two steps are concerned with calibrating the model, i.e., estimating the parameters of vehicle following in different regimes from a few data sets collected on US 183 and the third step is involved with prediction of the state of traffic based on the calibrated parameters for other data sets of US 183. The estimation of parameters in the vehicle following model requires knowledge of the following distance, the number of vehicles in the section, net number of vehicles entering a section and speed, of which only the speed and the net number of vehicles entering the section are readily available. Since information concerning speed is readily available, we decouple the tasks of calibrating the parameters of vehicle following from that of estimating the aggregate following distance and the number of vehicles in a section. The latter is considered in the first subsection, while the former is considered in the second subsection.

1. Use of historic traffic data to identify the structure of vehicle following for *Regimes 1 and 2*

From the traffic loop detector data, repetitive patterns are observed in traffic throughput with respect to the time of the day and the day of the week. The traffic data for the number of vehicles passing through a location and their aggregate speeds are used to identify the structure of vehicle following for the traffic flow regimes 1 and 2. Since the traffic data are available at discrete intervals of time, for the non-continuum model (see Equation 2.11), presented earlier, we will use k to denote the discrete time instant under consideration and h to denote the time step size. Since the number of vehicles exiting a section in a time step is available from the traffic data, we introduce another state, $N^{ex}_{i,sum}$, to represent the cumulative number of vehicles that have exited the section. We will use $\bar{H}(i)$ to represent a unit Heaviside function, i.e., $\bar{H}(i) = 0 \ \forall i \leq 0$ and $\bar{H}(i) = 1, \ \forall i \geq 1$. The following state space model may then be constructed for the purpose of identification:

$$\begin{aligned}
N_i(k+1) =& N_i(k) + h\big[\dot{N}_i^{en}(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\
N_{i,sum}^{ex}(k+1) =& N_{i,sum}^{ex}(k) + h\big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] \\
\bar{\Delta}_i(k+1) =& \bar{\Delta}_i(k) - h\big\{\big[\frac{(L_{car}+\bar{\Delta}_i(k))^2}{L_{s,i}}\big]\big[\dot{N}_i^{en}(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\
& - \bar{H}(i-1)(\beta_{i,i-1}\bar{v}_{i-1}(k) - \bar{v}_i(k))\big\}
\end{aligned} \tag{4.1}$$

The value of the section index i in the above equations has to be set equal to one for the purposes of a single section corroboration. In other words, we treat this section as a stand alone one with no section upstream to it. Hence in Equation 4.1, $\bar{H}(i-1)$ equals zero and thus there is no contribution of the $(\beta_{i,i-1}\bar{v}_{i-1}(k) - \bar{v}_i(k))$ term to the aggregate following distance update equation.

The term $h[\dot{N}_i^{en}(k) + \dot{\tilde{n}}_i(k)]$ is the net inflow of vehicles into the section in time interval $[kh, (k+1)h)$ and is measurable. The system output is taken to be $N_{i,sum}^{ex}$, the cumulative number of vehicles exiting the section. Thus, the model assumes that the rate of change of number of vehicles with respect to time in a section is known. Since, we do not know the initial conditions of the states N_i, $N_{i,sum}^{ex}$, and $\bar{\Delta}_i$, we design a state estimator using *Extended Kalman Filtering* technique. Similar approaches were used earlier to predict the flow of traffic for continuum models of traffic [57] - [60].

Below we describe the basic procedures involved in the extended Kalman filtering and its application, in identification purposes[1].

[1] Detailed definitions and proofs for state space modeling and Kalman filtering are provided in the Appendix A.

a. Extended Kalman Filtering:

The Kalman Filtering procedure for non-linear models involves computing, in real-time, the Taylor series approximation of the system function at the previous state estimate and that of the observation function at the corresponding predicted position. The Kalman filter so obtained is called the *extended Kalman filter*. While there are no guarantees of convergence of the estimates of the states to their true values, it still remains an effective tool as far as practice is concerned.

Below we describe the basic extended Kalman filtering equations and algorithm for a nonlinear model of a system [61]. Consider a non linear system in the following form:

$$\begin{aligned} \mathbf{x}_{k+1} =& \mathbf{f}_k(\mathbf{x}_k) + H_k(\mathbf{x}_k)\underline{\xi}_k \\ \mathbf{v}_k =& \mathbf{g}_k(\mathbf{x}_k) + \underline{\eta}_k \end{aligned} \tag{4.2}$$

where $\mathbf{x}$ is the state vector, $\mathbf{v}$ is the output vector, $\mathbf{f}_k$ and $\mathbf{g}_k$ are vector valued functions with ranges in $\Re^n$ and $\Re^q$ respectively and $1 \leq q \leq n$ and H_k is a matrix valued function with range in $\Re^n \times \Re^q$, such that for each k the first order partial derivatives of $\mathbf{f}_k(\mathbf{x}_k)$ and $\mathbf{g}_k(\mathbf{x}_k)$ with respect to all the components of $\mathbf{x}_k$ are continuous. Also, we consider zero mean Gaussian white noise sequences $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ with ranges in $\Re^p$ and $\Re^q$ respectively and $1 \leq p, q \leq n$.

Assume that the initial state of the system $\mathbf{x}_0$ has a known initial estimate, $\hat{\mathbf{x}}_0$, and variance Σ_0. We make two additional assumptions. First, the transition and measurement errors ($\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ respectively) are uncorrelated. This is reasonable since they arise from two different processes. Second we assume that the initial state, $\mathbf{x}_0$, is independent of the errors $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$. Again, this does not seem unreasonable. Putting the above stated assumptions mathematically, we get:

$$E(\underline{\xi}_k\underline{\xi}_l^T) = Q_k\delta_{kl} \qquad\qquad E(\underline{\eta}_k\underline{\eta}_l^T) = R_k\delta_{kl} \tag{4.3}$$

$$E(\underline{\xi}_k\underline{\eta}_l^T) = 0 \qquad E(\underline{\xi}_k\mathbf{x}_0^T) = 0 \qquad E(\underline{\eta}_k\mathbf{x}_0^T) = 0 \tag{4.4}$$

where Q_k and R_k are the variance matrices for random vectors $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ respectively, E denotes the expectation and the above conditions are satisfied for all k and l. Using the assumptions stated above, the following results about extended Kalman filtering can be stated[2]:

$$P_{0,0} = \Sigma_0 = Var(\mathbf{x}_0), \quad \hat{\mathbf{x}}_0 = E(\mathbf{x}_0) \tag{4.5}$$

$$For \quad k = 1, 2, ..., \tag{4.6}$$

$$P_{k,k-1} = \Big[\frac{\partial \mathbf{f}_{k-1}}{\partial \mathbf{x}_{k-1}}(\hat{\mathbf{x}}_{k-1})\Big] P_{k-1,k-1} \Big[\frac{\partial \mathbf{f}_{k-1}}{\partial \mathbf{x}_{k-1}}(\hat{\mathbf{x}}_{k-1})\Big]^T + H_{k-1}(\hat{\mathbf{x}}_{k-1}) Q_{k-1} H_{k-1}^T(\hat{\mathbf{x}}_{k-1}) \tag{4.7}$$

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{f}_{k-1}(\hat{\mathbf{x}}_{k-1}) \tag{4.8}$$

$$G_k = P_{k,k-1}\Big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\Big]^T \cdot \left[\Big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\Big] P_{k,k-1} \Big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\Big]^T + R_k\right]^{-1} \tag{4.9}$$

$$P_{k,k} = \left[I - G_k\Big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\Big]\right] P_{k,k-1} \tag{4.10}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1})). \tag{4.11}$$

In Kalman filtering terminology, $\hat{\mathbf{x}}_{k|k-1}$ represents a *one-step* prediction of the state $\mathbf{x}_k$. It represents the best knowledge of the values of the states at the time instant

[2] The notation $i|j$ indicates an estimate corresponding to the time instant i based on the information up to and including time instant j.

k, prior to obtaining the measurement values for the k^{th} time instant. Equation 4.8 shows how this can be obtained. $P_{k,k-1}$ and $P_{k,k}$ represent the variances of $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$. Equation 4.7 shows how $P_{k,k-1}$ depends on both the uncertainty in $\hat{\mathbf{x}}_{k-1|k-1}$ as well as the variances of the error Q_{k-1}.

The matrix G_k is called the *Kalman gain* matrix. Its interpretation becomes clear from Equation 4.11. Equation 4.11 shows that the *filtered* estimate $\hat{\mathbf{x}}_{k|k}$ can be represented as a sum of two terms. The first one is simply the one-step prediction (prior estimate) $\hat{\mathbf{x}}_{k|k-1}$. The second term represents the adjustment to be applied to the prior estimate in the light of the new measurements $\mathbf{v}_k$ that have just been obtained. The term $\mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1})$ represents, in a sense, a *predicted* cumulative sum of the number of vehicles exiting the section. $\mathbf{v}_k$ represents the cumulative sum of vehicles exiting the section actually measured by the loop detectors. $(\mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1}))$ therefore represents a "residual". In filtering theory, this sequence of residuals is termed as the *innovations* sequence. These innovations represent the "new" information in each measurement ($\mathbf{v}_k$, in our case). The Kalman gain G_k can now be interpreted as the weight given to the new information. From Equation 4.9, we can see that as the variance R_k increases, G_k decreases and the weight given to this new information decreases as it should. We now discuss how the extended Kalman filtering methodology can be applied to identify the structure of vehicle following for different traffic flow regimes.

b. Structure of vehicle following dynamics for traffic under different regimes:

Since the extended Kalman filtering technique requires the smoothness of the vector fields, $\mathbf{f}$ and $\mathbf{g}$, it is apparent that such a scheme cannot be applied in principle if there is frequent switching between different regimes of traffic. However, from the traffic data, one can observe that the duration of each regime is long enough for one to treat the problem of filtering for each regime individually. As long as smoothness of the vector field $\mathbf{f}$ is guaranteed in each regime and there is infrequent switching between different regimes, it is reasonable to circumvent the requirement of smooth $\mathbf{f}$ by considering problems of filtering for each regime separately. This is the approach we adopt here.

We apply the extended Kalman filtering methodology to identify the states in the state space model in Equation 4.1, with the value of i being set equal to one. The following state space system is constructed (in representation consistent with Equation 4.2):

$$\mathbf{x}_k = \begin{bmatrix} N_i(k) & N^{ex}_{i,sum}(k) & \bar{\Delta}_i(k) \end{bmatrix}^T$$

$$\mathbf{f}_k = \begin{bmatrix} N_i(k) + h\big[\dot{N}^{en}_i(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\ N^{ex}_{i,sum}(k) + h\big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] \\ \bar{\Delta}_i(k) - h\Big\{\big[\frac{(L_{car}+\bar{\Delta}_i(k))^2}{L_{s,i}}\big]\big[\dot{N}^{en}_i(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\ -\bar{H}(i-1)(\beta_{i,i-1}\bar{v}_{i-1}(k) - \bar{v}_i(k))\Big\} \end{bmatrix}$$

$$H_k = \begin{bmatrix} h & 0 & 0 \\ 0 & h & 0 \\ 0 & 0 & h \end{bmatrix} \tag{4.12}$$

$$\underline{\xi}_k = \begin{bmatrix} \xi^{N_i}_k & 0 & \xi^{\bar{\Delta}_i}_k \end{bmatrix}^T$$

with h being the time step. The cumulative count of the number of vehicles exiting the section at each time step is used as the output. Thus we have:

$$\mathbf{v}_k = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \mathbf{x}_k + \eta_k$$

where $\mathbf{v}_k$ is the output vector. Use of the Gaussian white noise sequences ($\xi_k^{N_i}$ and $\xi_k^{\bar{\Delta}_i}$) in the state transition equation helps us to treat the aggregate following distance and the number of vehicles in the section as different state variables. The aggregate speed data is used to estimate the states at each instant of time.

For all the simulation purposes, a fixed vehicle length of fifteen feet has been used. We now give the following structure to vehicle following. That is, we give a structure to the function f in the following equation:

$$\dot{v} = f(v, \Delta, \dot{\Delta})$$

The following structure of vehicle following for the different traffic regimes is hypothesized:

1. **Regime 1:** Observed from 7:00 AM to 10:00 PM except the time duration when there is a sharp drop in the traffic speed. The desired aggregate following distance in this regime is hypothesized to be linearly proportional to the speed of the vehicles and the equation describing the evolution of speed of traffic may be expressed as:
$$\dot{v}_i = \frac{1}{h_w}[\dot{\Delta}_i + \lambda_1(\Delta_i - h_w v_i)]$$

2. **Regime 2:** Observed between 4:30 PM to 6:30 PM. This regime can be further subdivided into two sub-regimes: (a) when the traffic speed drops sharply to a low speed and fluctuates about this minimum for some time and, (b) when the traffic speed again increases to a high speed (regime changes again to *Regime 1* after this). In this regime, we hypothesize a non-linear relationship between

the desired following distance and speed, i. e., $\bar{\Delta}_{des} = q(\bar{v})$.

$$\dot{v}_i = \frac{1}{\frac{dq}{dv}}[\dot{\Delta}_i + \lambda_2(\Delta_i - q(v))]$$

In traffic *Regime 1*, it is hypothesized that the drivers tend to maintain a following distance that varies linearly with the speed of the vehicle. This hypothesis reflects the observation that the drivers tend to accept smaller following distances while maintaining almost constant speeds. This happens when there is an increase in the demand on utilization of the freeway (more number of vehicles are present).

During *Regime 2*, it is hypothesized that the drivers react more sharply to the increasing demand on freeway utilization and are not able to maintain high speeds. As a result, aggregate traffic speed drops and drivers are able to maintain much smaller following distances, but with slower speeds. Since $q(v)$ associated with *Regime 2* is assumed to be smooth for EKF to be applicable, and since its structure is not known, a reasonable starting point is to express it using the first few terms of its Taylor's series:

$$\bar{\Delta}^{des} = \alpha_1 v^1 + \alpha_2 v^2 + ... + \alpha_p v^p. \tag{4.13}$$

Since the value of p in Equation 4.13 (highest order term) is unknown, we start by choosing its value as one and we increase it till we get a good fit for the estimated following distance. The following distance to be used in the Equation 4.13 is obtained when we estimate the states of the state space model in Equation 4.1. For many "Mondays" (for example) it was observed that for p = 1, 2, 3, 4 we do not get a good fit. p = 5 gives us a reasonably good fit. Terms higher than fifth order in the Taylor's series do not substantially improve the fit. Numerical accuracy issues also arise for higher order terms during parameter identification due to the units of following distance (feet) and speed (feet per minute). In Fig. 13 we plot six different predictions labelled according to highest order term till which the Taylor series was considered for *Regime 2a*.

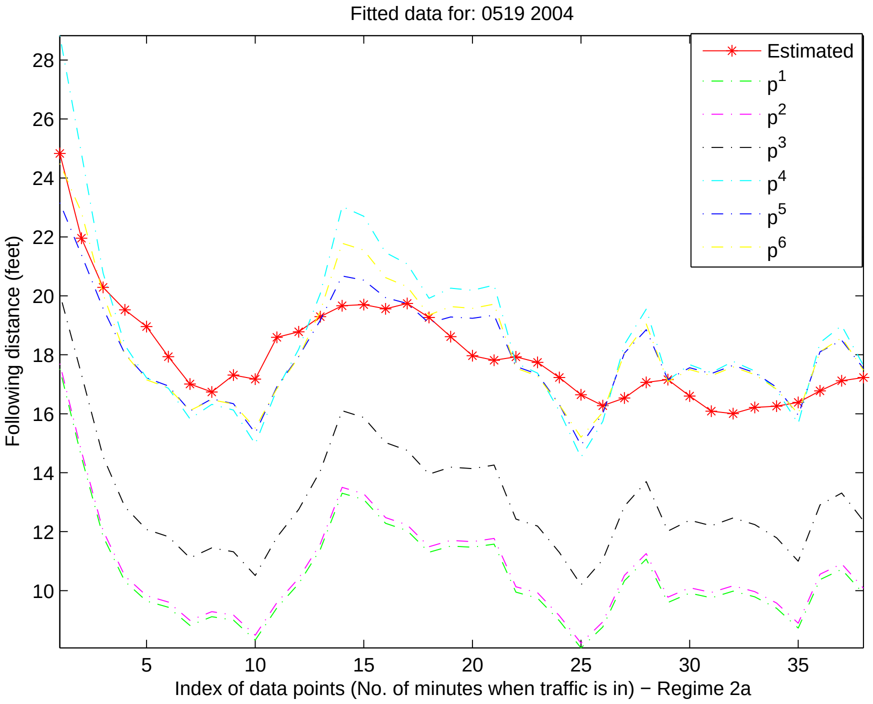

Fig. 13. Fitting data to identify structure of vehicle following for Regime 2a

After fixing a value of p, further analysis of the Taylor series for the non linear relationship between the desired following distance and the speed shows that the contributions from the lower order terms is not very significant. It was observed from the historical data that the contributions of lower order terms till $p - 1$ were not significant. Fig. 14 shows the percentage contribution of terms of different powers for data for six Monday(s) in 2004. For ease of maintaining traffic parameters and reducing the computational effort, only the p^{th} power term was considered in the relationship between the desired following distance and the speed as shown below:

$$\bar{\Delta}^{des} = \alpha v^p$$

The vehicle following law can then be synthesized as:

$$\dot{v}_i = \frac{1}{\alpha p v_i^{p-1}}[\dot{\Delta}_i + \lambda_2(\Delta_i - \alpha v_i^p)]$$

For *Regime 2a* and *Regime 2b* different values of α and p have to be used. Different values for α and p also make eminent sense from point of view of the observed traffic hysteresis as different values of aggregate traffic speed are observed for the same values of the occupancy levels during the congestion onset (*Regime 2a*) and recovery (*Regime 2b*) phases. In Fig. 15 we show the above mentioned hysteretic behavior observed in traffic.

In the above equations, the parameters, λ_1, λ_2 reflect the time constants associated with driving in their respective traffic regimes. These along with h_w (for *Regime 1*), α_{2a}, p_{2a} and α_{2b}, p_{2b} (for *Regime 2a and 2b* respectively) are the parameters to be estimated from the traffic data. The numerical values of the parameters associated with this structure are specific to the highway section under consideration. This structure of vehicle following dynamics is a constitutive relationship and can be modified to fit experimental data (traffic data). It should be noted that, values of estimated aggregate following distance are used in estimating the traffic parameters. The aggregate following distance is estimated by applying extended Kalman filtering algorithm to the state space model in Equation 4.12 as described earlier in this section.

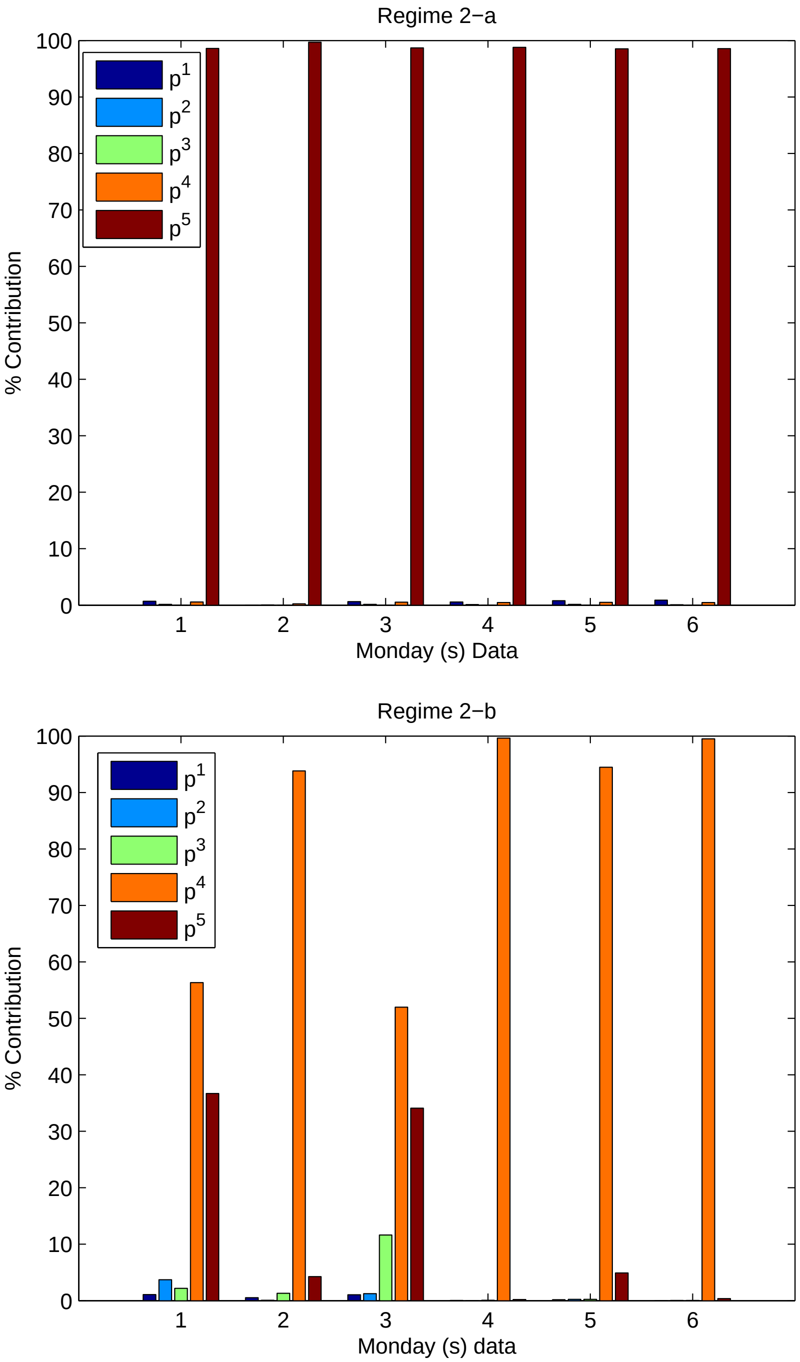

Fig. 14. Contribution of the terms at various orders of the power series for Regime 2-a and 2-b (Monday(s) data)

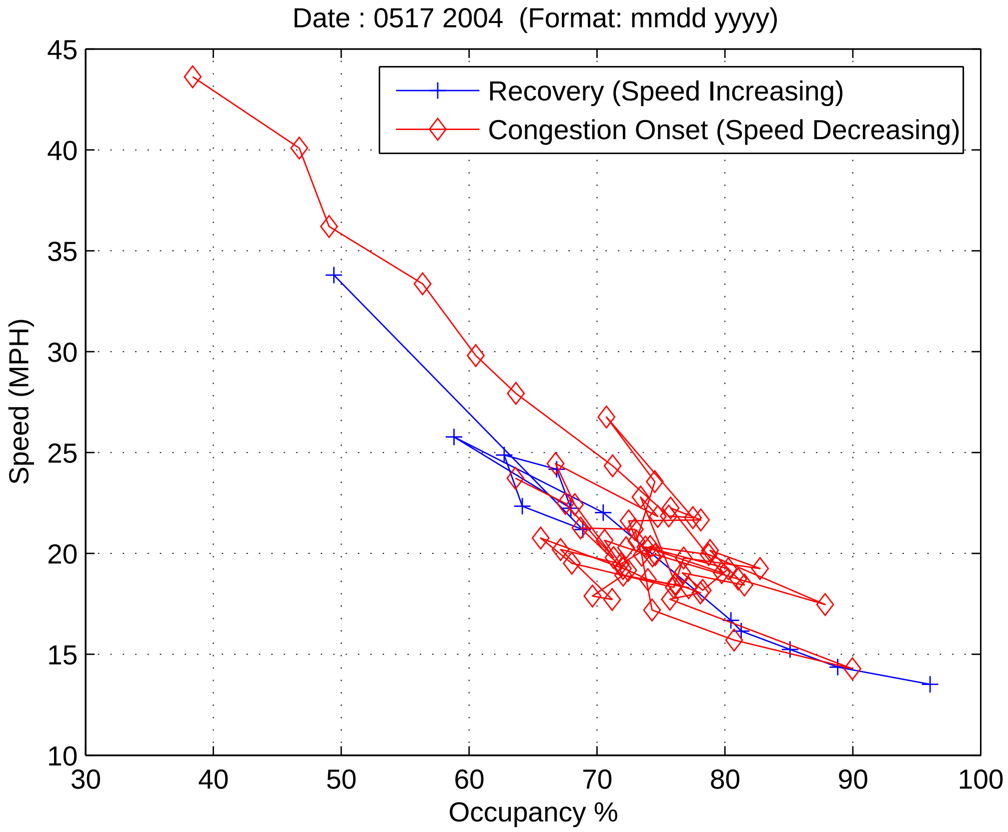

Fig. 15. Observed traffic hysteresis

2. Estimating the traffic parameters associated with different traffic regimes

In this subsection, we estimate the time constants λ_1, λ_2, aggregate time headway h_w associated with the traffic *Regime 1* and the parameters α and p for *Regime 2.* The values of estimated traffic states - the aggregate following distance and the number of vehicles in the section - and the aggregate speed data are used to estimate the parameters. For parameter estimation, objective functions are synthesized from the constitutive relationship for the vehicle following behavior of different traffic regimes. For example, in *Regime 1*, velocity updates can be obtained as

$$\bar{v}_i(k+1) = \frac{1}{h_w}\big(\bar{\Delta}_i(k+1) + (\lambda_1 h - 1)\bar{\Delta}_i(k)\big) - \bar{v}_i(k)(\lambda_1 h - 1).$$

Since, we have the aggregate speed data, we can then formulate the objective function to be minimized as

$$Minimize \qquad \sum_k \left(1 - \frac{\bar{v}_i(k+1)}{y(k+1)}\right)^2$$

where $y(k+1)$ is the corresponding speed data for *Regime 1.* Thus the objective functions are developed to be minimized in a least squares sense. The above mentioned objective function cannot be used if the traffic has stalled ($y(k) = 0$), and we will have to use a different objective function. Stalled traffic typically is a case of non-recurring congestion and can happen, for example if there a vehicle has broken down on the freeway or if there has been an accident. For recurring congestion, traffic data suggests sharp drops in aggregate traffic speeds, but zero traffic speed was observed only on very few days in the years 2003 and 2004. Since we are concerned with modeling repeatable traffic patterns (recurring congestion for example), we have neglected the days when there was stalled traffic for parameter estimation purposes.

A MATLAB application on multidimensional unconstrained nonlinear minimization approach based on Nelder-Mead simplex [62] method was used to minimize the non linear objective function. The estimated parameters show consistency over different working days. We list some sample estimated parameters in the results section.

3. Predicting traffic state in real-time

Estimated time constants and other parameters (h_w, α and p) associated with the different traffic regimes from historical data, are then used to predict the traffic state in real-time. Extended Kalman filtering is again used to estimate the states of traffic in real-time but now the state vector also includes the aggregate speed of the traffic (speed of the "representative" vehicles for the particular section). For example, for Regime 1, the state space representation of the traffic model (in the representative form consistent with Equation 4.2) will be:

$$\mathbf{x}_k = \begin{bmatrix} N_i(k) & N^{ex}_{i,sum}(k) & \bar{\Delta}_i(k) & \bar{v}_i(k) \end{bmatrix}^T$$

$$\mathbf{f}_k = \begin{bmatrix} N_i(k) + h\big[\dot{N}^{en}_i(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\ N^{ex}_{i,sum}(k) + h\big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] \\ \bar{\Delta}_i(k) - h\big\{\big[\frac{(L_{car}+\bar{\Delta}_i(k))^2}{L_{s,i}}\big]\big[\dot{N}^{en}_i(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\ -\bar{H}(i-1)(\beta_{i,i-1}\bar{v}_{i-1}(k) - \bar{v}_i(k))\big\} \\ \bar{v}_i(k) + \frac{h}{h_w}\big\{\big(\big[\frac{(L_{car}+\bar{\Delta}_i(k))^2}{L_{s,i}}\big]\big[\dot{N}^{en}_i(k) - \big[\frac{\bar{v}_i(k)N_i(k)}{L_{s,i}}\big] + \dot{\tilde{n}}_i(k)\big] \\ -\bar{H}(i-1)(\beta_{i,i-1}\bar{v}_{i-1}(k) - \bar{v}_i(k))\big) + \lambda_1(\bar{\Delta}_i(k) - h_w\bar{v}_i(k))\big\} \end{bmatrix}$$

$$H_k = \begin{bmatrix} h & 0 & 0 & 0 \\ 0 & h & 0 & 0 \\ 0 & 0 & h & 0 \\ 0 & 0 & 0 & h \end{bmatrix} \tag{4.14}$$

$$\underline{\xi}_k = \begin{bmatrix} \xi_k^{N_i} & 0 & \xi_k^{\bar{\Delta}_i} & \xi_k^{\bar{v}_i} \end{bmatrix}$$

$$\mathbf{v}_k = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \mathbf{x}_k + \eta_k$$

The input to the system is taken to be the net inflow of vehicles into the section, $h[\dot{N}_i^{en}(k) + \dot{\hat{n}}_i(k)]$, and the observed occupancy percentage. The output of the system is again taken to be the cumulative number of vehicles exiting the section. The observed percent occupancy is used in determining the traffic regime transitions as discussed earlier in the dissertation. Fig. 16 shows a traffic regime transition diagram based on the input occupancy levels. The traffic flow regime changes from *Regime 1* to *Regime 2a* when the occupancy becomes greater than some critical occupancy (denoted by "$Occ_{threshold}$" in Fig. 16) for the section under consideration. Once, when the traffic has already transitioned into *Regime 2a*, the regime transition between regimes *2a* and *2b* seems to take place based on the net number of vehicles entering the section. That is, the traffic regime changes from *2a* to *2b* when the number of vehicles exiting are more than the number of vehicles entering the section. Equivalently, the traffic can again switch to *Regime 2a* from *Regime 2b* if there are more number of vehicles entering the section than exiting. This fluctuation between *Regime 2a* and *Regime 2b* was observed to happen less frequently. This threshold value of the net number of vehicles entering the section which determines the switch between *Regime 2a* and *Regime 2b* is location specific and different sections will have a different threshold values (in Fig. 16 we denote this threshold value by "$num_{threshold}$ "). The traffic finally goes back to *Regime 1* once the occupancy levels fall below the critical occupancy for the section under consideration.

To corroborate the predicted traffic states, the predicted aggregate speed and the predicted number of vehicles at a detector station are plotted against the observed

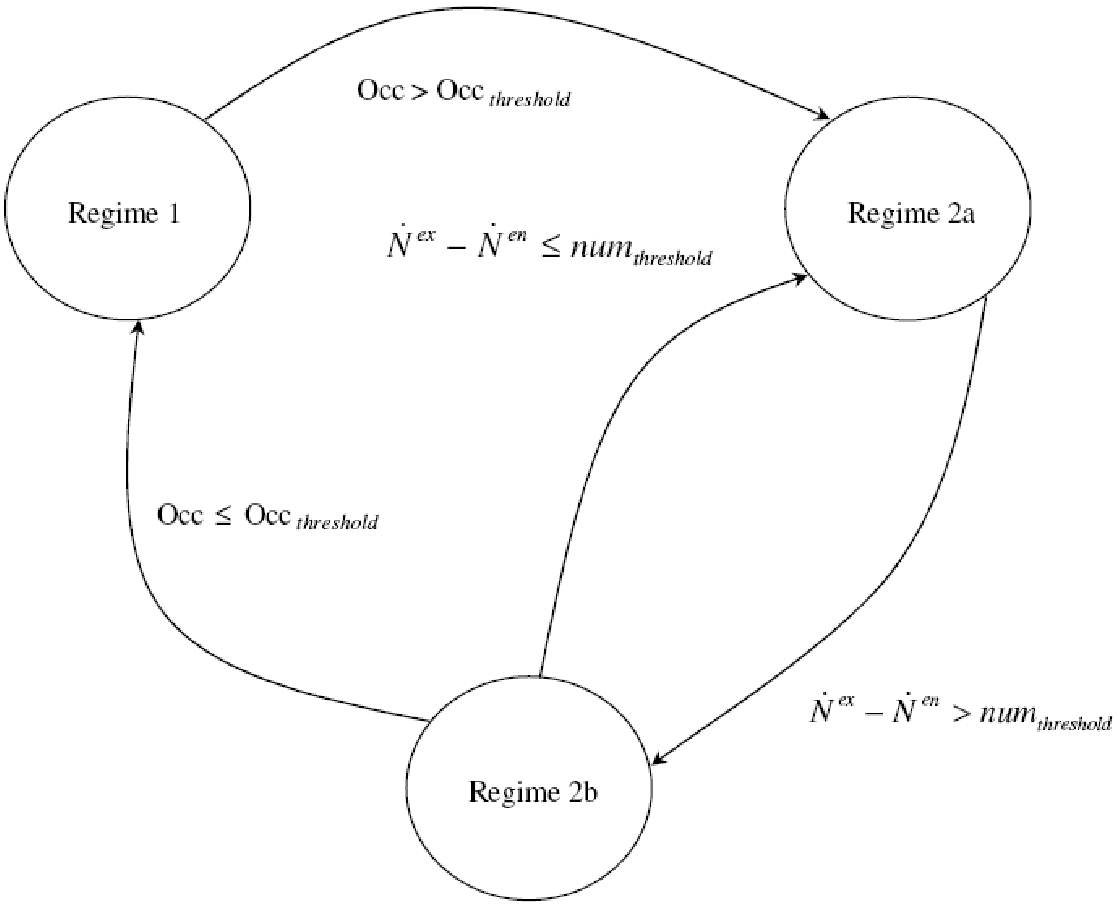
Regime 1
Regime 2a
Regime 2b
$\mathrm{Occ} > \mathrm{Occ}_{threshold}$
$\dot{N}^{ex} - \dot{N}^{en} \leq num_{threshold}$
$\mathrm{Occ} \leq \mathrm{Occ}_{threshold}$
$\dot{N}^{ex} - \dot{N}^{en} > num_{threshold}$

Fig. 16. Schematic of the traffic regime transition

values for a particular day. By applying the extended Kalman filtering technique with the data update frequency of one minute, we can make traffic state predictions for the next minute.

4. Results

In this section we provide some results for one minute traffic state predictions and typical values of the estimated traffic parameters for the following two stretches (sections) on US 183 north bound freeway in Austin. We also provide some other important details for the two sections, pertaining to corroboration purposes.

1. Section 1: From Ohlen Road (upstream station) to Mopac (downstream station).
    - The section is 3.75 lane miles long (or 1.25 miles actual distance).
    - The critical threshold occupancy level for traffic state transition between *Regime 1* and *Regime 2a* (or between *Regime 2b* and *Regime 1*) is observed to be 35%.
    - The threshold value ($num_{threshold}$ in Fig. 16) for net number of vehicles entering the section for regime switch between *Regime 2a* and *Regime 2b* is 10. That is, $\dot{N}^{ex} - \dot{N}^{en}$ has to be greater than 10 for regime transition from *Regime 2a* to *Regime 2b* or equivalently less than -10 for regime transition from *Regime 2b* to *Regime 2a*
2. Section 2: From Lamar Blvd. (upstream station) to Ohlen Road (downstream station). The corresponding threshold values (mentioned in the same order as above) for this section are:
    - The length of the section is 2.85 lane miles (or 0.95 miles actual distance).
    - Critical threshold occupancy level is 42%.
    - Threshold net number of vehicles is 15 for regime transition from *Regime 2a* to *Regime 2b*.

a. Estimated values for the parameters characterizing the flow of traffic in the different regimes

Below we give some typical values for the parameters that characterize the flow of traffic associated with *Regime 1* and *Regime 2* for both sections 1 and 2.

1. *Regime 1*: The vehicle following control law hypothesized for traffic *Regime 1* is:

$$\dot{v}_i = \frac{1}{h_w}[\dot{\Delta}_i + \lambda_1(\Delta_i - h_w v_i)]$$

Table II and Table III show the estimated values for h_w and λ_1 for section 1 and section 2 respectively for four different weeks. The units of λ_1 and h_w are $minute^{-1}$ and *minute* respectively. The parameters show repetitive trends with respect to the day of the week for both the sections. It should be noted that the values of both λ_1 and h_w for both the sections are pretty close. This is reasonable to expect, since the aggregate speed of traffic remains almost constant in *Regime 1* in both the sections. Also, the aggregate traffic speed during *Regime 1* is about the same in both sections 1 and 2.

The time headway (h_w) varies from about 9 to 24 seconds for both the sections. For example, on Mondays' in section 1, the time headway is about thirteen seconds. It can also be noted that the time headway is smaller on Mondays' and Fridays' than those compared with the other three working days of the week. These numbers for time headway are pretty reasonable for the high speeds observed in *Regime 1*. The small values of λ_1's and typical values of time headway and speed in the section suggests that drivers are ready to accept shorter following distances so that they can drive with almost constant speeds. Also, it can be observed that the rate of change of speed is more dependent on the relative speed ($\dot{\Delta}$).

Table II. Estimated parameters for *Regime 1* - Section 1

Regime 1								
Monday			Tuesday			Wednesday		
Date	λ_1	h_w	*Date*	λ_1	h_w	*Date*	λ_1	h_w
03-01-04	1.2E-7	0.170	03-02-04	9.9E-4	0.375	03-03-04	10.8E-4	0.212
03-08-04	1.3E-7	0.244	03-09-04	6.1E-4	0.288	03-10-04	8.9E-4	0.278
04-05-04	5.9E-8	0.236	04-06-04	7.8E-4	0.398	04-07-04	9.1E-4	0.319
05-03-04	1.3E-7	0.238	05-04-04	4.4E-4	0.300	05-05-04	6.7E-4	0.301

Thursday			Friday		
Date	λ_1	h_w	*Date*	λ_1	h_w
03-04-04	5.9E-4	0.349	03-05-04	1.3E-7	0.223
03-11-04	7.6E-4	0.253	03-12-04	1.3E-7	0.215
04-08-04	5.8E-4	0.294	04-09-04	1.9E-7	0.243
05-06-04	8.5E-4	0.3000	05-07-04	8.8E-8	0.191

Table III. Estimated parameters for *Regime 1* - Section 2

Regime 1								
Monday			Tuesday			Wednesday		
Date	λ_1	h_w	*Date*	λ_1	h_w	*Date*	λ_1	h_w
03-01-04	1.8E-7	0.167	03-02-04	9.2E-4	0.302	03-03-04	9.5E-4	0.210
03-08-04	1.5E-7	0.203	03-09-04	5.8E-4	0.298	03-10-04	9.1E-4	0.288
04-05-04	3.6E-7	0.200	04-06-04	8.3E-4	0.383	04-07-04	8.6E-4	0.332
05-03-04	1.1E-7	0.189	05-04-04	6.9E-4	0.325	05-05-04	7.8E-4	0.265

Thursday			Friday		
Date	λ_1	h_w	*Date*	λ_1	h_w
03-04-04	6.3E-4	0.301	03-05-04	1.7E-7	0.222
03-11-04	7.1E-4	0.248	03-12-04	1.6E-7	0.254
04-08-04	6.8E-4	0.285	04-09-04	2.1E-7	0.260
05-06-04	8.3E-4	0.199	05-07-04	5.9E-8	0.210

2. Regime 2: The vehicle following control law hypothesized for traffic *Regime 2* is:

   $$\dot{v}_i = \frac{1}{\alpha p v_i^{p-1}}[\dot{\Delta}_i + \lambda_2(\Delta_i - \alpha v_i^p)]$$

   Table IV and Table V show the estimated values for α, p and λ_2 for *Regime 2a* for four weeks for section 1 and section 2 respectively.

   For *Regime 2b*, smaller values of p are observed. Typically they are one less than the p's for the *Regime 2a*. For example, for section 1 on Monday's, p_{2b} is equal to 4. The, α_{2b}'s are also correspondingly smaller. This is due to the fact that the recovery from congestion is quite fast. Traffic remains in the congested regime (*Regime 2a*) for a larger time duration and when the recovery phase starts, the traffic is able to achieve faster speeds in a relatively smaller time duration.

Another interesting point to note is that we have not given estimates for the appropriate parameters for the flow of traffic in *Regime 2*, for Fridays. Friday, being the end of the working week, people tend to leave the offices all day long, after lunch. So there is no excess demand for freeway utilization. In fact it was observed that for some Fridays, slight congestion occurs around noon time. This point was also elaborated previously in the chapter on data analysis. This again stresses the point that traffic parameters are *location sensitive*, but the constitutive relations for the flow of traffic still remain the same.

Table IV. Estimated parameters for *Regime 2a* - Section 1

Regime 2a							
Monday				Tuesday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-01-04	1.2E-15	5	0.035	03-02-04	1.6E-8	3	0.384
03-08-04	2.4E-15	5	0.025	03-09-04	1.1E-8	3	0.655
04-05-04	5.8E-15	5	0.309	04-06-04	2.8E-8	3	0.238
05-03-04	2.6E-15	5	0.386	05-04-04	1.4E-8	3	0.313
Wednesday				Thursday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-03-04	1.4E-15	5	0.019	03-04-04	1.2E-15	5	0.016
03-10-04	1.4E-15	5	0.028	03-11-04	1.3E-15	5	0.028
04-07-04	1.2E-15	5	0.047	04-08-04	2.6E-15	5	0.028
05-05-04	1.7E-15	5	0.002	05-06-04	9.5E-15	5	0.442

Table V. Estimated parameters for *Regime 2a* - Section 2

Regime 2a							
Monday				Tuesday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-01-04	4.7E-12	4	0.241	03-02-04	1.9E-8	3	0.445
03-08-04	3.1E-12	4	0.118	03-09-04	2.1E-8	3	0.585
04-05-04	7.6E-12	4	0.203	04-06-04	2.6E-8	3	0.608
05-03-04	5.4E-12	4	0.289	05-04-04	1.7E-8	3	0.493
Wednesday				Thursday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-03-04	5.5E-12	4	0.103	03-04-04	3.1E-12	4	0.276
03-10-04	6.0E-12	4	0.098	03-11-04	2.8E-12	4	0.311
04-07-04	5.9E-12	4	0.146	04-08-04	2.6E-12	4	0.378
05-05-04	4.9E-12	4	0.206	05-06-04	4.1E-12	4	0.402

b. Traffic state prediction

In this subsection we provide results concerning the state of traffic. We plot the predicted aggregate traffic speed against the actual observed/measured aggregate speed of traffic. We also corroborate the predicted number of vehicles passing per minute at a particular detector station with the collected data.

Fig. 17 and Fig. 18 show the predictions after one minute on June 14th, 2004 and July 15th, 2004 for section 1. Fig. 19 show the predictions after one minute on June 14th, 2004 for section 2. For both the sections we predict both the aggregate traffic speed and the number of vehicles passing through a detector station (downstream end detector station). On the same plots (for speed and number of vehicles) we plot the actual traffic speeds and the vehicle count for the corresponding time of the day.

From the plots we observe that the non-continuum traffic flow model is able to predict the traffic state very well. Some spikes can be observed in the prediction of the number of vehicles crossing the detector station. These are due to the traffic flow regime switching.

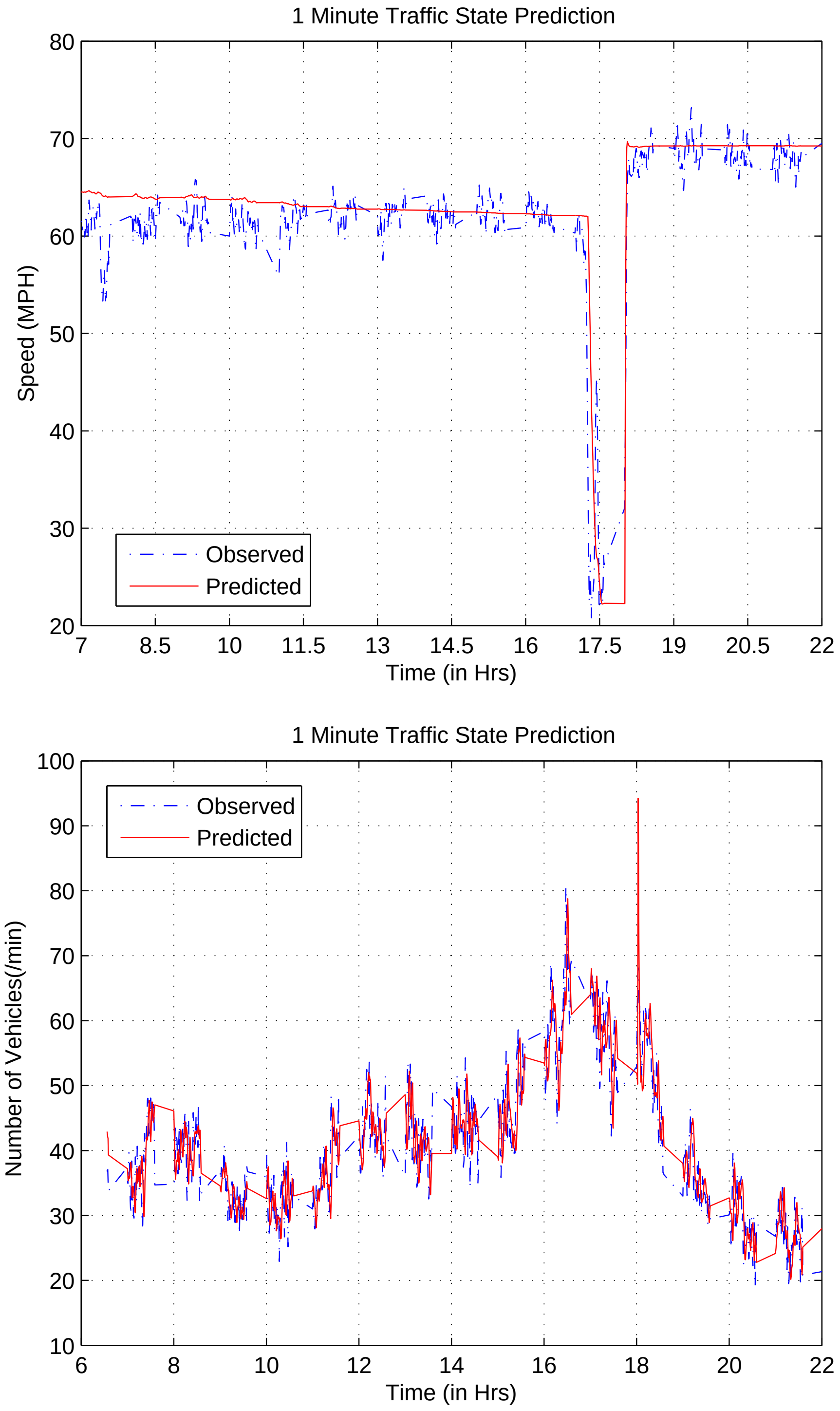

Fig. 17. 1 minute prediction of state of traffic, Section 1, June 14, 2004

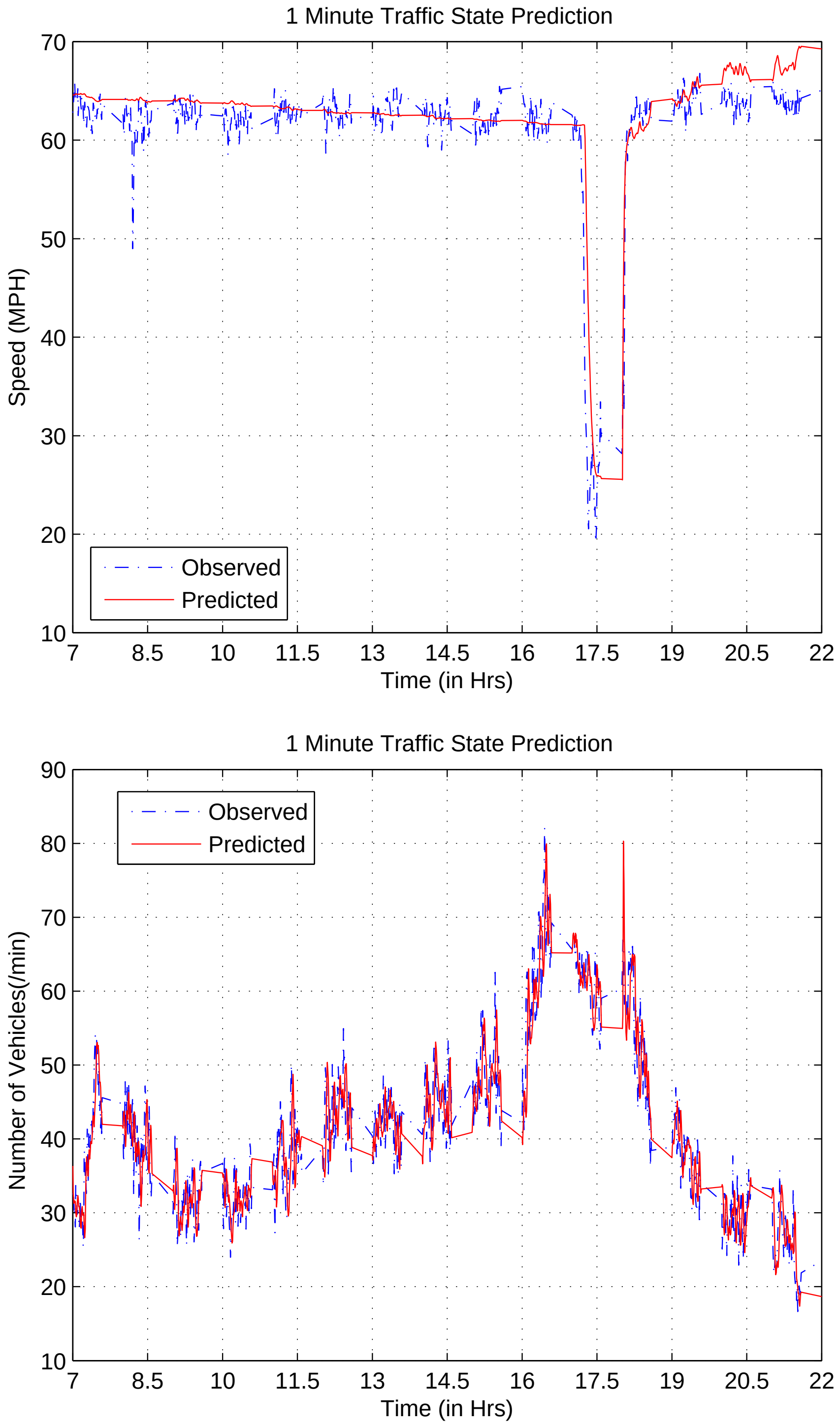

Fig. 18. 1 minute prediction of state of traffic, Section 1, July 15, 2004

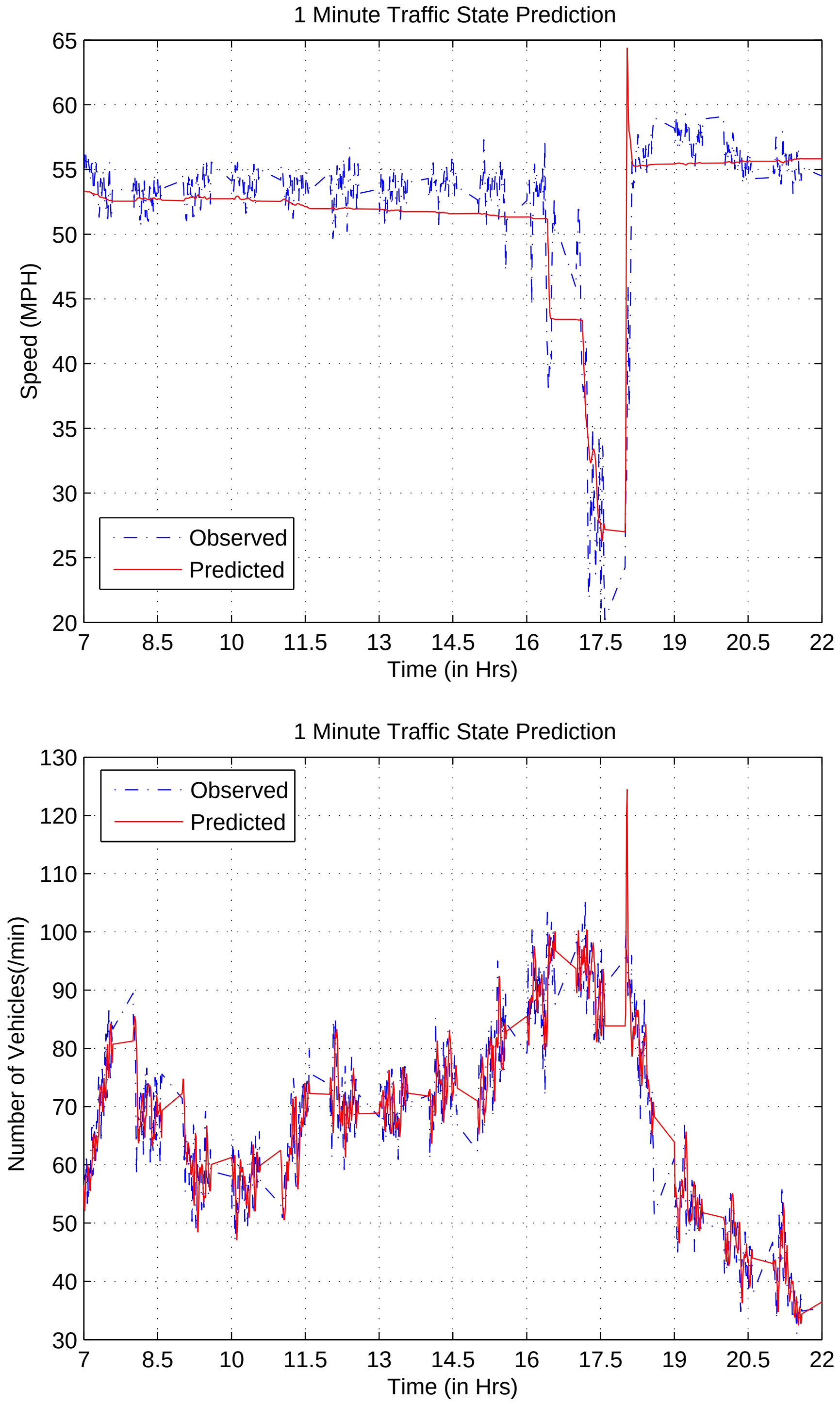

Fig. 19. 1 minute prediction of state of traffic, Section 2, June 14, 2004

C. Corroboration of proposed model: multiple sections in tandem

In this section we describe the non-continuum model corroboration for multiple sections in tandem. The corroboration procedure again consists of two basic steps. The first step is concerned with the calibration of the traffic flow model, i.e., estimating the model parameters. The second step is involved with the prediction of the states of traffic based on the calibrated model parameters. As an example we consider a case study of corroborating the non-continuum model for multiple linked sections with traffic data from sections 1 and 2 as described earlier in section B of this chapter.

1. Model calibration

Model calibration is again concerned with identifying the structure of vehicle following and then estimating the relevant traffic flow model parameters.

a. Setting up linked sections

For the model calibration purposes, we first have to set up linked section on the US 183 freeway. To that end, we again consider *section 1* and *section 2*. The indexing of the sections has been done while being consistent with the nomenclature assumed earlier. Thus *section 1* is the downstream section and *section 2* is the upstream one.

For model calibration purposes, we only need to re-calibrate the parameters for the upstream section (*section 2*). This is because, the only input that the downstream section takes from the upstream section (in the sense of the non-continuum traffic flow model) is the number of vehicles that enter the immediate downstream section under consideration. In the upstream section, on the other hand, the evolution of the traffic state depends on the state of the traffic in the downstream section. This dependence was captured in the aggregate following distance evolution equation in the traffic flow model. We discuss more about this one way traffic disturbance propagation in the subsequent section.

We can now construct a state space system which describes the traffic flow dynamics, for model calibration purposes, in the upstream section. The basic calibration procedure remains the same but there is an additional parameter which needs to be identified before the vehicle following parameters can be estimated. For the upstream section, we have the following the state space system (in representation consistent with quation 4.2 and the formulation consistent with Equation 4.12).

$$\mathbf{x}_k = \begin{bmatrix} N_2(k) & N^{ex}_{2,sum}(k) & \bar{\Delta}_2(k) \end{bmatrix}^T \tag{4.15}$$

$$\mathbf{f}_k = \begin{bmatrix} N_2(k) + h\big[\dot{N}^{en}_2(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)\big] \\ N^{ex}_{2,sum}(k) + h\big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] \\ \bar{\Delta}_2(k) - h\big\{\big[\frac{(L_{car}+\bar{\Delta}_2(k))^2}{L_{s,2}}\big]\big[\dot{N}^{en}_2(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)\big] - (\beta_{2,1}\bar{v}_1(k) - \bar{v}_2(k))\big\} \end{bmatrix}$$

The subscript 2 to the state variables in the Equation 4.15 denotes *section 2.* We notice that in this state space representation, we have the *speed-correction* factor ($\beta_{2,1}$) as an extra unidentified parameter, when compared to Equation 4.12. Before proceeding to estimate the relevant parameters in the vehicle following structure, we need to estimate the value of the speed-correction factor.

Since $\beta_{2,1}$ is a constant, we employ extended Kalman filtering to perform adaptive system identification purposes [61]. For adaptive identification purposes, there is a need to modify the state space system in Equation 4.12.

We augment the state vector by introducing $\beta_{2,1}$ as an extra state. We consider

$\beta_{2,1}$ as a random variable; that is we consider

$$\beta_{2,1}(k+1) = \beta_{2,1}(k) + \xi_k^{\beta_{2,1}} \tag{4.16}$$

where $\beta_{2,1}(k)$ is the value of $\beta_{2,1}$ at the k^{th} time instant and $E(\xi_k^{\beta_{2,1}}) = 0$. Thus we have the following state space representation of the traffic dynamics, for parameter identification purposes, in the section under current consideration (in representation consistent with Equation 4.2).

$$\mathbf{x}_k = \begin{bmatrix} N_2(k) & N_{2,sum}^{ex}(k) & \bar{\Delta}_2(k) & \beta_{2,1}(k) \end{bmatrix}^T$$

$$\mathbf{f}_k = \begin{bmatrix} N_2(k) + h\big[\dot{N}_2^{en}(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)\big] \\ N_{2,sum}^{ex}(k) + h\big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] \\ \bar{\Delta}_2(k) - h\big\{\big[\frac{(L_{car}+\bar{\Delta}_2(k))^2}{L_{s,2}}\big]\big[\dot{N}_2^{en}(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)\big] - (\beta_{2,1}(k)\bar{v}_1(k) - \bar{v}_2(k))\big\} \\ \beta_{2,1}(k) \end{bmatrix}$$

$$H_k = \begin{bmatrix} h & 0 & 0 & 0 \\ 0 & h & 0 & 0 \\ 0 & 0 & h & 0 \\ 0 & 0 & 0 & h \end{bmatrix} \tag{4.17}$$

$$\underline{\xi}_k = \big[\xi_k^{N_2} \quad 0 \quad \xi_k^{\bar{\Delta}_2} \quad \xi_k^{\beta_{2,1}}\big]^T$$

The cumulative count of the number of vehicles exiting the section at each time step is again utilized as the system measurement. Thus we have the following:

$$\mathbf{v}_k = [0 \quad 1 \quad 0 \quad 0]\mathbf{x}_k + \eta_k$$

where $\mathbf{v}_k$ is the the output vector.

After estimating the value of $\beta_{2,1}$, we follow the same procedure as was followed for the corroboration of a single stand alone section. That is we first estimate the parameters associated with the vehicle following in the two traffic flow regimes (*Regime 1* and *Regime 2*) and then finally utilize the estimated parameters to predict the state of traffic in real-time. The prediction of traffic state for multiple linked sections require more elaboration and we discuss it next.

2. Predicting the traffic state in real-time

Estimated time constants (h_w, α and p), associated with different traffic flow regimes, and speed-correction factor ($\beta_{i,i-1}$) are used to predict the traffic state in real-time. We again use extended Kalman filtering, but the state vector is now composed of states from all the inter-linked sections. Thus for the case study under current consideration, the traffic states from both *section 1* and *section 2* are included in the state space description of the traffic dynamics. For example, for *Regime 1*, the state space representation of the traffic flow model (in representation consistent with Equation 4.2) will be:

$$\mathbf{x}_k = \begin{bmatrix} N_1(k) & N^{ex}_{1,sum}(k) & \bar{\Delta}_1(k) & \bar{v}_1(k) & N_2(k) & N^{ex}_{2,sum}(k) & \bar{\Delta}_2(k) & \bar{v}_2(k) \end{bmatrix}^T$$

$$\mathbf{f}_k = \begin{bmatrix} N_1(k) + h\big[\dot{N}_1^{en}(k) - \big[\frac{\bar{v}_1(k)N_1(k)}{L_{s,1}}\big] + \dot{\tilde{n}}_1(k)\big] \\ N_{1,sum}^{ex}(k) + h\big[\frac{\bar{v}_1(k)N_1(k)}{L_{s,1}}\big] \\ \bar{\Delta}_1(k) - h\big\{[\frac{(L_{car}+\bar{\Delta}_1(k))^2}{L_{s,1}}][\dot{N}_1^{en}(k) - \big[\frac{\bar{v}_1(k)N_1(k)}{L_{s,1}}\big] + \dot{\tilde{n}}_1(k)]\big\} \\ \bar{v}_1(k) + \frac{h}{h_{w,1}}\big\{[\frac{(L_{car}+\bar{\Delta}_1(k))^2}{L_{s,1}}][\dot{N}_1^{en}(k) - \big[\frac{\bar{v}_1(k)N_1(k)}{L_{s,1}}\big] + \dot{\tilde{n}}_1(k)] \\ +\lambda_{1,1}(\bar{\Delta}_1(k) - h_{w,1}\bar{v}_1(k))\big\} \\ N_2(k) + h\big[\dot{N}_2^{en}(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)\big] \\ N_{2,sum}^{ex}(k) + h\big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] \\ \bar{\Delta}_2(k) - h\big\{[\frac{(L_{car}+\bar{\Delta}_2(k))^2}{L_{s,2}}][\dot{N}_2^{en}(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)] - (\beta_{2,1}\bar{v}_1(k) - \bar{v}_2(k))\big\} \\ \bar{v}_2(k) + \frac{h}{h_{w,2}}\big\{[\frac{(L_{car}+\bar{\Delta}_2(k))^2}{L_{s,2}}][\dot{N}_2^{en}(k) - \big[\frac{\bar{v}_2(k)N_2(k)}{L_{s,2}}\big] + \dot{\tilde{n}}_2(k)] \\ -(\beta_{2,1}\bar{v}_1(k) - \bar{v}_2(k)) + \lambda_{1,2}(\bar{\Delta}_2(k) - h_{w,2}\bar{v}_2(k))\big\} \end{bmatrix}$$

$$H_k = \begin{bmatrix} h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & h & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & h & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & h & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & h & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & h & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & h & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & h \end{bmatrix} \tag{4.18}$$

$$\underline{\xi}_k = \begin{bmatrix} \xi_k^{N_1} & 0 & \xi_k^{\bar{\Delta}_1} & \xi_k^{\bar{v}_1} & \xi_k^{N_2} & 0 & \xi_k^{\bar{\Delta}_2} & \xi_k^{\bar{v}_2} \end{bmatrix}$$

$$\mathbf{v}_k = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}\mathbf{x}_k + [\eta_{1,k} \quad \eta_{2,k}]^T$$

In Equation 4.18 $\lambda_{1,1}$ (and $h_{w,1}$) and $\lambda_{1,2}$ (and $h_{w,2}$) denote the time constants associated with *Regime 1* for section 1 and section 2 respectively. Also, the subscript

"1" and "2" associated with the state variables denote the corresponding state variables for section 1 and section 2 respectively. The inputs to this system are: the net inflow of vehicles into section 1 ($\dot{N}_1^{en} + \dot{\tilde{n}}_1$), the net inflow of vehicles into section 2 ($\dot{N}_2^{en} + \dot{\tilde{n}}_2$) and the observed occupancy percentage values for both section 1 and section 2. The observed percentage occupancy values are again used in determining the traffic regime transitions.

3. Results

In this subsection we present some results for the non-continuum traffic flow model corroboration for two sections in tandem. It should be noted, that the results, for both parameter estimation and traffic state prediction in real-time, for the downstream section (*section 1*) are the same as when it was treated as a single stand alone section. This is because in traffic flow, there is only unidirectional disturbance propagation and that is only in the upstream direction.

Table VI shows the estimated values of speed-correction factor ($\beta_{2,1}$) for four different weeks. It is interesting to note that the values of $\beta_{2,1}$, for almost all of the days, is very close to the ratio of the aggregate traffic speeds in the upstream and the downstream section. It is reasonable to expect this behavior since the speed-correction factor in a sense smoothes the sharp change in speed that happens at the boundary of the downstream and upstream section due to the "instantaneous" change in the speed of the representative vehicle when it crosses into the downstream section from the upstream section. If both the downstream and upstream sections had equal aggregate traffic speeds, then the value of $\beta_{2,1}$ is expected to be equal to one as there will be a "smooth" crossover of the representative vehicle into the downstream section.

Table VI. Estimated values of speed-correction factor ($\beta_{2,1}$)

$\beta_{2,1}$					
Monday		Tuesday		Wednesday	
Date	$\beta_{2,1}$	*Date*	$\beta_{2,1}$	*Date*	$\beta_{2,1}$
03-01-04	0.753	03-02-04	0.721	03-03-04	0.697
03-08-04	0.845	03-09-04	0.787	03-10-04	0.689
04-05-04	0.832	04-06-04	0.806	04-07-04	0.724
05-03-04	0.865	05-04-04	0.741	05-05-04	0.784

Thursday		Friday	
Date	$\beta_{2,1}$	*Date*	$\beta_{2,1}$
03-04-04	0.667	03-05-04	0.835
03-11-04	0.712	03-12-04	0.647
04-08-04	0.772	04-09-04	0.796
05-06-04	0.700	05-07-04	0.712

Table VII and Table VIII show the estimated parameters for *Regime 1* and *Regime 2a* for the same four weeks. It is worth noting that these estimated parameters are not much different, than when they were calculated by considering section 2 as a stand alone section. This is again to be expected as in the framework of the non-continuum traffic flow modeling, any section can be considered as a stand alone section by assigning it as the extreme downstream section. In Fig. 20 we plot the one minute traffic state prediction for June 14, 2004 on section 2. On Fig. 21 we plot the one minute traffic speed predictions for June 14, 2004 on section 2 as obtained by multiple

Table VII. Estimated parameters for *Regime 1* - Section 2, two sections in tandem

Regime 1								
Monday			Tuesday			Wednesday		
Date	λ_1	h_w	*Date*	λ_1	h_w	*Date*	λ_1	h_w
03-01-04	1.4E-7	0.172	03-02-04	9.0E-4	0.287	03-03-04	8.9E-4	0.237
03-08-04	1.5E-7	0.214	03-09-04	4.9E-4	0.265	03-10-04	9.3E-4	0.221
04-05-04	4.1E-7	0.184	04-06-04	5.6E-4	0.343	04-07-04	8.2E-4	0.325
05-03-04	4.7E-7	0.291	05-04-04	6.5E-4	0.305	05-05-04	5.9E-4	0.219

Thursday			Friday		
Date	λ_1	h_w	*Date*	λ_1	h_w
03-04-04	4.7E-4	0.269	03-05-04	1.9E-7	0.199
03-11-04	6.4E-4	0.226	03-12-04	7.9E-6	0.311
04-08-04	7.3E-4	0.312	04-09-04	2.7E-7	0.286
05-06-04	8.1E-4	0.185	05-07-04	5.2E-8	0.201

linked sections modeling and when section 2 was treated as a stand alone section. On the same figure, we also plot the actual observed traffic speed in section 2 on June 14, 2004. From the plot, we can note that the multiple linked section modeling approach provides slightly better predictions. This again is not unreasonable, since in multiple linked sections modeling approach, we account for the upstream traveling traffic disturbance propagation in real-time.

We are now in a position to explain the mechanism for traffic congestion as explained by the non-continuum traffic flow modeling approach. We describe the same in the following section.

Table VIII. Estimated parameters for *Regime 2a* - Section 2, two sections in tandem

Regime 2a							
Monday				Tuesday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-01-04	4.1E-12	4	0.219	03-02-04	8.9E-7	3	0.561
03-08-04	2.8E-12	4	0.143	03-09-04	1.7E-8	3	0.352
04-05-04	7.0E-12	4	0.211	04-06-04	3.0E-8	3	0.667
05-03-04	5.1E-12	4	0.268	05-04-04	2.6E-8	3	0.431
Wednesday				Thursday			
Date	α	p	λ_2	*Date*	α	p	λ_2
03-03-04	5.1E-12	4	0.134	03-04-04	3.0E-12	4	0.305
03-10-04	5.3E-12	4	0.119	03-11-04	2.6E-12	4	0.331
04-07-04	4.1E-12	4	0.166	04-08-04	2.5E-12	4	0.394
05-05-04	5.2E-12	4	0.246	05-06-04	3.7E-12	4	0.279

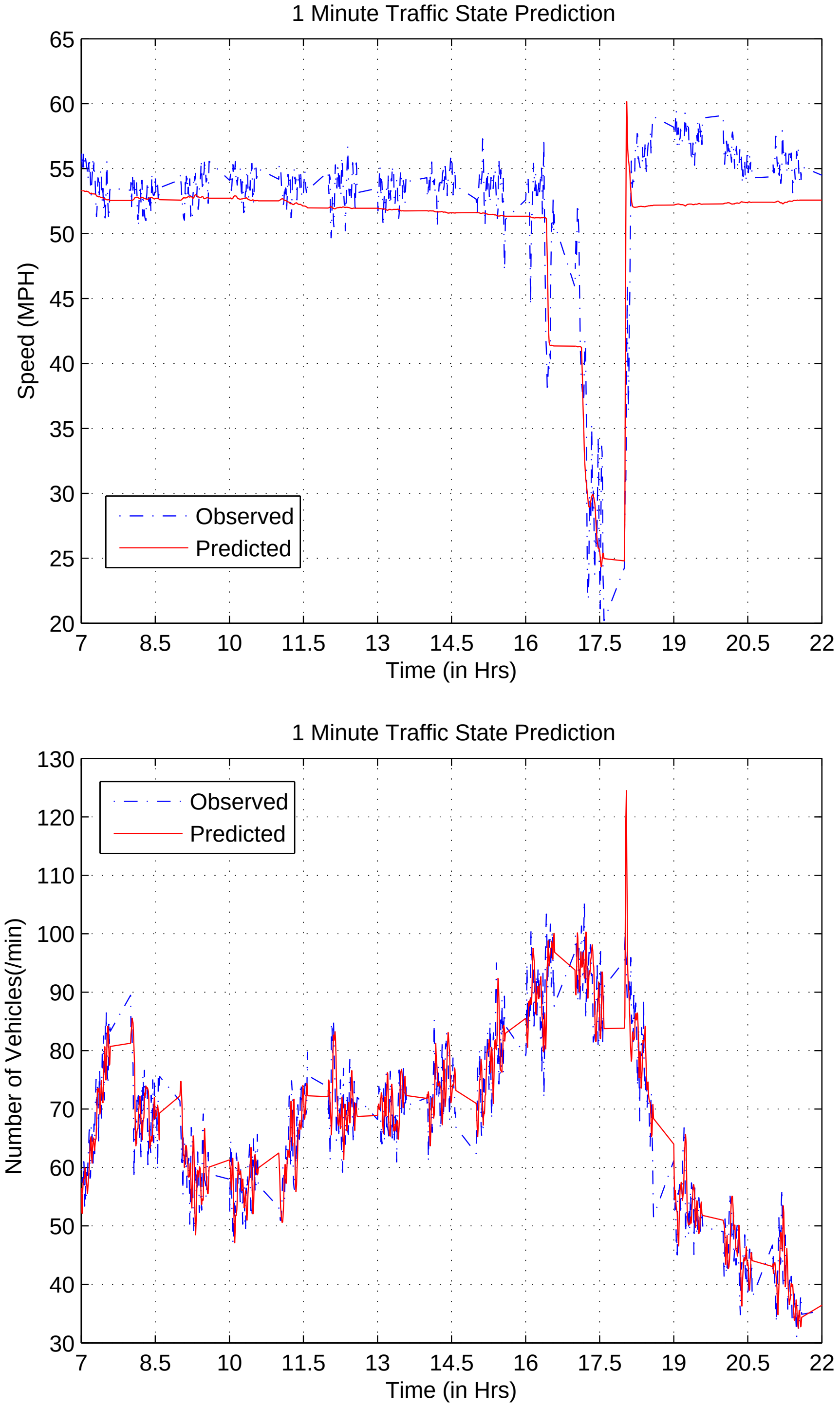

Fig. 20. 1 minute prediction of state of traffic, Section 2, two sections in tandem, June 14, 2004

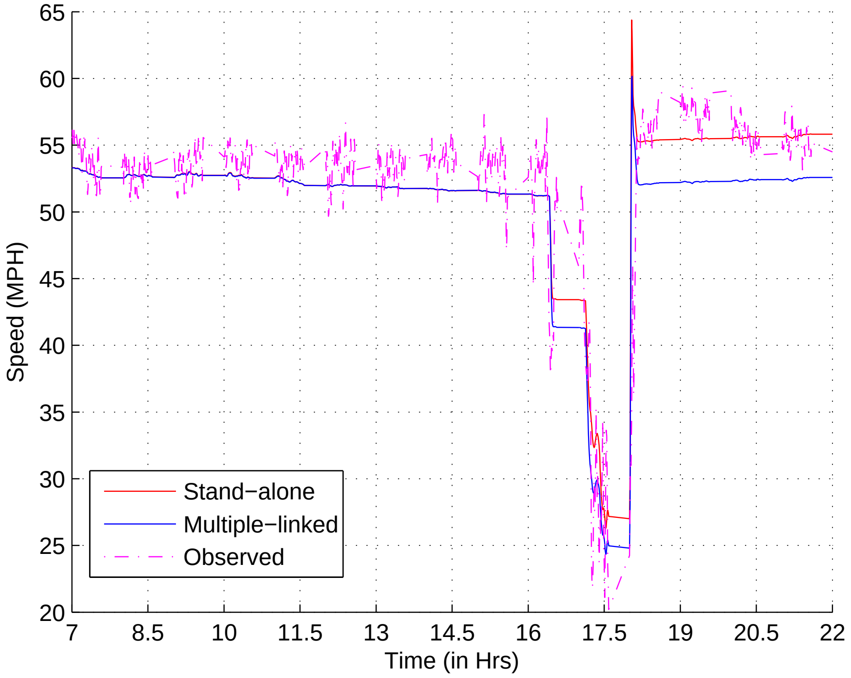

Fig. 21. Comparing "stand alone" and "multiple linked" 1 minute traffic speed prediction, Section 2, June 14, 2004

D. Explanation of the mechanism for traffic congestion based on a non-continuum traffic flow model

The mechanism for traffic congestion depicted by the non-continuum traffic flow model is as follows:

1. The rate of change of aggregate following distance in the i^{th} section changes with the introduction (or exiting) of vehicles from the entrance (or exit) ramps.
2. As a result of the rate of change of the aggregate following distance, the rate of change of aggregate traffic speed of vehicles in the i^{th} section changes.
3. As a consequence of the change of traffic speed of the vehicles in the i^{th} section, the rate of change of aggregate following distance in the immediately upstream ("$i + 1$") section changes (due to the speed differential component in the aggregate following distance equation). Thus, the aggregate speed and the aggregate following distance in the upstream section changes.

An important point to note in the non-continuum model is that, there is only one-way disturbance propagation (as compared to a disturbance propagation in both directions as predicted by the continuum traffic flow models). To understand this better, consider a decrease in aggregate traffic speed in the i^{th} section (this may be due to influx of vehicles from the entrance ramps). Now if the decrease in the aggregate traffic speed in the i^{th} section is such that the rate of change of number of vehicles exiting the i^{th} section does not change, then the downstream section does not feel the effect of the disturbance (the influx of vehicles from the entrance ramps) created in the i^{th} section.

In a freeway section, a decrease in speed can also result from a vehicle breaking down. Let us now consider the following scenario: there is a breakdown in an upstream

$(i+1)$ section and its immediate downstream (i) section is in congested regime (*Regime 2*). In this case, the number of vehicles entering the downstream (i) section will decrease and eventually the number of vehicles exiting the i^{th} section will be more than the number of vehicles entering. This will lead to a negative rate of change of number of vehicles with respect to time $(\dot{N}_i)$ which in turn will lead to an increase in the aggregate following distance in the i^{th} section. In other words, we will have a traffic regime transition from *Regime 2* to *Regime 1.* Thus, the disturbance is never propagated into the downstream sections. Infact, it can only be propagated into the upstream section(s) as explained by the congestion mechanism earlier. This overcomes a major limitation (decrease in speed in the current section implies a decrease in speed in the downstream section as well!) which is predicted by the continuum models. This prediction of the model and its agreement with what is observed in the flow of traffic cannot be over-emphasized. We reiterate the classical continuum model incorrectly predicts that there is a decrease in speed of the vehicles in the downstream section due to a decrease in the speed of vehicles in the upstream section.

Another important point to be made with respect to the non-continuum traffic flow model is that it does not involve ad-hoc spatial discretization. Also, the speed of the propagation of the disturbance to the traffic is not faster than that of the traffic - changes in aggregate following distance and speed in any section (and the adjoining sections as well) always lags the disturbance. It is for all these reasons that the non-continuum methodology for traffic flow modeling is a significant welcome departure from the existing macroscopic traffic flow modeling approaches.

E. Summary

We have presented the methodology for corroboration of the non-continuum traffic model with traffic data collected from the loop detectors deployed on the freeways. Structures for vehicle following behavior were identified and they were used in real-time estimation and prediction of traffic state. We have also explained the mechanism of traffic congestion as explained by the non-continuum traffic flow modeling approach. We discuss autoregressive integrated moving average (ARIMA) modeling of traffic time series data in Appendix B. In the same appendix we also present results regarding ARIMA modeling on the selected locations on US 183 north bound, i. e. section 1 and section 2. We then compare the non-continuum corroboration results with those obtained with ARIMA modeling. In the next chapter we extend the developed corroboration methodology for making short term traffic state forecasts.

CHAPTER V

SHORT-TERM TRAFFIC STATE FORECASTING AND TRIP TRAVEL TIME ESTIMATION AND PREDICTION

A. Introduction

Current practices in traffic management and control strategies are dominated by the emerging use of intelligent transportation systems (ITS) methodologies. The overall objective of ITS systems is to increase the operational efficiency and capacity of the transportation network by utilizing the advancements in technological and telecommunication systems. A continuous flow of information about the traffic state (example, aggregate traffic speed and number of vehicles crossing a fixed point on the freeway in a time interval) and its evolution over time is the basic idea behind creating an efficient ITS environment [63]. This traffic information is dynamic and has to be anticipative in nature. In other words, the traffic information (which is provided to the users of the transportation network) to be useful and applicable, must be updated in real-time and should provide projections/forecasts on the traffic conditions expected at some future instant in time. Short-term traffic state forecasting can then be defined as the process of estimating the anticipated traffic conditions at a future time, given historical and short-term feedback traffic information.

In the context of short-term traffic state forecasting there are two important considerations. They are forecasting horizon and forecasting time step respectively. Forecasting horizon denotes the extent of time ahead to which the forecast refers to. The forecasting step represents the time interval upon which the forecasts are made and indicates the frequency of predictions in the forecasting horizon. For example, an algorithm might predict the aggregate traffic speed ten minutes ahead in five minute intervals. The basic implications of the forecasting horizon is quite intuitive. For example, greater the forecasting horizon, the less accurate the forecast gets. Ishak et. al [64] studied the relationship between the accuracy of forecasts and the forecasting horizon and concluded that the prediction accuracy degrades with an increase in the forecasting horizon. Vythoulkas suggested that the quality of information extracted by users degrades if the forecasting horizon is less than ten minutes [65].

A review of the various short-term traffic state forecasting approaches can be found in the article by Vlahogianni et al. [66].

B. Short-term traffic forecasting by non-continuum traffic model

We now discuss how the non-continuum traffic model can be utilized in making short-term traffic state forecasts. During the corroboration of the traffic flow model, we observed that the model was able to make good one minute forecasts of the aggregate traffic speed and the number of vehicles crossing a fixed location on the freeway per minute. As already discussed, for all practical ITS applications we need much larger forecasting horizon.

By the application of extended Kalman filtering algorithm to the real-time traffic data we are able to obtain one step (one minute) traffic state predictions. An intuitive way of obtaining *multi-step* predictions, is to apply the state update vector to the filtered state at each time instant m times if m step prediction is required[1]. Though, this is a viable way for obtaining multi-step predictions it does not produce reliable forecasts. The main drawback is that these multi-step predictions will be based on the state of the traffic at the current time interval. For example, multi-step predictions made this way will fail to capture the onset of congestion (traffic regime switch from

[1] This is a very well known result in Kalman filtering theory. For more details please refer [61].

Regime 1 to *Regime 2*) that might happen ten minutes into the future, based on the occupancy levels information at the current time instant.

1. Proposed approach: Use of historical traffic data

In this subsection we propose a methodology for providing short-term traffic state forecasts by utilizing historical data archives. Setting up of the historical traffic data database from twenty five weeks of data was earlier presented in Chapter III of the dissertation. The database consists of five comma separated value (csv) files, each containing the full day's inductive loop detector traffic data for one working day (Monday through Friday).

The proposed approach is to obtain a weighted traffic data, with weights allocated to both the current, and the "future" traffic data obtained from the historical traffic database. In Fig. 22 we show the schematic of the method of using historical and real-time traffic information for the purposes of short-term traffic state forecasting. For the purposes of this study, we aim to make ten and fifteen minute traffic state forecasts. To that end, we allocate exponential weights to historical and real-time traffic information as follows.

Let k denote the current time instant. To make a traffic state forecast fifteen minutes into the future, that is at the time instant $k + 15$ (for example) we utilize the traffic data as follows:

1. traffic information at $(k + 15)^{th}$ time instant from the historical database,
2. traffic information at $(k + 10)^{th}$ time instant from the historical database, and
3. current traffic information at the k^{th} time instant.

We now allocate exponential weights to the above mentioned three *instances* of

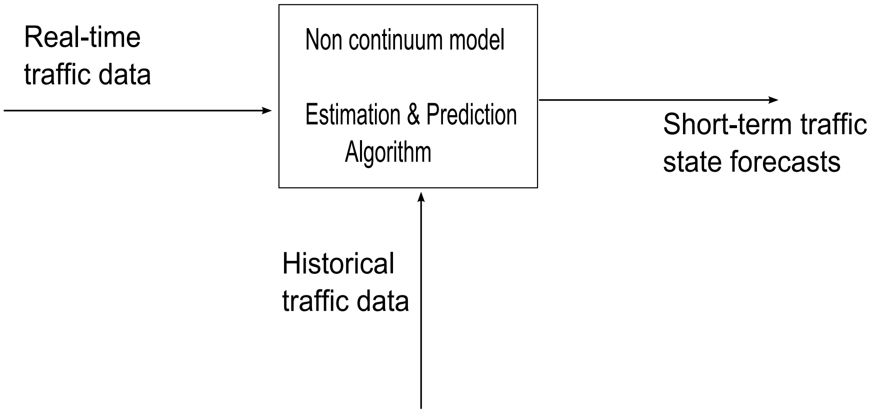
Real-time traffic data
Non continuum model
Estimation & Prediction Algorithm
Short-term traffic state forecasts
Historical traffic data

Fig. 22. Schematic of utilization of historical and real-time traffic for short-term forecastingdata

traffic data to obtain a “future instant” information. This is done by assigning the weights as follows:

$$W_i = \frac{e^{-i^2/3}}{\sum_{j=1}^{3} e^{-j^2/3}}, \quad i = 1, 2, 3 \tag{5.1}$$

In Equation 5.1, i refers to the three different time instants as described earlier and W_i is the corresponding associated weight. Thus we associate maximum weight to the fifteen minute future traffic information obtained from the historical database and the least to the current traffic information[2]. We can also assign similar weights to $(k + 10)^{th}$, $(k + 5)^{th}$ (both from the historical database) and k^{th} (current real-time data) time instants for traffic state forecasting on a ten minute horizon. It is important to note that in this procedure of utilizing the historical traffic information we are also using the feedback from the current traffic information.

With regard to this obtained “future instant” traffic information, it is important to note that on any given day if the traffic conditions are *drastically* different from the

[2] Specific values of these weights are: 0.5050, 0.3619 and 0.1331 respectively. It should be pointed out that this allocation weights has not been optimized. A more detailed study may have to be carried out find out the optimal weights.

historical averages, then the approach mentioned above may not produce desirable results. It should also be pointed out that the traffic behavior on the selected stretch of US 183 in Austin was observed to be very repeatable across all the working days[3].

If from the historical traffic data it is observed that on some days the traffic conditions are drastically different from the historical averages then the following alternate approach to obtain "future instant" traffic information may be followed. In this alternate approach we propose to use the difference between future traffic information obtained from the historical database and the current instant traffic information. To obtain a value for the traffic data, vehicle count for example, we multiply this difference with the ratio of the current traffic data obtained in real-time and the current instant data obtained from the historical database. We then associate the same weights as discussed earlier. Thus this approach puts more emphasis on the current real-time information. It should also be noted that if one had a historical database with traffic information about incidents and special events, then that information can also be incorporated to obtain predictions which have a capability of being applicable in most situations.

Once we have obtained the "future instant" traffic information then we can utilize the extended Kalman filtering algorithm to recursively estimate the traffic states. The input data which is required to estimate and predict the traffic state in real-time is now this *future instant* traffic data. We can now have three extended Kalman filters running in parallel. The first one *runs* on the real-time traffic data and provides one minute traffic state forecasts. The second and the third filters run on fifteen and ten minute future instant traffic information providing fifteen and ten minute traffic state forecasts respectively. The forecasting time step for all the three filters is one minute

[3] Traffic data has been discussed in detail in Chapter III.

(traffic data is available at one minute intervals).

2. Results

In Fig. 23 we show the ten minute traffic state forecasts for *section 1* on June 14, 2004. On the same figure we plot the actual observed traffic states. Fig. 24 shows the fifteen minute traffic state prediction for *section 1* on June 14, 2004. In Fig. 25 we plot the ten minute and fifteen minute speed predictions along with the observed traffic speed on June 14, 2004. We have zoomed into the plot area where the traffic regime switching takes place.

It can be observed that both ten minute and fifteen minute predictions match very well with the observed traffic states. For example, the fifteen minute speed predictions is able to capture the onset of congestion about eight minutes in advance.

C. Short term travel time prediction

Accurate Traffic surveillance systems are a core element in a transportation system. With the advent of intelligent of ITS, accurate estimation and prediction of section trip travel times over freeway networks has become an important issue. Accurate prediction of trip travel times is a very critical issue for many ITS applications, such as in-vehicle route guidance systems (RGS) and advanced traffic management systems (ATMS). With the development of the advanced traveler information systems (ATIS), short-term trip travel time prediction is becoming increasingly important [67], [68]. For example, for RGS applications, travel time information enables the generation of a path which would take the shortest amount of time (or an alternative path) connecting the current location and destinations, besides also suggesting directions dynamically in case of congestions or incidents.

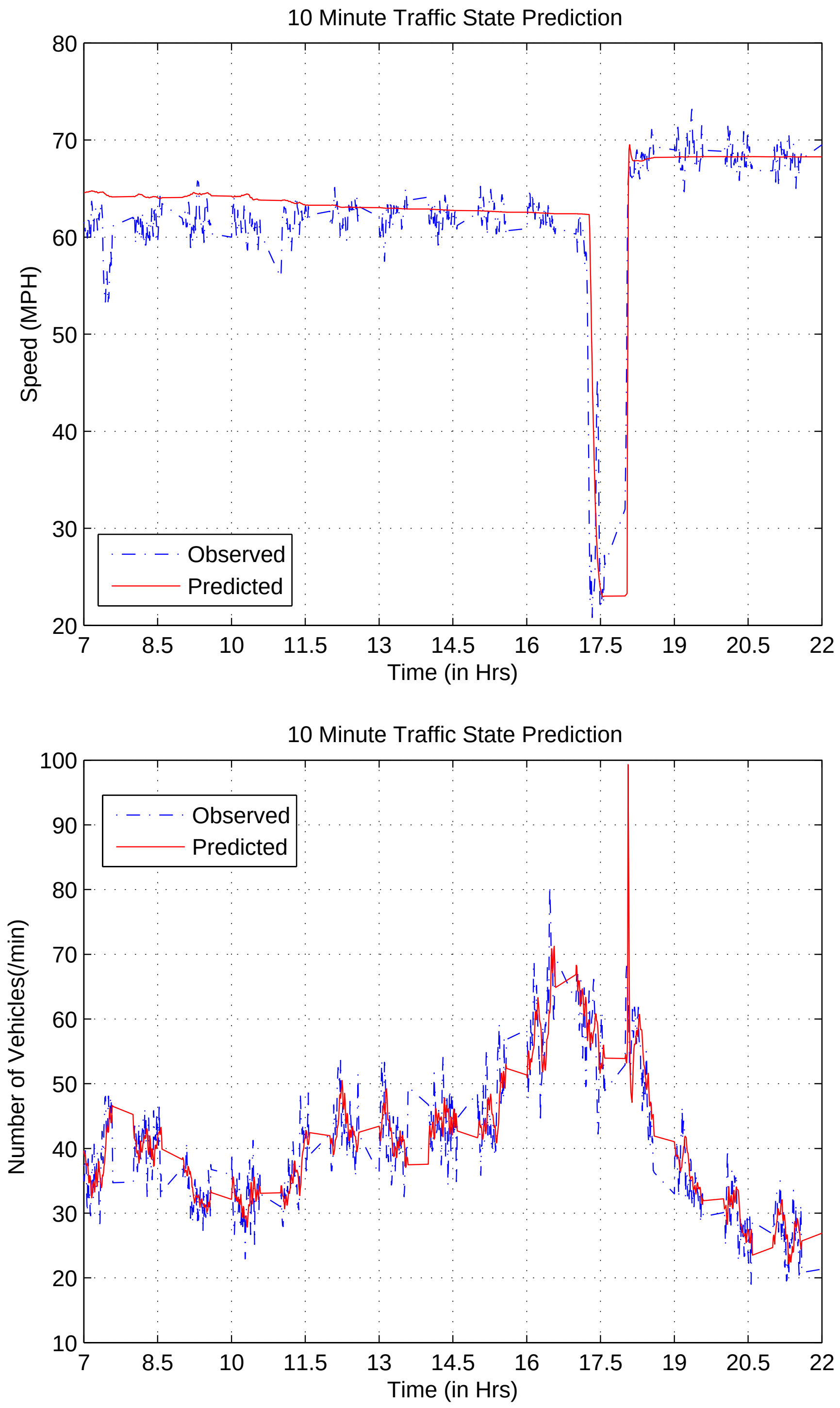

Fig. 23. 10 minute prediction of state of traffic, Section 1, June 14, 2004

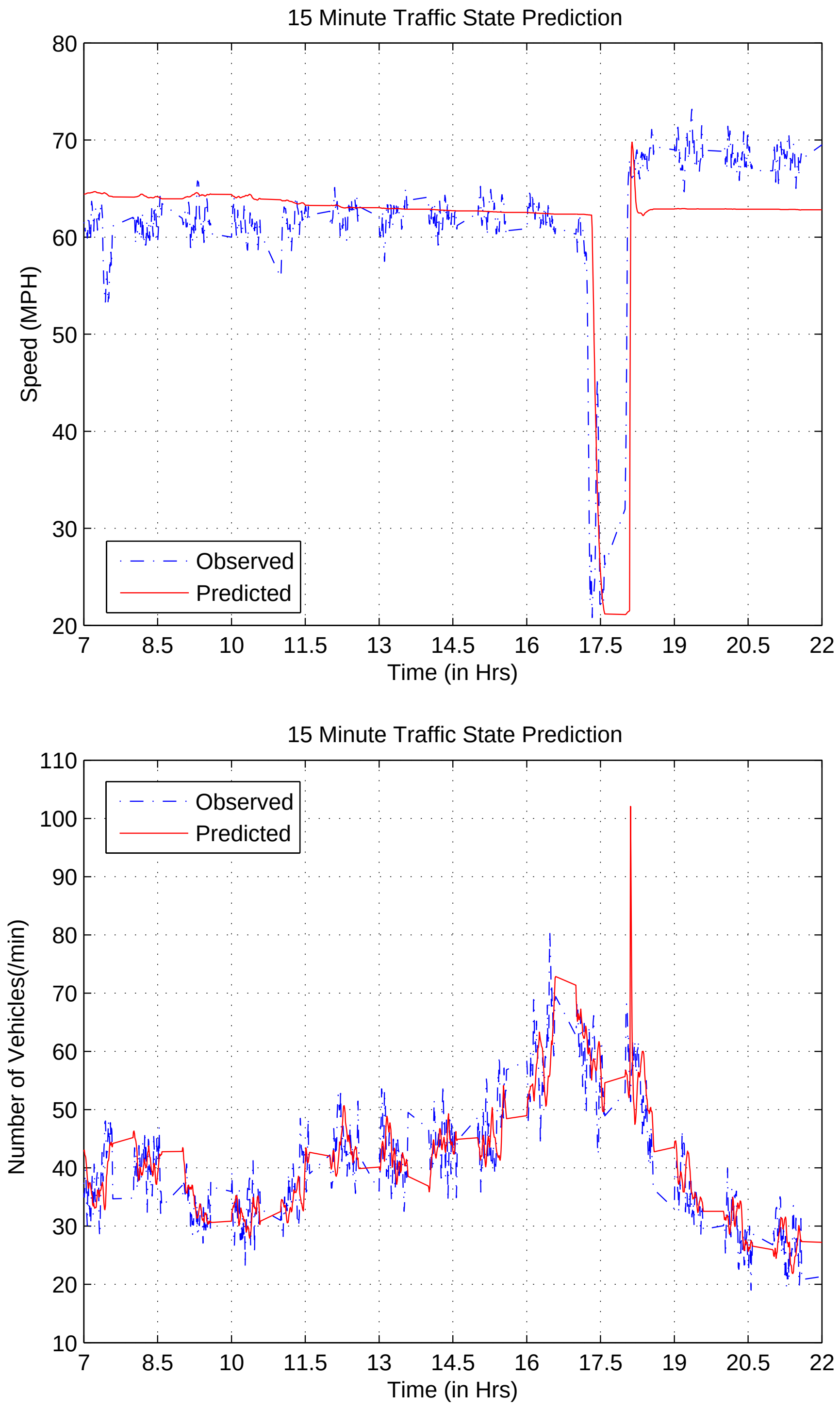

Fig. 24. 15 minute prediction of state of traffic, Section 1, June 14, 2004

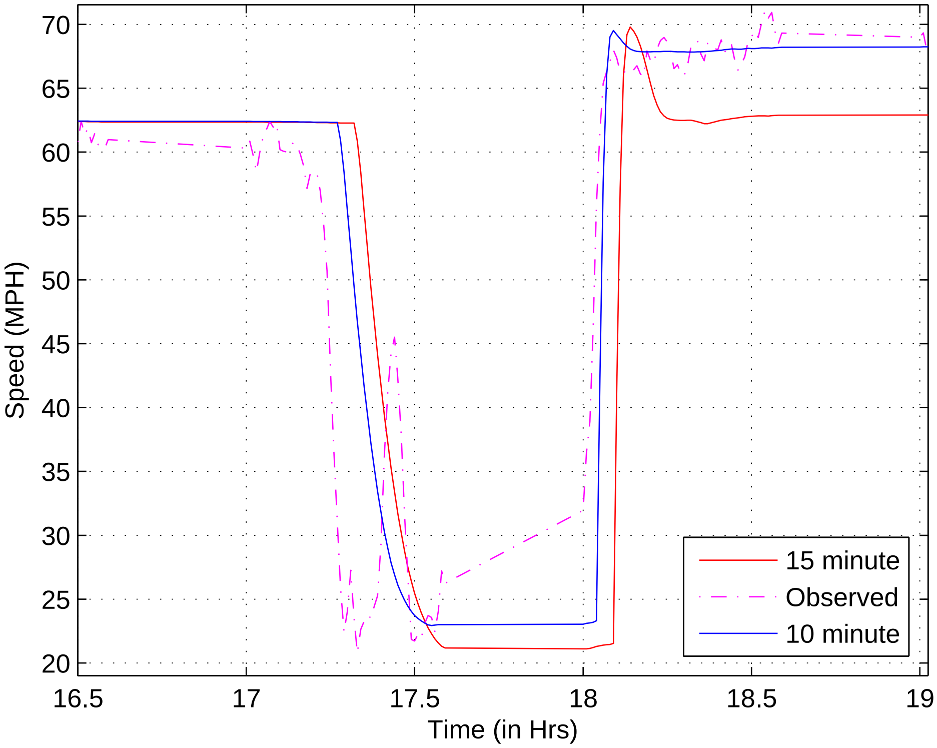

Fig. 25. Comparing 15 minute and 10 minute speed predictions, Section 1, June 14, 2004

For trip travel time estimation, many advanced techniques have been applied which involve the use of advanced sensor, telecommunications and image processing technologies [69]. However, the existing infrastructure in most urban contexts currently only supports point detection of traffic, mostly by inductive loop detectors.

There has been much research contributing to the field of trip travel time prediction which try to utilize the loop detector traffic data. In the context of prediction methodologies, various time series modeling approaches ([70] - [74]), artificial neural network models ([75], [76]), utilization of Kalman filtering and probe vehicles ([77], [78]) have been used. There have also been some efforts in trying to use input and exit flow and traffic speed measurements from loop detectors. Some of these studies, also try to exploit the relationship between traffic density and speed, in a continuum modeling context ([79] - [85]).

Owing to our motivation for developing a traffic model which can utilize the readily available loop detector data, for this study we have focussed our efforts on developing a speed-based section trip travel time estimation and prediction algorithm[4].

1. Speed-based section travel time estimation

Let us first define what we mean by travel time and section travel time. Travel time can be simply defined as the time spent by a vehicle in travelling from one point to another. For a freeway section, the section travel time can be defined as the amount of time it takes for a vehicle to reach the downstream end of the section after it starts from the upstream end. Trip travel times are represented by discretizing the temporal and spatial dimensions. Trip travel time can be calculated if one is able to estimate the average traffic speed. However, it is important to distinguish between space-mean

[4] We have interchangeably used section travel time and trip travel time.

and time-mean average speeds in the present context. The space-mean speed reflects the average speed over a spatial section of freeway, while the time-mean speed reflects the average speed of the traffic stream passing a specific fixed point on the freeway. The speed data observed by the loop detectors is the time-mean traffic speed.

The basic idea behind speed-based trip travel time estimation models is to construct a "virtual" trajectory of a hypothetical vehicle as it moves along the freeway. Most of the current speed-based models which attempt to estimate the travel time over a network of links [5] utilize the following basic relationship[6]:

$$t(k) = \frac{2l}{v_a(k) + v_b(k)} \tag{5.2}$$

where $t(k)$ is the trip travel time estimated at time instant k. l, $v_a(k)$ and $v_b(k)$ denote the length of the link and speed measurements at the upstream and downstream detector location at time instant k respectively. This is a very coarse estimation of trip travel time because the implicit assumption that the speed of the hypothetical vehicle does not change during its travel along the link.

D. Short-term trip travel time estimation and prediction using the non-continuum modeling approach

In the context of the non-continuum modeling methodology, the idea of the representative limiting vehicle provides us with an abstraction of the traffic at all instants in time. Thus, the trip travel time is implicitly contained within the definition of the

[5] These developed models uses a slightly different notion of a freeway section (called a "link"). A link is a freeway segment between two consecutive loop detector stations. As a result, when this definition is compared with our definition of a section, it becomes clear that two consecutive links constitute one section in the context of non-continuum traffic modeling.

[6] Let a and b denote the upstream and downstream ends of a link.

abstract representative vehicle. We can define the section travel time by the following relationship:

$$TT_{i,t_k} = \frac{L_{s,i}}{\bar{w}_i(t_k)} \tag{5.3}$$

where TT_{i,t_k} denotes the section travel time for section i estimated at time t_k. $L_{s,i}$ and $\bar{w}_i(t_k)$ denote the section length and the space-mean speed of the traffic in the i^{th} section and at time t_k respectively. It is important to note that $\bar{w}_i$ is never zero since every vehicle traverses the section in a finite amount of time, though it may have zero speed for some time period in between. Looking at Equation 5.3 it might feel intuitive to replace $\bar{w}_i$ by the speed of the representative vehicle, $\bar{v}_i$ at each time instant t_k. However during the time periods of congestion (and this is the most relevant case), the speed of the representative vehicle ($\bar{v}_i$) may temporarily be zero (or very small) leading to arbitrarily high errors when used in place of the space-mean speed ($\bar{w}_i$) in Equation 5.3.

1. Proposed approach

An approach which can avoid this difficulty and leads to much better results is as follows. Consider a virtual test vehicle which is started periodically from the upstream end of the section and is moved along the section with the speed of the representative vehicle at each time interval until the it exits the section. Thus in our non-continuum modeling context, a virtual test vehicle is started at one minute intervals and it moves with a speed $\bar{v}_i(t_k)$ for the time period $[t_k, t_{k+1})$ before shifting its speed to $\bar{v}_i(t_{k+1})$ at $t = t_{k+1}$. We can then have the following recursive scheme until the virtual test vehicle exits the section. If $pos_{tv,j}(t_k) \leq L_{s,i}$ and $L_{s,i} - pos_{tv,j}(t_k) \geq T.\bar{v}_i(t_k)$, then we have:

$$\begin{aligned} pos_{tv,j}(t_{k+1}) &= pos_{tv,j}(t_k) + T.\bar{v}_i(t_k) \\ TT_{i,t_k} &= TT_{i,t_k} + T \end{aligned} \tag{5.4}$$

If $\quad pos_{tv,j}(t_k) \leq L_{s,i} \quad$ and $\quad L_{s,i} - pos_{tv,j}(t_k) \ngeq T.\bar{v}_i(t_k)$, then

$$\begin{aligned} pos_{tv,j}(t_{k+1}) &= L_{s,i} \\ TT_{i,t_k} &= TT_{i,t_k} + \frac{L_{s,i} - pos_{tv,j}(t_k)}{\bar{v}_i(t_k)} \end{aligned} \tag{5.5}$$

In equations 5.4 and 5.5 above $pos_{tv,j}(t_k)$ and T are the position of the virtual test vehicle at time instant t_k and the time step (one minute in our context) respectively.

It should be noted that the trip travel time is available only once the virtual test vehicle has exited the section under consideration. This methodology extends itself easily for estimation of travel time over multiple linked sections. When a vehicle crosses over from the upstream to the downstream section, the speed of the virtual test vehicle changes from that in the upstream to the aggregate traffic speed in the downstream section. Thus the vehicle then traverses the downstream section for a part of the time step (that is one minute interval), if any, with the current speed of the downstream section and then updates its speed at the start of the new time step.

2. Trip travel time prediction

For ITS applications, prediction of trip travel times over a ten to fifteen minute forecasting horizon is required. In our proposed speed-based approach for estimating trip travel times, it is very convenient to extend the trip travel time estimation algorithm to produce trip travel time forecasts.

To be able to make a 15 minute forecasting horizon (for example) trip travel time prediction, we simply have to replace the current aggregate traffic speed ($\bar{v}_i(t_k)$) with the fifteen minute predicted aggregate traffic speed. The forecasting time step of one minute for short-term traffic state predictions (as discussed earlier) helps in making the predictions for trip travel time with the same time step.

3. Results

We now present some results regarding the trip travel time estimation and prediction. Efforts were made to collect empirical measurements for trip travel time on US 183 in Austin. Fig. 26 shows the estimated trip travel time for traversing from the upstream end of section 2 to the downstream extreme of section 1 (or multiple linked sections) for Monday, June 14, 2004. On the same figure we have plotted the actual trip travel time observed on Monday, January 29, 2007. Fig. 27 and Fig. 28 show the ten minute and fifteen minute trip travel time predictions for June 14, 2004 plotted along with the observed travel time on January 29, 2007.

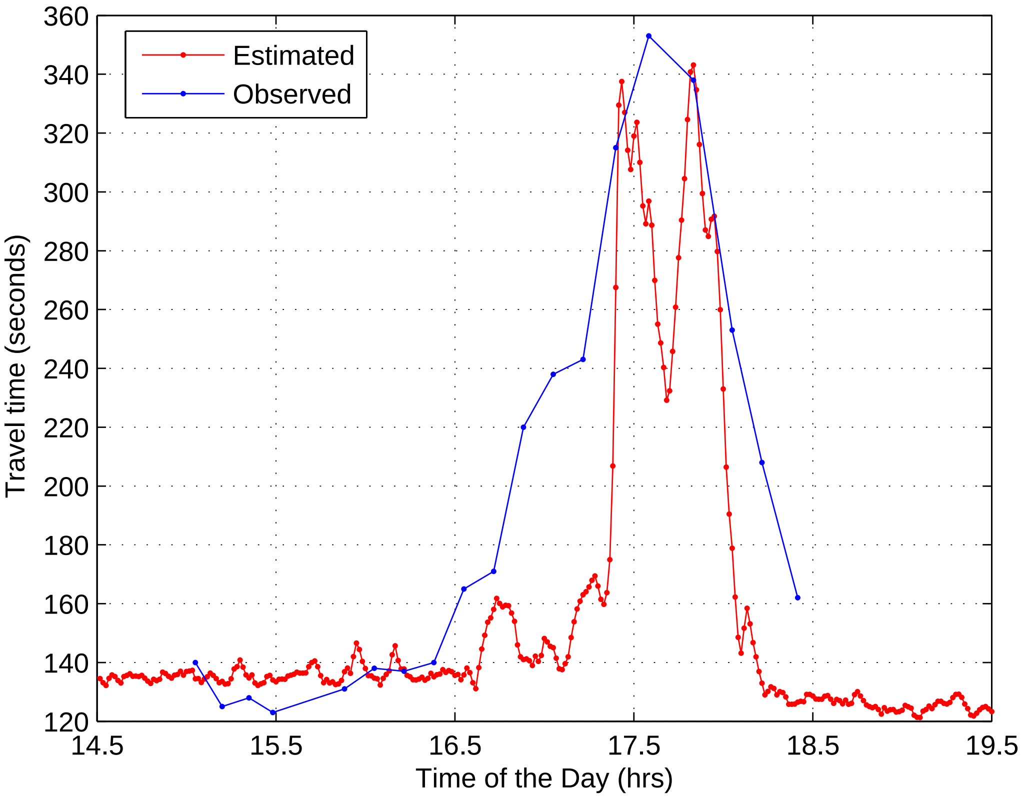

Fig. 26. Estimated trip travel time on June 14, 2004: Multiple linked sections

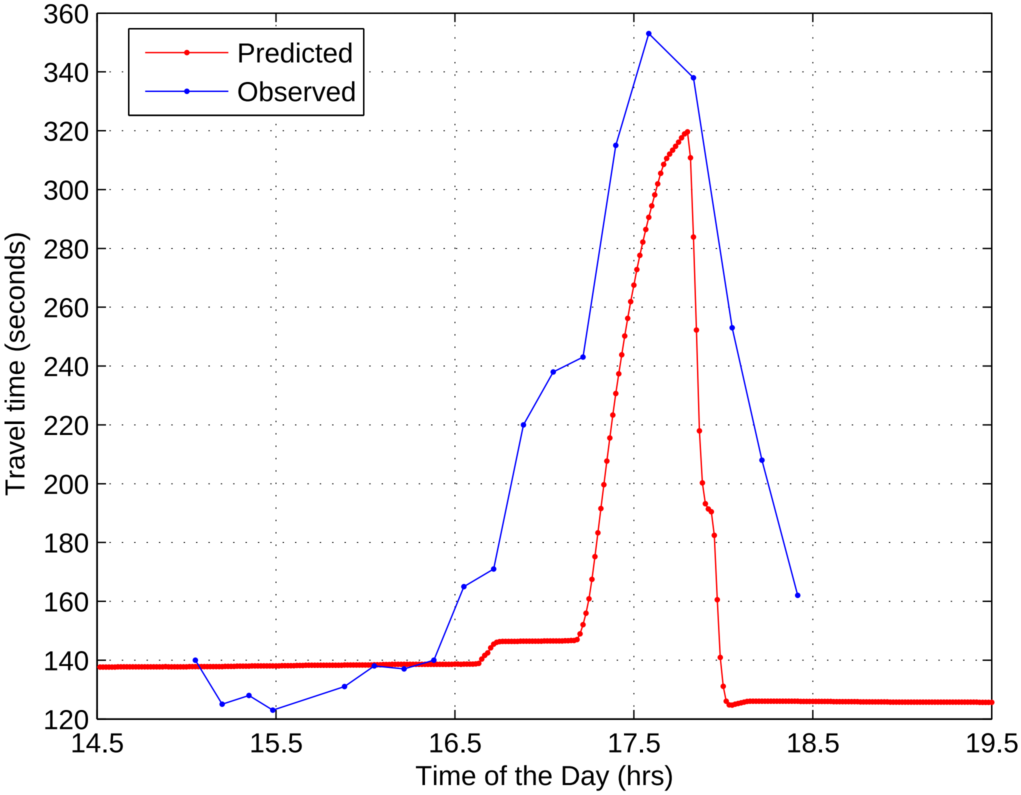

Fig. 27. 10 minute prediction of trip travel time on June 14, 2004: Multiple linked sections

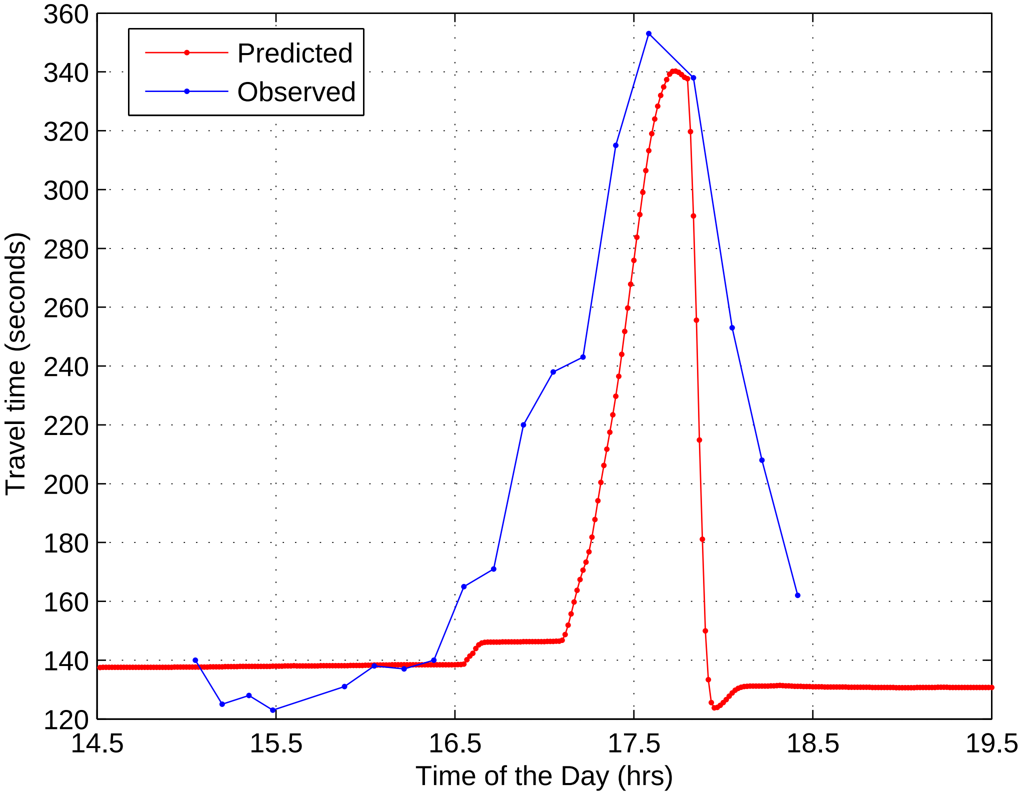

Fig. 28. 15 minute prediction of trip travel time on June 14, 2004: Multiple linked sections

The results plotted in figures 26, 27 and 28 have to be interpreted as follows. The travel time (plotted on the $y-axis$) corresponding to a time instant (plotted on the $x-axis$) is the time taken for the virtual test vehicle to reach the downstream end (end point) of the trip after it starts at the time instant under consideration from the upstream end (starting point) of the trip. For fifteen minute (or ten minute) prediction of trip travel time the travel time plotted is the travel time estimated to be taken by a virtual test vehicle which starts fifteen minutes (or ten minutes) into the future. For example, in Fig. 28, the reported travel time at 5:45 PM is about 340 seconds. This travel time is the estimated travel time for the virtual test vehicle which starts at 5:45 PM when the current time is 5:30 PM. Since, the forecasting time step is one minute and the virtual test vehicle takes 340 seconds to traverse the length of the freeway under consideration, we can report this predicted travel time at 5:36 PM. The travel time may be reported in the following format:

"`Current time: 5:36 PM. The predicted travel time from Lamar Blvd. to Mopac at 5:45 PM is 5 minutes and 40 seconds.`"

These predicted travel times may be updated in real-time at internet web-sites or may be displayed at the variable message sign boards along the freeway. Depending on the scope of the ITS application, the update rates can be kept fixed or variable. For example, they can be updated in real-time every ten minutes for relatively free flow conditions and five minutes for congested traffic. Following remarks are very important when validating the proposed methodology results with the actual measured travel time:

- The loop detector clocks and the clock used during trip time measurement are not synchronized, thus creating either a lead or a lag in the measured time when compared to the time when data was collected and processed on June 14, 2004.

- The exact locations of the trap loop detectors on the freeway and the exit/entrance ramps is not known. It has been suggested that the locations of the loop detectors on the freeway are supposed to be within 500 *feet* of the nearest cross street. Thus it becomes difficult to determine the exact locations of the start point (upstream end of section 2) and the end point (downstream end of section 1) of the trip under consideration.
- The travel time readings were measured by driving a vehicle repeatedly between the entrance ramp at Lamar Blvd. (upstream end of section 2) and the exit ramp at Mopac (downstream end of section 1). This resulted in fewer measurements. To that end, the measured travel time data has been interpolated linearly between two consecutive measurements.

Despite the above mentioned difficulties, we can observe from figures 26, 27 and 28 that the proposed methodology is able to make estimations and predictions which compare well with the actual measured travel time. These results again help reinstate the point that the aggregate behavior of traffic is pretty well defined and is observed to be repeatable[7].

E. Summary

We presented the methodology for short term traffic state forecasting using the non-continuum traffic flow model. An algorithm for trip travel time prediction was also presented. It can be seen that the results for traffic state and trip travel time prediction are very encouraging.

[7] Model calibrated with traffic data from years' 2003 and 2004 predicts travel time for a day in 2007 quite well.

CHAPTER VI

CONCLUSION

In this chapter we present an assessment of the contributions of this research to the state-of-the-art of the macroscopic traffic flow modeling. We then provide a brief discussion of some practical issues that arise in the implementation of this research for *on-field* ITS applications. We conclude the chapter with suggestions for further research.

A. Contributions to state-of-the-art

This research represents an advancement of the state-of-the-art of the macroscopic traffic flow models and their applications in real-time estimation and prediction of the traffic conditions. The non-continuum traffic model presented here is a significant departure from the current macroscopic traffic flow modeling approaches. Specifically,

- The non-continuum approach overcomes the philosophical difficulties in modeling traffic as a continuum. Aggregate vehicle following behavior is integrated in the macroscopic traffic flow model. This integration is of utmost importance since it is essential to understand how the control at the microscopic level affects the macroscopic dynamics. Moreover, with the non-continuum approach we directly obtain a spatially discrete model which is suited for estimation and control purposes.

- In the presented methodology, estimation and prediction of the traffic state can be conducted simultaneously.

- The presented model has been subjected to empirical testing based on real dual loop detector data with very encouraging results. With this elaborate process of corroboration, the model is ready for implementation within a prototype freeway management and/or traveller management systems.

B. Application issues

In this section, we discuss the various issues related to the implementation of the non-continuum model within an ITS framework.

1. Data issues

Sensor failures are not uncommon occurrences. Various bad or missing data issues as applicable to US 183 traffic data were discussed in detail in Chapter III. It is important that the traffic state estimation and prediction methodology is robust with respect to sensor failures[1]. The methodology presented in this thesis can accommodate missing observations arising as a result of sensor failures. The easiest way to do this would be to assign a historical average value[2]. The estimation and prediction methodology would otherwise remain the same.

2. Computational issues

Given the size of the most real-life freeway networks, computational considerations assume an important role in practical ITS applications. In the present research effort we corroborated the non-continuum model for a freeway length of 2.2 miles. For our case study, there were no computational issues - all estimation and prediction

[1] In this dissertation, for corroboration purposes we identified two sections with good and meaningful data for both the years 2003 and 2004. What we imply here by sensor failures is that there is an odd occurrence of failure and the sensor can be corrected in due course in time.

[2] This can be done by setting up a historical database in the same way as has been done in this research effort.

calculations are able to complete within the one minute forecasting time step.

For real-life ITS applications in real-time, the freeway stretch has to be much longer. This will lead to many linked sections and thus more traffic states to estimate and predict. To that end, it is very important to consider modifications to the extended Kalman filtering algorithm which can reduce the computational load. The most *time-consuming* operation in the Kalman filtering process is the computation of the Kalman gain matrices (see Equation 4.9). In particular "Sequential" and "Square-Root" algorithms have been proposed which attempt to avoid a direct computation of the inverse of the matrix

$$\Big[\big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\big]P_{k,k-1}\big[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\big]^T + R_k\Big]^{-1}$$

in Equation 4.9[3].

C. Further research directions

We now discuss some possible directions for further research.

1. Freeway management applications

Further research may be carried out to extend the applications of the non-continuum traffic flow model.

1. Further research may be carried out in designing and testing various ramp metering strategies. One way to do this is by utilizing synthetic traffic data in conjunction with traffic data collected from the sensors deployed on the freeways for simulation purposes.

[3] The reader is referred to [61] for more details.

2. More empirical testing of the model may be carried out. This can be done via calibrating the model for much longer stretch of the freeway. This will also involve in developing variants of the Kalman filtering algorithm as discussed before. This exercise can also lead to possibly identifying more structures of vehicle following depending on the location of the selected freeway stretch.
3. Research work can be carried out to increase the forecasting horizon to be more than fifteen minutes. There is also some research scope in identifying optimal weights to be assigned for historical and current traffic information. This can possibly lead to much better predictions of traffic state and trip travel time.

2. Possible model refinements

One possible model refinement is to include more than one representative vehicle in any section. This may be accomplished by developing an aggregate model for each individual lane. It is possible to consider each lane separately since lane changes and passing can be automatically taken into account via the aggregation procedure as they will lead to changes in the following distance. This research endeavor can possibly lead to a much more realistic description of the macroscopic traffic flow and its corroboration with traffic data can be a challenging and a rewarding effort.

D. Summary

A comprehensive methodology for the corroboration of the non-continuum traffic flow model with traffic data collected from loop detectors deployed on the US 183 freeway in Austin, Texas has been presented in this thesis. The model is able to estimate and predict the traffic state in real-time. An algorithm for predicting trip travel times has also been developed. The results thus far are very encouraging and indicate that the developed traffic state estimation and prediction methodology is robust enough to work with loop detector traffic data in real-time and is ready to be for a prototype implementation.

REFERENCES

[1] J. P. Kinzer, “Application of the theory of probability to problems of highway traffic,” in *Proc. Institute of Traffic Engineers*, vol. 5, 1934, pp. 118-124.

[2] B. D. Greenshields, “A study of traffic capacity,” in *Proc. Highway Research Board*, Washington, D.C., vol. 14, 1935, pp. 448-477.

[3] F. A. Haight, *Mathematical Theories of Traffic Flow.* New York: Academic Press, 1963.

[4] H. Greenberg, “An analysis of traffic flow,” *Operations Research*, vol. 7, no. 1, pp. 79-85, 1959.

[5] Federal Highway Administration, “Focus on Congestion Relief,” http://www.fhwa.dot.gov/congestion, Accessed January 2005.

[6] D. Schrank and T. Lomax, “The 2005 Urban Mobility Report,” http://tti.tamu.edu/documents/mobility_report_2005.pdf, Accessed April 2005.

[7] C. Chen, Z. F. Jia and P. Varaiya, “Causes and cures of highway congestion,” *IEEE Control Systems Magazine*, pp. 26-32, 2001.

[8] S. Darbha and K. R. Rajagopal, “Limit of a collection of dynamical systems: An application to modeling the flow of traffic,” *Mathematical Models and Methods in Applied Sciences*, vol. 12, no. 10, pp. 1381-1399, 2002.

[9] D. Helbing, “Traffic and related self-driven many-particle systems,” *Reviews of Modern Physics*, vol. 73, no. 4, pp. 1067-1141, 2001.

[10] M. Papageorgiou, “Some remarks on macroscopic traffic flow modeling,” *Transportation Research Part A*, vol. 32, no. 5, pp. 323-329, 1998.

[11] C. F. Daganzo, “Estimation of gap acceptance parameters within and across the population from direct roadside observation,” *Transportation Research Part B*, vol. 15, pp. 1-15, 1981.

[12] P. G. Gipps, “A model for the structure of lane changing decisions,” *Transportation Research Part B*, vol. 20, no. 5, pp. 403-414, 1986.

[13] L. A. Pipes, “An operational analysis of traffic dynamics,” *Journal of Applied Physics*, vol. 24, pp. 274-281, 1953.

[14] T. W. Forbes, H. J. Zagorski and E. L. Deterline, “Measurement of driver reactions to tunnel conditions,” in *Proc. Highway Research Board*, vol. 37, 1958, pp. 345-357.

[15] W. Leutzbach, *An Introduction to the Theory of Traffic Flow.* Berlin: Springer-Verlag, 1988.

[16] M. Jepsen, “On the speed-flow relationship in road traffic: A model of driver behavior,” in *Proc. Third International Symposium on Highway Capacity*, Copenhagen, Denmark, 1998, pp. 297-319.

[17] P. G. Gipps, “A behavioral car-following model for computer simulation,” *Transportation Research Part B*, vol. 15, pp. 105-111, 1981.

[18] T. Dijker, P. H.L. Bovy and R. G. M. M. Vermijs, “Car following under congested conditions: empirical findings,” *Transportation Research Record*, 1644, pp. 20-28, 1998.

[19] R. E. Chandler, R. Herman, and E. W. Montroll, "Traffic dynamics: Studies in car following," *Operations Research*, vol. 6, pp. 165-184, 1958.

[20] D. C. Gazis, R. Herman, and R. W. Rothery, "Nonlinear follow the leader models for traffic flow," *Operations Research*, vol. 9, pp. 545-567, 1961.

[21] K. Nagel, "Particle hopping models and traffic flow theory," *Phys. Rev. E*, 53, pp. 4655-4672, 1996.

[22] K. Nagel, "From particle hopping models to traffic flow theory," *Transportation Research Record*, 1644, pp. 1-9, 1998.

[23] D. C. Gazis, R. Herman, and R. B. Potts, "Car following theory of steady state traffic flow," *Operations Research*, vol. 7, pp. 499-505, 1959.

[24] L. A. Pipes, "Car following models and the fundamental diagram of road traffic," *Transportation Research*, vol. 1, pp. 21-29, 1967.

[25] G. F. Newell, "A simplified car-following theory: a lower order model," *Transportation Research Part B*, vol. 36, pp. 195-205, 2002.

[26] B. S. Kerner and S. L. Klenov, "A microscopic model for phase transitions in traffic flow," *Journal of Physics A: Mathematical and General*, vol. 35, pp. L31-L43, 2002.

[27] G. F. Newell, "A theory of traffic flow in tunnels," in *Theory of Traffic Flow*, R. Herman, Ed. Amsterdam: Elsevier, 1961, pp. 193-206.

[28] Q. Yang and H. N. Koutsopoulos, "A microscopic traffic simulator for evaluation of dynamic traffic management systems," *Transportation Research Part C*, vol. 4, no. 3, pp. 113-129, 1996.

[29] A. Halati, L. Henry, and S. Walker, “CORSIM-corridor traffic simulation model,” in *Proc. Traffic Congestion and Traffic Safety Conference*, Chicago, IL, 1998, pp. 570-576.

[30] M. J. Lighthill and G. B. Whitham , “On kinematic waves II: A theory of traffic flow on long crowded roads,” in *Proc. Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 229, Issue 1178, 1955, pp. 317-345.

[31] P. I. Richards, “Shock waves on the highway,” *Operations Research*, pp. 42-51, 1956.

[32] H. J. Payne, “Models for freeway traffic and control,” in *Simulation Council Proceedings Series: Mathematical Models of Public Systems*, La Jolla, CA, 1971, (Edited by G. A. Bekey), vol. 1, pp. 51-61.

[33] G. Liu, A. S. Lyrintzis and P. G. Michalopoulos, “Improved high-order model for freeway traffic flow,” *Transportation Research Record*, 1644, pp. 57-46, 1998.

[34] G. F. Newell, “Mathematical models for freely-flowing highway traffic,” *Journal of the Operations Research Society of America*, vol. 3, no. 2, pp. 176-186, 1955.

[35] I. Prigogine and C. F. Andrews, “A Boltzman-like approach for traffic flow,” *Operations Research*, vol. 8, pp. 789-797, 1960.

[36] I. Prigogine, “A Boltzman-like approach to the statistical theory of traffic flow,” in *Theory of Traffic Flow*, R. Herman, Ed. Amsterdam: Elsevier, 1961, pp. 158-164.

[37] I. Prigogine and R. Herman, *Kinetic theory of vehicular traffic*. New York: American Elsevier, 1971.

[38] N. Bellomo and M. Pulvirenti, *Modelling in Applied Sciences: A Kinetic Theory Approach*. Boston: Birkhauser, 2000.

[39] C. F. Daganzo, "The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B*, vol. 28, no. 4, pp. 269-287, 1994.

[40] G. F. Newell, "A simplified theory of kinematic waves in highway traffic, part I: general theory," *Transportation Research Part B*, vol. 27, no. 4, pp. 281-287, 1993.

[41] S. P. Hoogerdoorn and P. H. L. Bovy, "State-of-the-art of vehicular traffic flow modelling," *Proc. Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 215, Part 4, pp. 283-304, 2001.

[42] C. A. Truesdell and K. R. Rajagopal, *An Introduction to the Mechanics of Fluids.* New York: Springer, 2000.

[43] G. F. Newell, "A simplified car following theory: A lower order model," *Transportation Research Part B*, vol. 36, pp. 196-205, 2002.

[44] D. Swaroop and J. K. Hedrick, "String stability of interconnected systems," *IEEE Trans. Automatic Control*, vol. 41, no. 3, pp. 349-357, 1996.

[45] W. Maes, "Traffic data collection system for the Belgian motorway network - measures of effectiveness aspects," in *Proc. International Symposium on Traffic Control Systems*, Berkeley, CA, 1979, vol. 2D - Analysis and Evaluation, pp. 45-73.

[46] J. Treiterer and J. A. Myers, "The hysteresis phenomena in traffic flow," in *Proc. Sixth Symposium on Transportation and Traffic Theory*, D. J. Buckley Ed. New

York: Elsevier, 1974, pp. 13-38.

[47] S. Darbha and K. R. Rajagopal, “Intelligent cruise control systems and traffic flow stability,” California PATH Research Report, UCB-ITS-PRR-98-36, 1998.

[48] A. Angel, M. Hickman, P. Mirchandani, and D. Chandani, “Methods of analyzing traffic imagery collected from aerial platforms,” *IEEE Trans. on Intelligent Transportation Systems*, vol. 4, no. 2, pp. 99-107, 2003.

[49] J. Kell, I. Fullerton, and M. Mills, *Traffic Detector Handbook*. Second Edition, Federal Highway Administration, Washington, D.C. 1990.

[50] Texas Transportation Institute, and Texas Department of Transportation, “20030522, 2003 Austin Freeway Operations Traffic Data Archive,” Texas Transportation Institute, College Station, TX.

[51] Texas Transportation Institute, and Texas Department of Transportation, “20030522, 2004 Austin Freeway Operations Traffic Data Archive,” Texas Transportation Institute, College Station, TX.

[52] M. Papageorgiou, *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control*. New York: Springer-Verlag, 1983.

[53] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, “Three categories of traffic data: Historical, current, and predictive,” in *Proc. 9th IFAC Symposium on Control in Transportation Systems*, Braunschweig, Germany, 2000, pp. 250-255.

[54] H. M. Zhang, “A mathematical theory of traffic hysteresis,” *Transportation Research Part B*, vol. 33, pp. 1-23, 1999.

[55] H. M. Zhang and T. Kim, "A car following theory for multiphase vehicular traffic flow," *Transportation Research Part B*, vol. 39, pp. 385-399, 2005.

[56] C. F. Daganzo, "A behavioral theory of multilane traffic flow. Part 1: Long homogenous freeway sections," *Transportation Research Part B*, vol. 36, pp. 131-158, 2002.

[57] M. Cremer and M. Papageorgiou, "Parameter identification for a traffic flow model," *Automatica*, vol. 17, no. 6, pp. 837-843, 1981.

[58] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through kalman filtering theory," *Transportation Research Part B*, vol. 18, no. 1, pp. 1-11, 1984.

[59] Y. Wang and M. Papageorgiou, "Real time freeway traffic state estimation based on extended Kalman filter: A general approach," *Transportation Research Part B*, vol. 39, pp. 141-167, 2005.

[60] J. Meier and H. Wehlan, "Section wise modeling of traffic flow and its application in traffic state estimation," in *Proc. IEEE Intelligent Transportation Systems Conference*, Oakland, CA, 2001, pp. 440-445.

[61] C. K. Chui and G. Chen, *Kalman Filtering: With Real Time Applications.* 3^{rd} Ed., Springer Series in Information Sciences, New York: Springer-Verlag, 1999.

[62] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, pp. 308-313, 1965.

[63] H. C. Lieu, "Traffic estimation and prediction system," *TR News, Transportation Research Board*, vol. 208, pp. 3-6, 2000.

[64] S. Ishak and H. Al-Deek, "Performance of short-term time series traffic prediction model," *Journal of Transportation Engineering*, 128(6), pp. 490-498, 2002.

[65] P. C. Vythoulkas, "Alternative approaches to short term traffic forecasting for use in driver information systems," in *Proc. 12th International Symposium on Traffic Flow Theory and Transportation*, Berkeley, CA, 1993, pp. 485-506.

[66] E. I. Vlahogianni, J. C. Golias and M. C. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transport Review*, vol. 24, no. 5, pp. 533-557, 2004.

[67] S. Chien and M. Chen, "Determining the number of probe vehicles for freeway travel time estimation using microscopic simulation," *Transportation Research Record*, 1719, pp. 61-68, 2001.

[68] S. Chien and M. Chen, "Dynamic freeway travel time prediction using probe vehicle data: Link-based vs path-based," *Transportation Research Record*, 1768, pp. 157-161, 2001.

[69] S. M. Turner, "Advanced technologies for travel time data collection," *Transportation Research Record*, 1551, pp. 51-58, 1996.

[70] T. Oda, "An algorithm for prediction of travel time using vehicle sensor data," in *Proc. IEEE 3rd International Conference on Road Traffic Control*, London, UK, 1990, pp. 40-44.

[71] A. Hobeika and C. Kim, "Traffic flow prediction systems based on upstream traffic," in *Proc. IEEE Vehicle Navigation & Information Systems Conference*, Yokohama, Japan, 1994, pp. 345-350.

[72] P. Pant, M. Polycarpou, M. P. Sankaranarayanan, B. Li, X. Hu and A. Hossain, "Travel time prediction system (TIPS) for freeway work zones," in *ASCE Proc. ICTTS'98 Conference on Traffic and Transportation Studies*, Beijing, China, 1998, vol. 98, pp. 20-29.

[73] T. Nakata and J. Takeuchi, "Mining traffic data from probe car system for travel time prediction," in *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 817-822.

[74] J. Yang, "A study of travel time modeling via time series analysis," in *Proc. IEEE Conference on Control Applications*, Toronto, Canada, 2005, pp. 855-860.

[75] D. Park, L. Rilett and G. Han, "Forecasting multiple period freeway link travel times using neural networks with expanded input nodes," in *Proc. ASCE 5th International Conference on Applications of Advanced Technologies in Transportation*, Newport Beach, CA, 1998, pp. 325-332.

[76] L . Rilett and D. Park, "Direct forecasting of freeway corridor travel time times using spectral basis neural networks," *Transportation Research Record*, 1617, pp. 163-170, 1999.

[77] S. I. Chien and C. M. Kuchipudi, "Dynamic travel time prediction with real-time and historic data," *Journal of Transportation Engineering*, no. 6, pp. 608-616, 2003.

[78] J. Yang, "Travel time prediction using the GPS test vehicle and Kalman fitering techniques," in *Proc. American Control Conference*, Portland, Oregon, 2005, pp. 2128-2133.

[79] M. Cremer, "On the calculation of individual travel times by macroscopic models," in *Proc. IEEE Vehicle Navigation & Information Systems Conference*, Seattle, WA, 1995, pp. 187-193.

[80] D. H. Nam and D. R. Drew, "Traffic dynamics: Method for estimating freeway travel time in real time from flow measurements," *Journal of Transportation Engineering*, vol. 122, no. 3, pp. 185-191, 1996.

[81] B. Coifman and M. Cassidy, "Automated travel time measurement using vehicle lengths from loop detector speed traps," California PATH Research Report, UCB-ITS-PRR-2000-12, 2000.

[82] B. Coifman, "Estimating travel times and vehicle trajectories on freeways using dual loop detectors," *Transportation Research Part A*, vol. 36, pp. 351-364, 2002.

[83] J. Oh, R. Jayakrishnan and W. Recker, "Section travel time estimation from point detection data," Center for Traffic Simulation Studies, Paper UCI-ITS-TS-WP-02-15, 2002, http://repositories.cdlib.org/itsirvine/ctss/UCI-ITS-TS-WP-02-15, Accessed June 2004.

[84] X. Zhang and J. A. Rice, "Short term travel time prediction," *Transportation Research Part C*, vol. 11, pp. 187-210, 2003.

[85] J. A. Rice and E. Zwet, "A simple and effective method for predicting travel times on freeways," in *Proc. IEEE Intelligent Transportation Systems Conference*, Oakland, CA, pp. 227-232, 2001.

[86] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering (ASME)*, 82D, pp. 35-45, 1960.

[87] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day, 1976.

[88] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelation in autoregressive integrated moving average time series models," *Journal of American Statistical Association*, vol. 65, pp. 1509-1526, 1970.

[89] T. Y. Der, "Some investigations of forecasting freeway occupancies," M.S. thesis, Illinois Institute of Technology, Chicago, 1977.

[90] M. Eldor, "Demand predictors for computerized freeway control systems," in *Proc. 7th International Symposium on Transportation and Traffic Theory*, Institute of Systems Science Research, Kyoto, Japan, 1977, pp. 341-370.

[91] M. Levin and Y. D. Tsao, "On forecasting freeway occupancies and volumes," *Transportation Research Record*, 773, pp. 47-49, 1980.

[92] M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time series dataa by using Box-Jenkins techniques," *Transportation Research Record*, 722, pp. 1-9, 1982.

[93] S. L. Dhingra, P. P. Mujumdar and R. H. Gajjar, "Applications of time series techniques for forecasting truck traffic attracted by the Bombay Metropolitan Region," *Journal of Adanced Transportation*, vol. 27, no. 3, pp. 227-249, 1993.

[94] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transportation Research Record*, 1678, pp. 179-188, 1999.

APPENDIX A

STATE SPACE MODELING AND KALMAN FILTERING

In this appendix we give a brief overview of State Space Modeling and the Kalman Filter[4]. We first introduce the state space model and the basic Kalman filtering algorithm as applied to linear systems. Finally we introduce the *Extended Kalman Filtering* algorithm as applied for filtering applications of non linear systems.

State Space Model

State-space models are essentially a notational convenience for estimation and control problems. For a linear system, its state space description typically consists of two equations - the *transition*[5] equation and the *measurement*[6] equation.

$$\text{Transition Equation:} \quad \mathbf{x}_{k+1} = A_k \mathbf{x}_k + \Gamma_k \underline{\xi}_k \tag{A.1}$$

$$\text{Measurement Equation:} \quad \mathbf{v}_k = C_k \mathbf{x}_k + \underline{\eta}_k \tag{A.2}$$

where $\mathbf{x}_k$ is the vector which represents the state of the system during the time interval k. $\mathbf{v}_k$ is the vector of measurements made in the time interval k. A_k, Γ_k and C_k are known constant matrices. $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ are respectively, the (unknown) system and measurement noise sequences, with known statistical information such as mean, variance and covariance.

We typically make the following assumptions about the model:

[4] For more extensive coverage of the material, the reader is referred to [61].

[5] Or *system*

[6] Or *observation*

1. $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ are *zero mean*, *independent*[7] and *gaussian* processes with

   $$E(\underline{\xi}_k\underline{\xi}_l^T) = Q_k\delta_{kl} \text{ and } E(\underline{\eta}_k\underline{\eta}_l^T) = R_k\delta_{kl}; \quad \delta_{kl} = 1 \text{ if } k = l \text{ and } 0 \text{ otherwise.}$$

2. The initial state of the system $\mathbf{x}_0$ is independent of $\underline{\xi}_k$ and $\underline{\eta}_k$ in the sense that $E(\mathbf{x}_0\underline{\xi}_k^T) = 0$ and $E(\mathbf{x}_0\underline{\eta}_k^T) = 0$ for all $k = 0, 1, ...$

Given these assumptions, the *filtering* problem is to estimate the quantity $\hat{\mathbf{x}}_{k|k} = E(\mathbf{x}_k|\overline{\mathbf{v}}_k)$[8] where $\overline{\mathbf{v}}_k$ denotes the vector $[\mathbf{v}_0...\mathbf{v}_k]^T$. A *one-step prediction* problem is to estimate the quantity $\hat{\mathbf{x}}_{k|k-1} = E(\mathbf{x}_k|\overline{\mathbf{v}}_{k-1})$. Finally the smoothing problem is to estimate the quantity $\hat{\mathbf{x}}_{k|j} = E(\mathbf{x}_k|\overline{\mathbf{v}}_j)$ where $j > k$.

The Kalman Filter

The Kalman filter is named after Rudolph E. Kalman, who in 1960 [86] published his famous paper describing a recursive solution to the discrete-data linear filtering problem. We wish to determine an optimal estimate $\hat{\mathbf{x}}_k$ of the state of the system $\mathbf{x}_k$ at the time k. The main idea is to obtain this optimal estimate recursively: given a prior optimal estimate of the state of the system at time k denoted by $\hat{\mathbf{x}}_{k|k-1}$, we wish to obtain an updated estimate $\hat{\mathbf{x}}_{k|k}$ after measurement $\mathbf{v}_k$ is known[9]. The optimality is in the sense of least squares to minimize the estimated error covariance.

Least squares preliminaries:

We now briefly discuss how the principle of least squares can be applied in designing a state estimator for a linear stochastic system. Consider the measurement equation in Equation A.2 which shows the observed data contaminated with noise. The goal

[7] Independent in the sense that $E(\underline{\xi}_k\underline{\eta}_l^T) = 0$ for all k and l.

[8] We will denote $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k}$ as the estimate of $\mathbf{x}_k$ in the time interval k.

[9] The presentation here is based on [61]. Readers interested in a more rigrous derivation are referred to [86].

is to obtain an optimal estimate $\hat{\mathbf{y}}_k$ of the state vector $\mathbf{x}_k$ from the information $\{\mathbf{v}_k\}$. This is done by minimizing the quantity

$$F(\mathbf{y}_k, W_k) = E(\mathbf{v}_k - C_k\mathbf{y}_k)^T W_k(\mathbf{v}_k - C_k\mathbf{y}_k)$$

over all n-vectors $\mathbf{y}_k$ where W_k is a positive definite and symmetric matrix, called a *weight matrix.* That is, we wish to find a $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_k(W_k)$ such that

$$F(\hat{\mathbf{y}}_k, W_k) = \min_{\mathbf{y}_k} F(\mathbf{y}_k, W_k) \tag{A.3}$$

where *min* denotes the *minimum.* In addition, we wish to determine the optimal weight $\hat{W}_k$. It can be proved that the optimal weight matrix is $\hat{W}_k = R_k^{-1}$, and the optimal estimate of $\mathbf{x}_k$ using this optimal weight is[10]

$$\hat{\mathbf{x}}_k := \hat{\mathbf{y}}_k(R_k^{-1}) = (C_k^T R_k^{-1} C_k)^{-1} C_k^T R_k^{-1} \mathbf{v}_k \tag{A.4}$$

We call $\hat{\mathbf{x}}_k$ the *least squares* optimal estimate of $\mathbf{x}_k$[11].

Derivation of Kalman filtering algorithm

To find the minimum variance estimate of the state vector, we incorporate the information of all measurements $\mathbf{v}_j$, $j = 0, 1, \cdots, k$, in determining the estimate $\hat{\mathbf{x}}_k$ of $\mathbf{x}_k$. To accomplish this, we introduce the vectors:

$$\overline{\mathbf{v}}_j = [\mathbf{v}_0^T \cdots \mathbf{v}_j^T]^T, \quad j = 0, 1, ... \tag{A.5}$$

[10] For detailed proof please refer to [61].

[11] This estimate is also the minimum variance estimate of $\mathbf{x}_k$.

It can then be shown with some algebraic manipulation that the state space description of the linear system can be written as

$$\overline{\mathbf{v}}_j = H_{k,j}\mathbf{x}_k + \overline{\underline{\epsilon}}_{k,j}, \tag{A.6}$$

where

$$H_{k,j} = \begin{bmatrix} C_0\Phi_{0,k} \\ \vdots \\ C_j\Phi_{j,k} \end{bmatrix} \quad \text{and} \quad \overline{\underline{\epsilon}}_{k,j} = \begin{bmatrix} \underline{\epsilon}_{k,0} \\ \vdots \\ \underline{\epsilon}_{k,j} \end{bmatrix}$$

with $\Phi_{l,k}$ being the transition matrices defined by

$$\Phi_{l,k} = \begin{cases} A_{l-1}\cdots A_k & \text{if } l > k \\ I & \text{if } l = k \end{cases}$$

$\Phi_{l,k} = \Phi_{k,l}^{-1}$ if $l < k$ and

$$\underline{\epsilon}_{k,l} = \underline{\eta}_l - C_l \sum_{i=l+1}^{k} \Phi_{l,i}\Gamma_{i-1}\underline{\xi}_{i-1}$$

For real-time applications, we need a recursive relation for the optimal state estimate. We will derive a recursive formula that gives $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k}$ from the "prediction" $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k-1}$ from the estimate $\hat{\mathbf{x}}_{k-1} = \hat{\mathbf{x}}_{k-1|k-1}$. At each step, we will only use the incoming bit of the measurement information so that very little data storage is necessary.

The prediction-correction formulation

To compute $\hat{\mathbf{x}}_k$ in real-time, we will derive the recursive formula

$$\begin{cases} \hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - C_k\hat{\mathbf{x}}_{k|k-1}) \\ \hat{\mathbf{x}}_{k|k-1} = A_{k-1}\hat{\mathbf{x}}_{k-1|k-1}, \end{cases} \tag{A.7}$$

where G_k will be called the *Kalman gain* matrices. The starting point is the initial estimate $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}_{0|0}$. Since $\hat{\mathbf{x}}_0$ is an unbiased estimate of the initial state $\mathbf{x}_0$, we can use $\hat{\mathbf{x}}_0 = E(\mathbf{x}_0)$, which is a constant vector. The Kalman gain matrices G_k also have to computed recursively. These two recursive processes together are called the *Kalman filtering process.*

Let $\hat{\mathbf{x}}_{k|j}$ be the (optimal) minimum variance least squares estimate of $\mathbf{x}_k$ by choosing the weight matrix to be

$$W_{k,j} = (Var(\underline{\epsilon}_{k,j}))^{-1}$$

using $\overline{\mathbf{v}}_j$ in Equation A.6. It can now be verified that

$$W_{k,k-1}^{-1} = \begin{bmatrix} R_0 & & 0 \\ & \ddots & \\ 0 & & R_{k-1} \end{bmatrix} + Var \begin{bmatrix} C_0 \sum_{i=1}^{k} \Phi_{0,i}\Gamma_{i-1}\underline{\xi}_{i-1} \\ \vdots \\ C_{k-1}\Phi_{k-1,k}\Gamma_{k-1}\underline{\xi}_{k-1} \end{bmatrix} \tag{A.8}$$

and

$$W_{k,k}^{-1} = \begin{bmatrix} W_{k,k-1}^{-1} & 0 \\ 0 & R_k \end{bmatrix} \tag{A.9}$$

Now it can be followed from the least squares formulation that (Equation A.4)

$$\hat{\mathbf{x}}_{k|j} = (H_{k,j}^T W_{k,j} H_{k,j})^{-1} H_{k,j}^T W_{k,j} \overline{\mathbf{v}}_j \tag{A.10}$$

Our first goal is to relate $\hat{\mathbf{x}}_{k|k-1}$ with $\hat{\mathbf{x}}_{k|k}$. To do so, we observe that

$$\begin{aligned} H_{k,k}^T W_{k,k} H_{k,k} &= \begin{bmatrix} H_{k,k-1}^T C_k^T \end{bmatrix} \begin{bmatrix} W_{k,k-1} & 0 \\ 0 & R_k^{-1} \end{bmatrix} \begin{bmatrix} H_{k,k-1} \\ C_k \end{bmatrix} \\ &= H_{k,k-1}^T W_{k,k-1} H_{k,k-1} + C_k^T R_k^{-1} C_k \end{aligned} \tag{A.11}$$

and

$$H_{k,k}^T W_{k,k} \overline{\mathbf{v}}_k = H_{k,k-1}^T W_{k,k-1} \overline{\mathbf{v}}_{k-1} + C_k^T R_k^{-1} \mathbf{v}_k. \tag{A.12}$$

Using Equations A.10, A.11 and A.12, we have

$$\begin{aligned}&(H^T_{k,k-1}W_{k,k-1}H_{k,k-1} + C^T_k R^{-1}_k C_k)\hat{\mathbf{x}}_{k|k-1}\\&= H^T_{k,k-1}W_{k,k-1}\overline{\mathbf{v}}_{k-1} + C^T_k R^{-1}_k C_k \hat{\mathbf{x}}_{k|k-1}\end{aligned} \tag{A.13}$$

and

$$\begin{aligned}&(H^T_{k,k-1}W_{k,k-1}H_{k,k-1} + C^T_k R^{-1}_k C_k)\hat{\mathbf{x}}_{k|k}\\&= (H^T_{k,k}W_{k,k}H_{k,k})\hat{\mathbf{x}}_{k|k}\\&= H^T_{k,k-1}W_{k,k-1}\overline{\mathbf{v}}_{k-1} + C^T_k R^{-1}_k \mathbf{v}_k.\end{aligned} \tag{A.14}$$

Subtracting Equation A.13 from Equation A.14 we get

$$\begin{aligned}&(H^T_{k,k-1}W_{k,k-1}H_{k,k-1} + C^T_k R^{-1}_k C_k)(\hat{\mathbf{x}}_{k|k} - \hat{\mathbf{x}}_{k|k-1})\\&= C^T_k R^{-1}_k(\mathbf{v}_k - C_k\hat{\mathbf{x}}_{k|k-1})\end{aligned} \tag{A.15}$$

Now we define

$$G_k = (H^T_{k,k-1}W_{k,k-1}H_{k,k-1} + C^T_k R^{-1}_k C_k)^{-1} C^T_k R^{-1}_k \tag{A.16}$$

$$= (H^T_{k,k}W_{k,k}H_{k,k})^{-1} C^T_k R^{-1}_k. \tag{A.17}$$

Then we have

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - C_k\hat{\mathbf{x}}_{k|k-1}) \tag{A.18}$$

Since $\hat{\mathbf{x}}_{k|k-1}$ is a one step prediction and $(\mathbf{v}_k - C_k\hat{\mathbf{x}}_{k|k-1})$ is the error between the real measurement and the prediction, Equation A.18 is in fact a “prediction-correction” formula with the Kalman gain matrix G_k as a weight matrix. To complete the recursive process, we need an equation that gives $\hat{\mathbf{x}}_{k|k-1}$ from $\hat{\mathbf{x}}_{k-1|k-1}$. Equation A.19 gives that relation[12].

$$\hat{\mathbf{x}}_{k|k-1} = A_{k-1}\hat{\mathbf{x}}_{k-1|k-1} \tag{A.19}$$

We now need a recursive scheme to obtain the Kalman gain matrices G_k. We write

[12] Detailed proof can be found in [61].

$G_k = P_{k,k}C_k^T R_k^{-1}$ where $P_{k,k} = (H_{k,k}^T W_{k,k} H_{k,k})$ and set $P_{k,k-1} = (H_{k,k-1}^T W_{k,k-1} H_{k,k-1})$. Then using Equation A.11 we obtain $P_{k,k}^{-1} = P_{k,k-1}^{-1} + C_k^T R_k^{-1} C_k$.

It can be proved that

$$G_k = P_{k,k-1}C_k^T (C_k P_{k,k-1} C_k^T + R_k)^{-1} \tag{A.20}$$

so that,

$$P_{k,k} = (I - G_k C_k)P_{k,k-1}. \tag{A.21}$$

Furthermore, it can be shown that

$$P_{k,k-1} = A_{k-1}P_{k-1,k-1}A_{k-1}^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T. \tag{A.22}$$

Hence using Equations A.21 and A.22 with the initial matrix $P_{0,0}$, we obtain a recursive scheme to compute $P_{k-1,k-1}$, $P_{k,k-1}$, G_k and $P_{k,k}$ for $k = 1, 2, \cdots$. Moreover using Equation A.6 and Equation A.10 it can also be shown that

$$P_{k,k-1} = E(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^T \tag{A.23}$$

and that

$$P_{k,k} = E(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^T. \tag{A.24}$$

In particular, we have

$$P_{0,0} = E(\mathbf{x}_0 - E\mathbf{x}_0)(\mathbf{x}_0 - E\mathbf{x}_0)^T = Var(\mathbf{x}_0). \tag{A.25}$$

Finally, combining all the results obtained above, we arrive at the Kalman filtering process in Equation A.26 for the linear stochastic system with state space description as in Equation A.1 and Equation A.2:

$$\begin{cases} P_{0,0} = & Var(\mathbf{x}_0), \quad \hat{\mathbf{x}}_0 = E(\mathbf{x}_0) \\ \text{For } k = & 1, 2, \cdots, \\ P_{k,k-1} = & A_{k-1}P_{k-1,k-1}A_{k-1}^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T \\ \hat{\mathbf{x}}_{k|k-1} = & A_{k-1}\hat{\mathbf{x}}_{k-1|k-1} \\ G_k = & P_{k,k-1}C_k^T(C_kP_{k,k-1}C_k^T + R_k)^{-1} \\ P_{k,k} = & (I - G_kC_k)P_{k,k-1} \\ \hat{\mathbf{x}}_{k|k} = & \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - C_k\hat{\mathbf{x}}_{k|k-1}) \end{cases} \tag{A.26}$$

Let us now consider the state space description of a general linear deterministic/stochastic system (where a deterministic control input is also present)

$$\begin{cases} \mathbf{x}_{k+1} & = A_k\mathbf{x}_k + B_k\mathbf{u}_k + \Gamma_k\underline{\xi}_k \\ \mathbf{v}_k & = C_k\mathbf{x}_k + D_k\mathbf{u}_k + \underline{\eta}_k \end{cases} \tag{A.27}$$

where B_k and D_k are constant matrices. $\{\mathbf{u}_k\}$ is the sequence of control inputs. The system in Equation A.27 can be decomposed into the sum of a linear deterministic system:

$$\text{Deterministic:} \begin{cases} \mathbf{z}_{k+1} & = A_k\mathbf{z}_k + B_k\mathbf{u}_k \\ \mathbf{s}_k & = C_k\mathbf{z}_k + D_k\mathbf{u}_k \end{cases} \tag{A.28}$$

and a (purely) stochastic system:

$$\text{Stochastic:} \begin{cases} \mathbf{d}_{k+1} & = A_k\mathbf{d}_k + \Gamma_k\underline{\xi}_k \\ \mathbf{w}_k & = C_k\mathbf{w}_k + \underline{\eta}_k \end{cases} \tag{A.29}$$

with $\mathbf{x}_k = \mathbf{z}_k + \mathbf{d}_k$ and $\mathbf{v}_k = \mathbf{s}_k + \mathbf{w}_k$. The solution $\mathbf{z}_k$ of the linear deterministic system is well known and is given as follows:

$$\mathbf{z}_k = (A_{k-1} \cdots A_0)\mathbf{z}_0 + \sum_{i=1}^{k}(A_{k-1} \cdots A_{i-1})B_{i-1}\mathbf{u}_{i-1} \tag{A.30}$$

Hence, the optimal estimate of the state vector $\hat{\mathbf{x}}_k$ in Equation A.27 can be obtained by

$$\hat{\mathbf{x}}_k = \hat{\mathbf{d}}_k + \mathbf{z}_k. \tag{A.31}$$

Thus, by superimposing the deterministic solution with the Kalman filtering equations for a purely stochastic system (Equation A.26) we can obtain the following Kalman filtering process for the linear stochastic/deterministic system (Equation A.27) as:

$$\begin{cases} P_{0,0} = & Var(\mathbf{x}_0), \quad \hat{\mathbf{x}}_0 = E(\mathbf{x}_0) \\ \text{For } k = & 1, 2, \cdots, \\ P_{k,k-1} = & A_{k-1} P_{k-1,k-1} A_{k-1}^T + \Gamma_{k-1} Q_{k-1} \Gamma_{k-1}^T \\ \hat{\mathbf{x}}_{k|k-1} = & A_{k-1} \hat{\mathbf{x}}_{k-1|k-1} + B_{k-1} \mathbf{u}_{k-1} \\ G_k = & P_{k,k-1} C_k^T (C_k P_{k,k-1} C_k^T + R_k)^{-1} \\ P_{k,k} = & (I - G_k C_k) P_{k,k-1} \\ \hat{\mathbf{x}}_{k|k} = & \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - D_k \mathbf{u}_k - C_k \hat{\mathbf{x}}_{k|k-1}) \end{cases} \tag{A.32}$$

Extended Kalman filter

The Kalman filtering process described earlier was designed for linear systems. For a general non linear system, a linearization procedure is performed in deriving the filtering equations. We will consider a real-time linear Taylor series approximation of the system function at the previous state estimate and that of the observation function at the corresponding prediction position. The Kalman filter so obtained is generally called as the *extended Kalman filter*.

Consider a non linear system in the following form:

$$\begin{aligned} \mathbf{x}_{k+1} =& \mathbf{f}_k(\mathbf{x}_k) + H_k(\mathbf{x}_k) \underline{\xi}_k \\ \mathbf{v}_k =& \mathbf{g}_k(\mathbf{x}_k) + \underline{\eta}_k \end{aligned} \tag{A.33}$$

where $\mathbf{x}$ is the state vector, $\mathbf{v}$ is the output vector, $\mathbf{f}_k$ and $\mathbf{g}_k$ are vector valued functions and H_k is a matrix valued function, such that for each k the first order partial derivatives of $\mathbf{f}_k(\mathbf{x}_k)$ and $\mathbf{g}_k(\mathbf{x}_k)$ with respect to all the components of $\mathbf{x}_k$ are continuous. Also, we consider zero mean Gaussian white noise sequences $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$. Similar to the assumptions made for the linear model, we make the following assumptions

$$\begin{cases} E(\underline{\xi}_k\underline{\xi}_l^T) = Q_k\delta_{kl}, \quad E(\underline{\eta}_k\underline{\eta}_l^T) = R_k\delta_{kl} \\ \quad E(\underline{\xi}_k\underline{\eta}_l^T) = 0, \quad E(\underline{\xi}_k\mathbf{x}_0^T) = 0 \\ \quad E(\underline{\eta}_k\mathbf{x}_0^T) = 0 \end{cases} \tag{A.34}$$

where Q_k and R_k are the variance matrices for random vectors $\{\underline{\xi}_k\}$ and $\{\underline{\eta}_k\}$ respectively.

The initial estimate $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}_{0|0}$ and the first prediction $\hat{\mathbf{x}}_{1|0}$ are chosen to be:

$$\begin{cases} \hat{\mathbf{x}}_0 & = E(\mathbf{x}_0) \\ \hat{\mathbf{x}}_{1|0} & = \mathbf{f}_0(\hat{\mathbf{x}}_0) \end{cases} \tag{A.35}$$

We will now formulate $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k}$, consecutively for $k = 1, 2, \cdots$, using the predicted estimate

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{f}_k(\hat{\mathbf{x}}_k) \tag{A.36}$$

and the linear state space description

$$\begin{cases} \mathbf{x}_{k+1} & = A_k\mathbf{x}_k + \mathbf{u}_k + \Gamma_k\underline{\xi}_k \\ \mathbf{w}_k & = C_k\mathbf{x}_k + \underline{\eta}_k \end{cases} \tag{A.37}$$

where A_k, $\mathbf{u}_k$, Γ_k,$\mathbf{w}_k$ and C_k are to be determined in real-time as follows.

Suppose that $\hat{\mathbf{x}}_j$ has been determined so that $\hat{\mathbf{x}}_{j+1|j}$ is also defined as in Equation A.36, for $j = 1, 2, \cdots, k$. We now consider the linear approximation of $\mathbf{f}_k(\mathbf{x}_k)$ at $\hat{\mathbf{x}}_k$

and that of $\mathbf{g}_k(\mathbf{x}_k)$ at $\hat{\mathbf{x}}_{k|k-1}$ as follows:

$$\begin{cases} \mathbf{f}_k(\mathbf{x}_k) & \simeq \mathbf{f}_k(\hat{\mathbf{x}}_k) + A_k(\mathbf{x}_k - \hat{\mathbf{x}}_k) \\ \mathbf{g}_k(\mathbf{x}_k) & \simeq \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1}) + C_k(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) \end{cases} \tag{A.38}$$

where[13]

$$A_k = \left[\frac{\partial \mathbf{f}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_k)\right] \qquad \textit{and} \qquad C_k = \left[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\right]$$

Hence, by setting

$$\begin{cases} \mathbf{u}_k & = \mathbf{f}_k(\hat{\mathbf{x}}_k) - A_k(\hat{\mathbf{x}}_k) \\ \Gamma_k & = H_k(\hat{\mathbf{x}}_k) \\ \mathbf{w}_k & = \mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1}) + C_k\hat{\mathbf{x}}_{k|k-1} \end{cases} \tag{A.39}$$

the nonlinear model in Equation A.33 may be approximated by the linear model in Equation A.37.

It is to be noted that the linearization procedure adopted above is possible only if $\hat{\mathbf{x}}_k$ has been determined. Since we assume that $\hat{\mathbf{x}}_0$ is known, the system description in Equation A.37 is valid for $k = 0$. From this, we define $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_{1|1}$ as the optimal estimate of $\mathbf{x}_1$ in the linear model in Equation A.37, using the measurements $[\mathbf{v}_0^T \; \mathbf{w}_1^T]^T$. Now, by applying Equation A.36, Equation A.37 is established for $k = 1$, so that

[13] Here, for any vector valued function

$$\mathbf{h}(\mathbf{x}_k) = \begin{bmatrix} h_1(\mathbf{x}_k) \\ \vdots \\ h_m(\mathbf{x}_k) \end{bmatrix}$$

where $\mathbf{x}_k = [x_k^1 \cdots x_k^n]^T$ we denote, as usual,

$$\left[\frac{\partial \mathbf{h}}{\partial \mathbf{x}_k}(\mathbf{x}_k^*)\right] = \begin{bmatrix} \frac{\partial h_1}{\partial x_k^1}(\mathbf{x}_k^*) & \cdots & \frac{\partial h_1}{\partial x_k^n}(\mathbf{x}_k^*) \\ \vdots & & \vdots \\ \frac{\partial h_m}{\partial x_k^1}(\mathbf{x}_k^*) & \cdots & \frac{\partial h_m}{\partial x_k^n}(\mathbf{x}_k^*) \end{bmatrix}$$

$\hat{\mathbf{x}}_2 = \hat{\mathbf{x}}_{2|2}$ can be determined analogously, using the data $[\mathbf{v}_0^T \ \mathbf{w}_1^T \ \mathbf{w}_2^T]^T$, etc. From the Kalman filtering results for general linear deterministic/stochastic state space descriptions, we can write the "correction" formula as:

$$\begin{aligned} \hat{\mathbf{x}}_k &= \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{w}_k - C_k\hat{\mathbf{x}}_{k|k-1}) \\ &= \hat{\mathbf{x}}_{k|k-1} + G_k\big((\mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1}) + C_k\hat{\mathbf{x}}_{k|k-1}) - C_k\hat{\mathbf{x}}_{k|k-1}\big) \\ &= \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1})) \end{aligned}$$

where G_k is the Kalman gain matrix for the linear model in Equation A.37.

The resulting extended Kalman filtering process can be summarized as follows in Equation A.40.

$$\left\{ \begin{array}{rl} P_{0,0} = & Var(\mathbf{x}_0), \quad \hat{\mathbf{x}}_0 = E(\mathbf{x}_0) \\ For \quad k = & 1, 2, ..., \\ P_{k,k-1} = & \left[\frac{\partial \mathbf{f}_{k-1}}{\partial \mathbf{x}_{k-1}}(\hat{\mathbf{x}}_{k-1})\right] P_{k-1,k-1} \left[\frac{\partial \mathbf{f}_{k-1}}{\partial \mathbf{x}_{k-1}}(\hat{\mathbf{x}}_{k-1})\right]^T + H_{k-1}(\hat{\mathbf{x}}_{k-1}) Q_{k-1} H_{k-1}^T(\hat{\mathbf{x}}_{k-1}) \\ \hat{\mathbf{x}}_{k|k-1} = & \mathbf{f}_{k-1}(\hat{\mathbf{x}}_{k-1}) \\ G_k = & P_{k,k-1}\left[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\right]^T \cdot \left[\left[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\right] P_{k,k-1}\left[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\right]^T + R_k\right]^{-1} \\ P_{k,k} = & \left[I - G_k\left[\frac{\partial \mathbf{g}_k}{\partial \mathbf{x}_k}(\hat{\mathbf{x}}_{k|k-1})\right]\right] P_{k,k-1} \\ \hat{\mathbf{x}}_{k|k} = & \hat{\mathbf{x}}_{k|k-1} + G_k(\mathbf{v}_k - \mathbf{g}_k(\hat{\mathbf{x}}_{k|k-1})) \end{array} \right. \tag{A.40}$$

APPENDIX B

ARIMA MODELING OF TRAFFIC TIME SERIES DATA AND ITS COMPARISON WITH THE NON-CONTINUUM MODEL PERFORMANCE

In this appendix we present a brief overview of Autoregressive Integrated Moving Average (ARIMA) approach towards modeling of time series data and its application towards traffic state predictions. We also compare the results of the one step traffic state prediction using ARIMA modeling approach with those obtained by using the non-continuum modeling approach.

ARIMA modeling: Box Jenkins approach

The Box-Jenkins approach [87] is used to construct a one step predictor model to predict freeway traffic state variables. Let X_t represent a non-seasonal time series of observations taken at equally spaced time intervals. The time series X_t is either stationary or reducible to a stationary form Z_t, by computing the difference for some integer number of times d such that:

$$Z_t = (1 - B)^d X_t \tag{B.1}$$

where B is the back-shift operator defined as $BX_t = X_{t-1}$.
A time series can be represented by the following general class of linear models:

$$\Phi_p(B)(1 - B)^d(X_t - \mu) = \Theta_q(B)a_t \tag{B.2}$$

where p, d, q are non negative integers and μ is the mean of the series. $\Phi_p(B)$ is the autoregressive operator of order p defined as

$$\Phi_p(B) = 1 - \phi_1 B ... - \phi_p B^p$$

where $\phi_1, \cdots, \phi_p$ are the autoregressive (AR) coefficients. $\Theta_q(B)$ is the moving average operator of order q defined as

$$\Theta_q(B) = 1 - \theta_1 B ... - \theta_q B^q$$

where $\theta_1, \cdots, \theta_q$ are the moving average (MA) coefficients. a_t are the random disturbances assumed to be independent and identically distributed (iid) with zero mean and variance σ_a^2. The model in Equation B.2 is called an autoregressive integrated moving average (ARIMA) model of order (p, d, q).

ARIMA models are fitted to a particular data set by a three stage iterative procedure:

1. Preliminary Identification
2. Estimation
3. Diagnostic check

In preliminary identification, the values of p, d, and q are determined by inspecting the autocorrelations and partial autocorrelations of the series or its differences, or both and by comparing with those of some basic stochastic processes. The sample autocorrelation function is given by

$$r_k = \frac{\sum_{t=1}^{n-K}[(X_t - \bar{X})(X_{t+k} - \bar{X}]}{\sum_{t=1}^{n}(X_t - \bar{X})^2}, \quad K = 1, 2, ...$$

where $\bar{X}$ is the sample mean and n is the number of observations. The autocorrelation function of a stochastic process provides a measure of how long a disturbance in the system affects the state of the system in the future. In general:

- **Moving Average Processes**: The autocorrelation function of a moving average process of order q has a cutoff after lag q (memory of lag q), while its partial autocorrelation function tails off.

- **Autoregressive Process**: The autocorrelation function of an autoregressive process of order p tails off in the form of damped exponentials or damped sine waves, while its partial autocorrelation function has a cutoff after lag p.

For mixed processes, both the autocorrelation and partial autocorrelation function tail off. Failure of the autocorrelation function to die out rapidly suggests that differencing is needed ($d > 0$).

Once the values of p, d, and q have been determined, the autoregressive and moving average parameters are estimated by using non linear least squares techniques. Finally the goodness of the model fit is checked. If the form of the chosen model is satisfactory, then the resulting residuals, $\hat{a}_t$, should be uncorrelated random deviations with zero mean[14]. To test for this, Box and Pierce [88] developed an overall test of autocorrelations for lags 1 thorough K. According to the test, the variable Q defined below, is approximately distributed as a chi-square variable with $(K - p - q)$ degrees of freedom.

$$Q = n \sum_{i=1}^{K} r_i^2(\hat{a}) \tag{B.3}$$

In the above equation, n is the number of observations minus the degree of differencing and $r_i(\hat{a})$ is the residual autocorrelation for lag i. K is typically chosen in the range of 15 to 30. Thus, given the time series of the residues, we reject the iid (residues being independent and identically distributed) hypothesis at level of significance α if $Q > \chi^2_{1-\alpha}(h)$, where $\chi^2_{1-\alpha}(h)$ is the $1 - \alpha$ quantile of the chi-square distribution with h degrees of freedom[15].

[14] Residuals are generally defined as the difference between the observation and fitted value.

[15] $h = K - p - q$ in our context.

Modeling US 183 traffic time series data

We consider two locations on US 183 North Bound and analyze the traffic speed time series data. We identify these two locations as Station 23 (at Lazy Lane in section 2) and Station 29 (at Metric Blvd. in section 1). In Fig. 29 we plot the traffic speeds observed at Station 23 and Station 29 on June 14, 2004. We call this as the representative speed data.

Preliminary identification and estimation

We now calculate the sample autocorrelations and partial autocorrelation functions. Fig. 30 and Fig. 31 show the sample autocorrelations and partial autocorrelation functions for the representative speed data and for the data which has been differenced once. Note that in these figures, the blue lines, very close and parallel to the x-axis, represent the error bounds for the data. The bounds are determined based on $\pm 2/\sqrt{n}$. These error bounds are for 95% confidence limits. If the values of the autocorrelation and partial autocorrelation functions lie within these limits, then we consider that they are not significantly different from zero. It can be observed from Fig. 30 and Fig. 31 that the sample autocorrelations of the raw speed time series data for both Station 23 and Station 29 damp off very slowly as lag increases. This suggests that differencing is required.

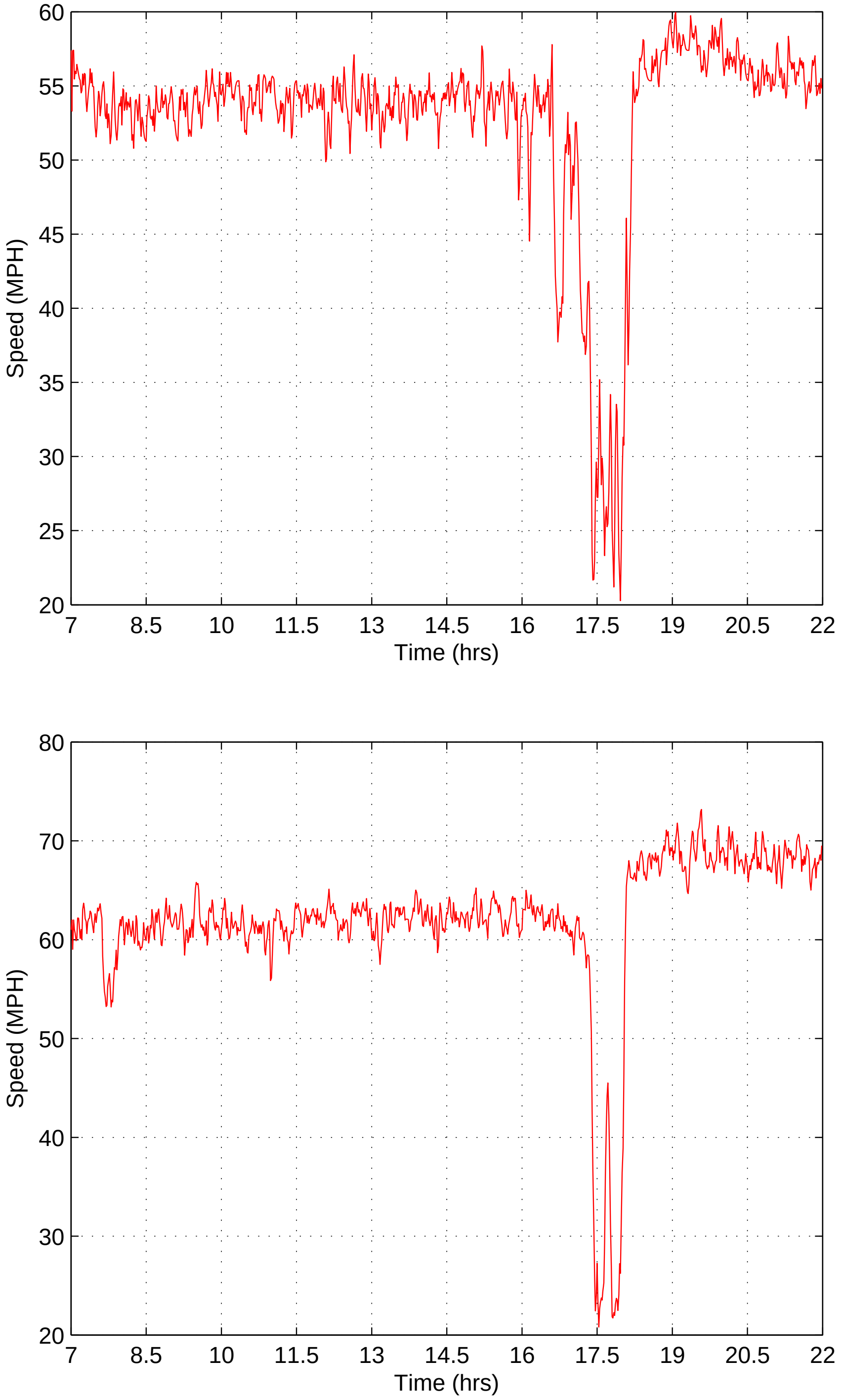

Fig. 29. Representative traffic speeds at stations 23 (top) and 29 (down), June 14, 2004

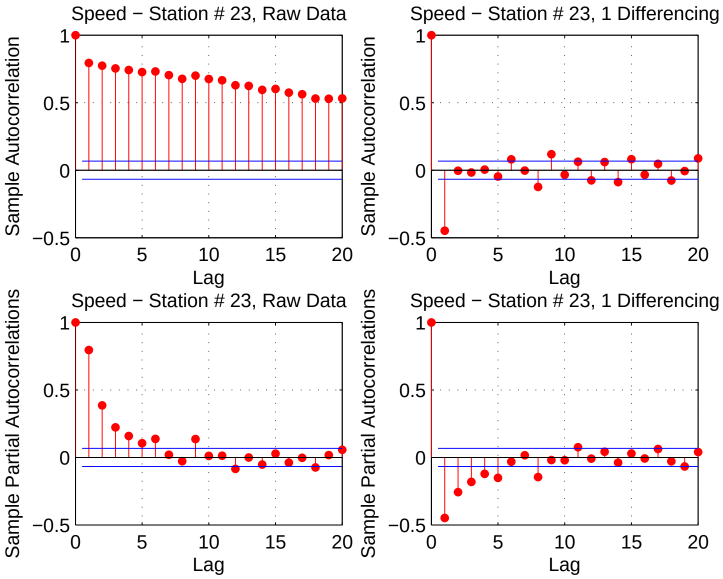

Fig. 30. Sample autocorrelation and partial autocorrelations, speed data, Station 23, June 14, 2004

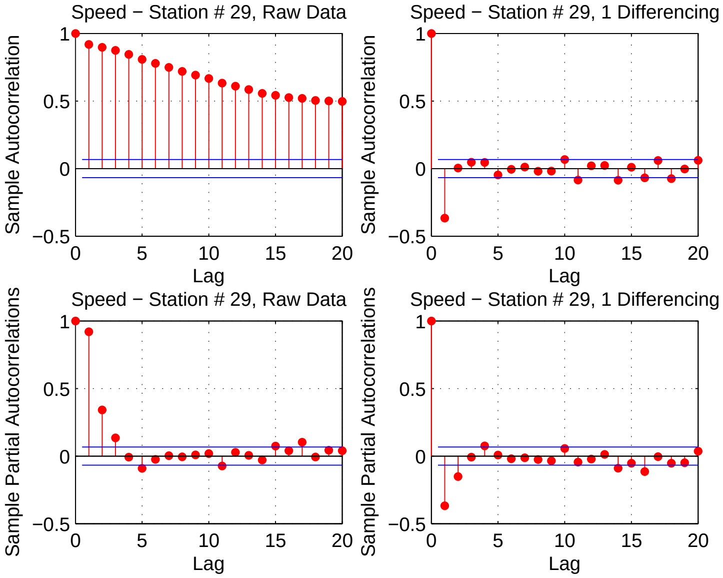

Fig. 31. Sample autocorrelation and partial autocorrelations, speed data, Station 29, June 14, 2004

The sample autocorrelations of the first differences, however, indicate that only the spikes at lag 1 is large in relation to the standard error[16]. The partial autocorrelations of the first differences gradually tail off for both the time series. This suggests that the stochastic process generating the time series is ARIMA (0,1,1); that is the first differences of traffic speed data can be represented by a first order moving average model. Similar autocorrelation and partial autocorrelation function plots were observed for many working days for both Station 23 and Station 29. Thus the speed data at both the detector locations can be modeled as ARIMA (0,1,1) models, albeit with different coefficients.

ARIMA (0,1,1) model can be represented as:

$$(1-B)(X_t-\mu)=(1-\theta_1 B)a_t \tag{B.4}$$

or simply

$$X_t-X_{t-1}=Z_t=a_t-\theta_1 a_{t-1}. \tag{B.5}$$

Two weeks (Monday - Thursday each week) speed time series data, with fifteen hours each day (900 data points each day), was analyzed. In Table IX we list the average parameter (θ_1) values obtained for two weeks of data. In the same table we also list the corresponding standard error.

Diagnostic Check

Diagnostic checking was carried out by inspecting the residuals ($\hat{a}_t$). Fig. 32 and Fig. 33 show the plot of residuals and their autocorrelation functions for speed data with one differencing for Station 23 and Station 29 respectively. From the autocorrelation

[16] The standard error of a sample of sample size n is the sample's standard deviation divided by $\sqrt{n}$.

Table IX. Average θ_1 values and its standard error for Station 23 and Station 29 speed data

Day	Station 23		Station 29	
	Coeffecient	Standard Error	Coeffecient	Standard Error
Monday	0.7456	0.0223	0.6165	0.0176
Tuesday	0.6968	0.0479	0.6278	0.0157
Wednesday	0.7474	0.0247	0.7081	0.0361
Thursday	0.6913	0.0632	0.6261	0.0542

plots we observe that there are no systematic patterns and are all quite small. The average of the residuals and their estimated standard error for speed data from Station 23 are 0.0058 and 0.1293 respectively. Similarly for speed data from Station 29, the average and standard error are 0.0111 and 0.1177 respectively. This strongly suggests that the residuals have zero mean.

We also conducted the chi-square distribution test for the residuals (see Equation B.3). For $K = 25$, the calculated value of Q for Station 23 and Station 25 were 31.36 and 32.83 respectively. Since these values are less than $\chi^2_{0.95}(24)$, we can accept the fitted model as adequate.

There have been several studies where researchers have studied the volume and occupancy loop detector time series data and have attempted to model them as ARIMA models of various orders ([89] - [94]). Each study concluded different model orders, indicating that the performance and the orders was subject to the traffic data used in the study.

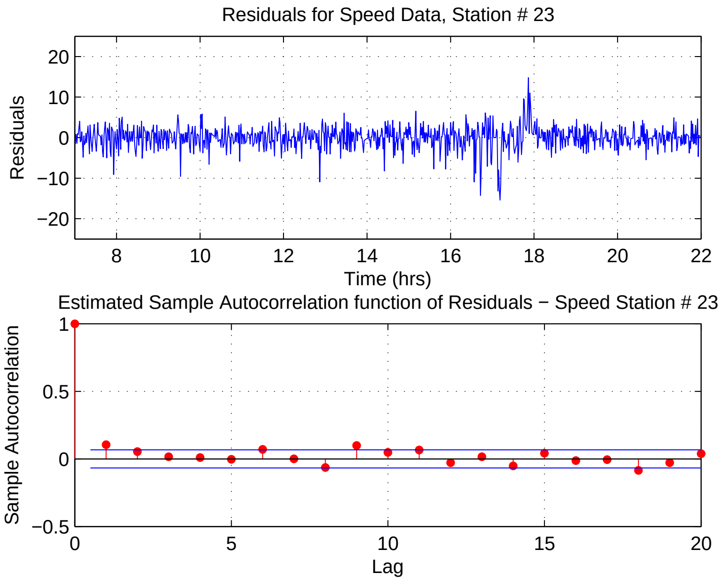

Fig. 32. Residual plots and sample autocorrelations, speed data, Station 23, June 14, 2004

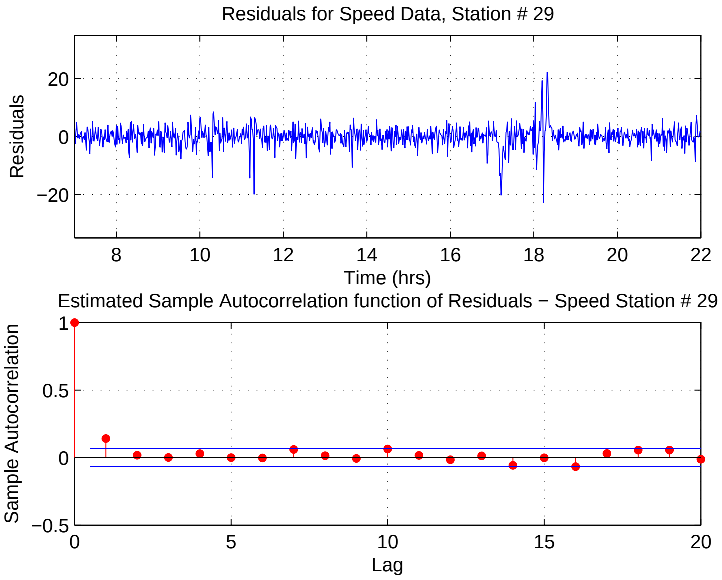

Fig. 33. Residual plots and sample autocorrelations, speed data, Station 29, June 14, 2004

One minute predictions

The ARIMA model can be used to make one step predictions in real-time.

Let $\hat{Z}_{t-1}(1)$ be the one step ahead forecast made at time $(t-1)$ for Z_t, which when observed will be represented by Equation B.5. If $\hat{Z}_{t-1}$ is the minimum mean square error forecast, then its value is determined by the conditional expectation of Z_t, given the history (H_t) of the series up to time t. That is:

$$\begin{aligned}\hat{Z}_{t-1} &= E(Z_t/H_t) \\ &= -\theta_1 a_{t-1} \qquad \text{(B.6)}\end{aligned}$$

Therefore, the forecast error at time $(t-1)$ is determined by subtracting Equation B.6 from Equation B.5:

$$\begin{aligned}e_{t-1}(1) &= Z_t - \hat{Z}_{t-1} \\ &= a_t \qquad \text{(B.7)}\end{aligned}$$

Hence, the white noise that generated the ARIMA (0,1,1) process is the one step ahead forecast error. Thus, we obtain the following update expression for the model:

$$\hat{Z}_t(1) = -\theta_1 e_{t-1}(1) \qquad \text{(B.8)}$$

Fig. 34 show the one step (one minute) forecasts of the traffic speed at Station 23 and Station 29.

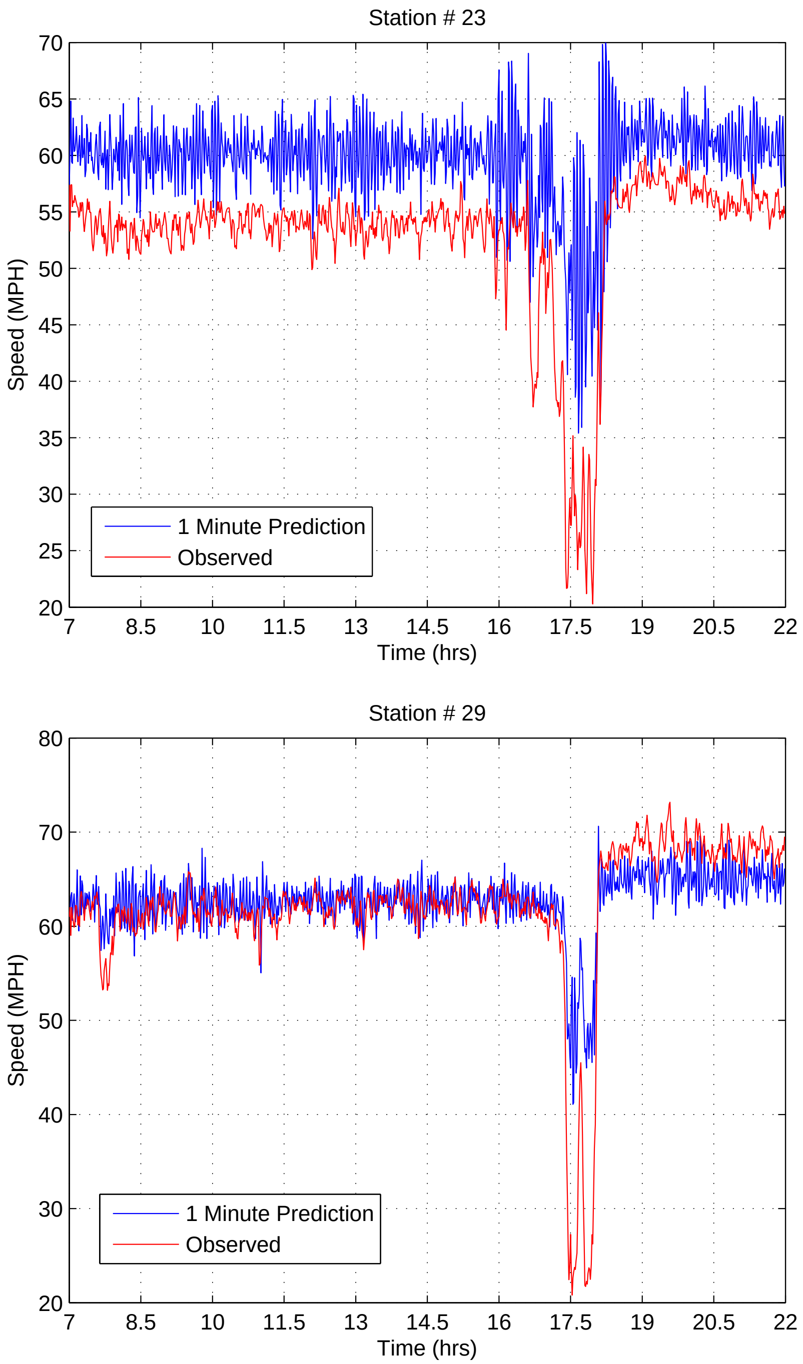

Fig. 34. One minute ARIMA model prediction, Station 23 and Station 29, June 14, 2004

Fig. 35 and Fig. 36 show the comparison of one step (one minute) forecasts of the traffic speed by non-continuum and ARIMA modeling approaches at Station 23 and Station 29 respectively. It can be seen, that for the selected locations, the non-continuum traffic flow model is able to predict the traffic speed much better than the ARIMA model. One important point to note is that the speed prediction results obtained from the ARIMA model are very noisy. Also, the fitted ARIMA model was not able to fully capture the sharp speed drops which occur during peak congestion.

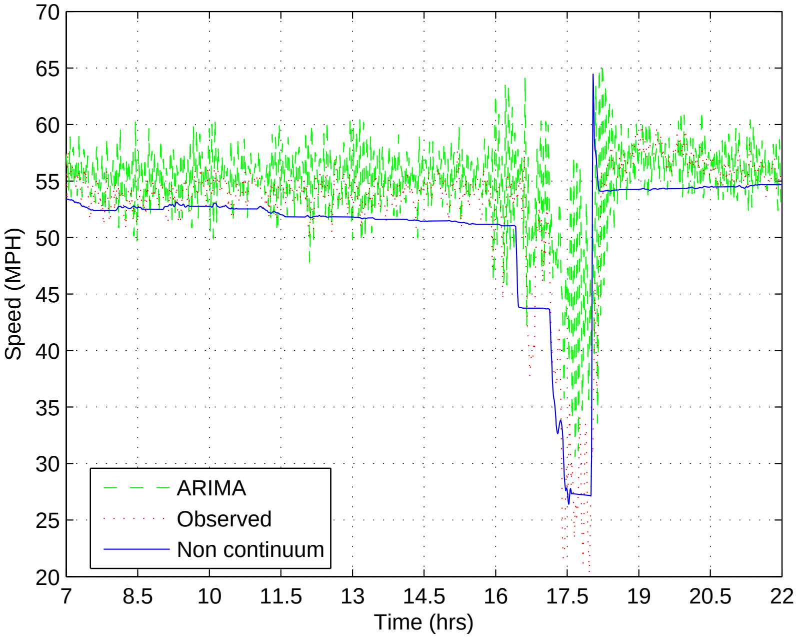

Fig. 35. One minute predictions - Comparing non-continuum and ARIMA models, Station 23, June 14, 2004

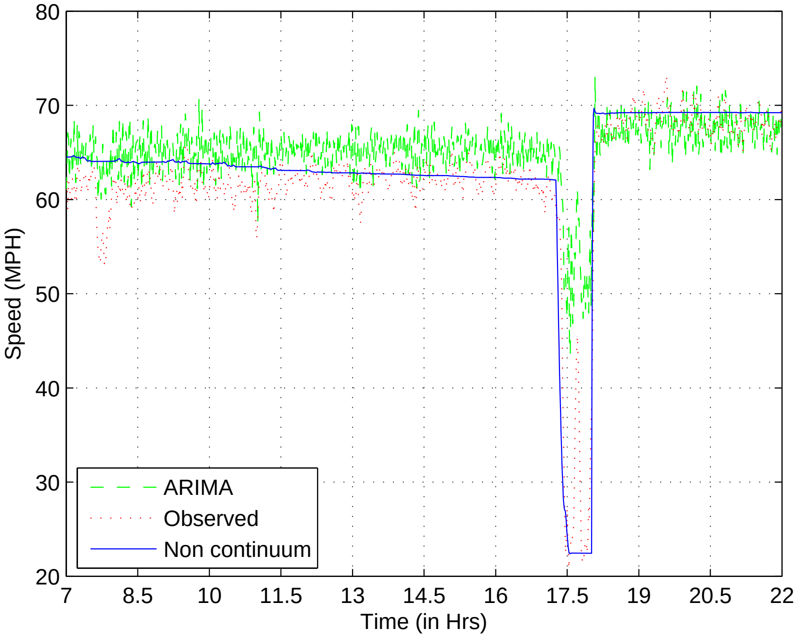

Fig. 36. One minute predictions - Comparing non-continuum and ARIMA models, Station 29, June 14, 2004

VITA

Vipin Tyagi was born in Delhi, the capital of India. He was brought up in Uttar Pradesh (a state in the northern part of India) and completed his high school in the city of Ghaziabad. He joined the Indian Institute of Technology Bombay, Mumbai, India in August 1997 and graduated with a Bachelor of Technology and a Master of Technology in Mechanical Engineering in 2002. He joined the Department of Mechanical Engineering at Texas A&M University in August 2002 to pursue his doctorate in mechanical engineering and graduated with a Ph.D. in May 2007. He may be contacted through Professor Swaroop Darbha, Department of Mechanical Engineering, Texas A&M University, College Station, Texas 77843, USA.