Off-line Calibration of Dynamic Traffic Assignment Models

by

Ramachandran Balakrishna

Bachelor of Technology in Civil Engineering
Indian Institute of Technology, Madras, India (1999)
Master of Science in Transportation
Massachusetts Institute of Technology (2002)

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

Doctor of Philosophy in the Field of Transportation Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006

Author .....
Department of Civil and Environmental Engineering
May 12, 2006

Certified by .....
Moshe E. Ben-Akiva
Edmund K. Turner Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by .....
Haris N. Koutsopoulos
Associate Professor of Civil and Environmental Engineering,
Northeastern University
Thesis Supervisor

Accepted by .....
Andrew Whittle
Chairman, Departmental Committee for Graduate Students

Off-line Calibration of Dynamic Traffic Assignment Models

by

Ramachandran Balakrishna

Submitted to the Department of Civil and Environmental Engineering
on May 12, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in the Field of Transportation Systems

Abstract

Advances in Intelligent Transportation Systems (ITS) have resulted in the deployment of surveillance systems that automatically collect and store extensive network-wide traffic data. Dynamic Traffic Assignment (DTA) models have also been developed for a variety of dynamic traffic management applications. Such models are designed to estimate and predict the evolution of congestion through detailed models and algorithms that capture travel demand, network supply and their complex interactions. The availability of rich time-varying traffic data spanning multiple days thus provides the opportunity to calibrate a DTA model's many inputs and parameters, so that its outputs reflect field conditions.

The current state of the art of DTA model calibration is a sequential approach, in which supply model calibration (assuming known demand inputs) is followed by demand calibration with fixed supply parameters. In this thesis, we develop an off-line DTA model calibration methodology for the simultaneous estimation of all demand and supply inputs and parameters, using sensor data. We adopt a minimization formulation that can use any general traffic data, and present approaches to solve the complex, non-linear, stochastic optimization problem. Case studies with DynaMIT, a DTA model with traffic estimation and prediction capabilities, are used to demonstrate and validate the proposed methodology. A synthetic traffic network with known demand parameters and simulated sensor data is used to illustrate the improvement over the sequential approach, the ability to accurately recover underlying model parameters, and robustness in a variety of demand and supply situations. Archived sensor data and a network from Los Angeles, CA are then used to demonstrate scalability. The benefit of the proposed methodology is validated through a real-time test of the calibrated DynaMIT's estimation and prediction accuracy, based on sensor data not used for calibration. Results indicate that the simultaneous approach significantly outperforms the sequential state of the art.

Thesis Supervisor: Moshe E. Ben-Akiva
Title: Edmund K. Turner Professor of Civil and Environmental Engineering

Thesis Supervisor: Haris N. Koutsopoulos
Title: Associate Professor of Civil and Environmental Engineering, Northeastern University

Acknowledgments

This thesis would not have been possible without contributions from various quarters. Foremost, I would like to acknowledge the support and inputs from my thesis supervisors, Professors Moshe Ben-Akiva and Haris N. Koutsopoulos. They have set extremely high standards for me, and have taught by example.

My doctoral committee has been an invaluable source of suggestions, advice and encouragement. I would like to thank Prof. Nigel Wilson, Dr Kalidas Ashok and Dr Tomer Toledo for their support and guidance.

Faculty and friends have contributed immensely through informal discussions outside the classroom/lab. Their genuine interest in my work has been a source of encouragement, and has helped me place this research in perspective. I thank Professors Nigel Wilson, Patrick Jaillet, Cindy Barnhart, Joe Sussman, Ikki Kim, Michel Bierlaire and Brian Park, and PhD students Hai Jiang and Yang Wen, for their insights. Other friends including Dr Arvind Sankar, Prof. Lakshmi Iyer, Dr K. V. S. Vinay and lab-mates Vikrant Vaze and Varun Ramanujam have routinely buttonholed me on my latest results, which has helped me clarify concepts in my own mind.

I am grateful to Dr Henry Lieu and Raj Ghaman of the Federal Highway Administration, whose funding supported much of this research. The data for the Los Angeles analysis was provided by the California PATH program, and Gabriel Murillo and Verej Janoyan of the LA Department of Transportation. The tireless Dr Scott Smith of the Volpe Center was instrumental in getting the outputs from this thesis out into the real world.

CEE has a long list of able administrators who have cheerfully and pro-actively attended to many a potential issue before they arose. I would especially like to thank Leanne Russell, Anthee Travers, Cynthia Stewart, Donna Hudson, Pat Dixon, Pat Glidden, Ginny Siggia and Sara Goplin for their constant assistance that resulted in a smooth run through grad school.

Lab-mates come and go, but the memories will live on forever. I shared an office and many cherished moments with Dr Constantinos Antoniou, who continues to be

my fountain of knowledge on a wide range of transportation and IT topics. Together, we proved the sufficiency of plain, vanilla e-mail for high-volume, real-time communications across continents. Bhanu Prasad Mahanti, Ashish Gupta, Anita Rao, Gunwoo Lee, Akhil Chauhan, Charisma Choudhury, Vaibhav Rathi, Vikrant Vaze, Varun Ramanujam, Maya Abou Zeid, Emmanuel Abbe, Caspar Chorus, Carmine Gioia and Gianluca Antonini have all made the ITS lab a vibrant and social environment.

Room-mates Prahladh Harsha, Jeff Hwang and Rajappa Tadepalli provided something to look forward to upon returning home from the lab. At least until I got married and moved out! Friends made at MIT and before have watched out for me in various ways. I thank Arvind Sankar, Lakshmi Iyer, K. V. S. Vinay, Bharath Krishnan, Rajappa Tadepalli and Padmashree Ramachandran, Aravind Srinivasan and Karunya Ramasamy, Srini Sundaram, Anand Sivaraman and Chaitanya Ullal for their help and companionship.

I thank Vikrant Agnihotri, Vikram Sivakumar and Varun Ramanujam for their diligent organization of cricket games at MIT, and everyone who showed up so that we not only had quorum, but also an enthusiastic knock-about.

I am grateful to Radha Kalluri, Ray Goldsworthy, Aarthi Chandrasekharan, Srinath Gaddam, Charu Varadharajan and everybody else at (and associated with) MIT Natya for providing a cultural edge to my MIT stay (not to mention the motivation to pick up my violin a few times). I thank Kripa Varanasi, Prahladh Harsha and Ganesh Davuluri for volunteering as guinea pigs in my experiments as a violin teacher.

I cannot imagine how I would have finished this project without the whole-hearted support of my family: my wife Krithika, parents and grandparents and sister Poornima have stood by me throughout.

And finally, my thanks to God for the strength to survive and experience this great journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

As the world rapidly becomes a global society and economy, the need to physically move goods and people from one place to another has never been greater. Individuals and families make more trips today for a variety of reasons: they commute to work, perform household tasks and travel to participate in recreational activities. The simultaneous boom in manufacturing and retail services has led to a surge in the shipment of raw materials and finished goods across and between entire continents. Of these two components, the transportation of people is of particular interest, since the entities interact directly with their environment, make their own decisions and prefer to operate within individual-specific environmental parameters.

Road (highway) traffic systems involve perhaps the most complex set of interactions related to transportation. Individuals in such systems need to be in constant control of their vehicles. They also make continuous decisions relating to route and lane choice, speed, acceleration and deceleration, overtaking, merging and response to information and control messages. Driver behavior under a variety of traffic conditions (such as congestion, delays and accidents) and personal circumstances (the need to keep an appointment, for example) add yet another dimension that can often perturb traffic flow and increase stress. Transportation systems must therefore be planned, operated and managed with care in order to ensure smooth flow, taking into consideration expected demand, stochasticity and potential disruptions.

Network modeling has historically played an important role in analyzing the costs and benefits of important transportation infrastructure proposals. The ever-expanding list of stakeholders and the growing awareness of diverse socio-political, environmental and quality-of-life issues have made transportation network planning, design and operations a complex yet necessary process. The existence of complicated interactions between the various parties involved, together with the irreversibility of infrastructure investment, drives the need to carefully weigh the targeted system benefits against all potential undesirable long- and short-term consequences.

Transportation has a strong influence on land use. Decisions by traffic planners clearly have long-term impacts on network performance, land use evolution and regional urban development. For example, the construction of a high level-of-service

highway or transit link has the potential to draw businesses and residents towards the new transportation corridor. Industrial growth and induced vehicular traffic could also lead to congestion and environmental effects that may play major roles in households' decisions to own automobiles or change home location. In addition, the physical alignment of the proposed facility may itself be the focus of intense political and environmental concern that can delay urban development and cause financial over-runs. Accurate analysis methodologies are therefore essential to ensuring the long-term sustainability of urban growth while also providing short-term benefits and opportunities to the local communities.

Constraints on the availability of land and financial resources along with strong opposition to the addition of noisy facilities close to residential areas have forced a rethinking of urban mobility planning. Transportation systems have recently been the subject of a paradigm shift away from building new capacity, towards the enhancement, better management and utilization of existing infrastructure. Enhancements to increase network capacity include lane expansions, signal and ramp metering optimization and ramp re-design. It is also believed that more efficient demand and incident management through the deployment of Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) may mitigate congestion. The success of such measures relies heavily on planners' abilities to accurately model a wide range of ATMS and ATIS technologies, evaluate network performance, and capture driver behavior (particularly their response to improved information). The modeling capabilities to support such decisions at both the planning and operational levels target short-term and within-day effects, given long-term decisions. Critically, these models must capture *dynamic* traffic evolution in order to replicate the formation and dissipation of queues and spillback (under both recurrent congestion and during unexpected perturbations such as incidents).

Static approaches based on the traditional four-step modeling approach predict trip rates (the trip generation step) over large time horizons potentially spanning several years. They also forecast the corresponding origin-destination (OD) demands and the level of use of various transportation modes, through the trip distribution and

modal split steps. Finally, these models estimate congestion levels and link volumes. This final step, generally referred to as traffic assignment, assumes that network flows and travel times remain unchanged over the entire study period such as the morning or evening peak period. Steady-state OD flows are then loaded onto the network, yielding link flows that are expected to capture average conditions across the entire period. The simplest network loading technique is all-or-nothing assignment through which the entire flow between any OD pair is assigned to the path with the minimum travel time (or generalized cost). Capacitated variations of this shortest-path approach have been tried, in order to capture congestion effects. Other popular methods include User Equilibrium (UE), Stochastic User Equilibrium (SUE) and System Optimal (SO). A UE assignment is based on the hypothesis that drivers, and other users of the transportation system, evaluate and maximize their perceived utilities across all feasible (or reasonable) routes. Under this assumption, no driver can reduce her travel cost by switching to another route from her choice set. Deterministic network link costs and drivers' perfect perception of the same are also assumed. SUE approaches introduce probabilistic route choice models that recognize stochasticity in drivers' perceptions of route costs. Finally, the SO approach minimizes the total system travel cost across all drivers. Such a situation, even if it exists, is not an equilibrium, since some drivers can potentially benefit by switching to another route. Further, SO assignment is not reflective of expected driver behavior, since it requires communication and cooperation between all drivers.

Static analyses (such as the four step process) are well-suited for long-range planning purposes such as major infrastructure investments, land use planning (including industrial and residential zoning), and airport and facility location. A particular advantage of these methods is their ability to project OD demand and mode utilization based on current data. However, the resolution along the time dimension is too coarse to allow the modeling of within-day and day-to-day effects. For example, a static model may yield daily average vehicular flows on the links comprising the study network, but cannot capture within-day dynamic demand profiles, the formation and dissipation of queues, and network performance under incidents or other real-time

perturbations.

The advent of simulation techniques and powerful, inexpensive computers have seen a gradual shift away from static traffic assignment approaches, towards less tractable yet more realistic models that capture *dynamic* demand patterns, incorporate stochastic driver behavior, model traffic dynamics, and explicitly replicate demand-supply interactions. Dynamic models also allow the modeling of drivers' *en-route* decisions, which include response to traveler information disseminated through variable message signs or other means. Transportation analysts are increasingly adopting such complex modeling and simulation methods to design transportation infrastructure and optimize its operations. The rapid transition of the state-of-the-art of dynamic transportation network modeling to the state of the practice in recent times has been possible due to three primary factors:

- Modern traffic network sensing technologies such as pavement loop detectors, automatic vehicle identification (AVI) and remote traffic microwave systems (RTMS) are yielding richer, more up-to-date and easily collected traffic data that provide opportunities to estimate more realistic models of traffic and driver behavior.

- Research in modeling and simulation techniques have resulted in models that can replicate network demand and supply processes, driver behavior mechanisms, and their complex interactions. Further, these advanced models have been validated through real sensor data that is becoming more widespread with large-scale traffic surveillance deployment.

- Rapid progress in the capabilities of modern computers are playing a critical role in thrusting ITS prototypes onto the practical realm. Faster computers are demonstrating the advantages of data- and processor-intensive simulation models for both off-line planning analyses as well as real-time system optimization and operations.

Next we present the structure of a generic Dynamic Traffic Assignment (DTA) model, followed by an overview of the evolution of dynamic traffic assignment ap-

proaches. This discussion aims to convey the growing model complexity during the development of DTA theory, and motivates the need for calibrating such complex systems before they are applied to real-life scenarios.

1.1 Generic structure of DTA models

DTA models replicate various traffic phenomena through complex demand and supply model components that interact systematically to simulate the performance of the network. The structure of a generic DTA model is outlined in Figure 1-1. Demand models estimate and predict origin-destination (OD) trip patterns, and simulate the behavior of individual drivers (including pre-trip departure time, mode and route choice, and response to information). Supply models capture traffic phenomena through detailed representations of the capacities of network elements, the traffic dynamics resulting from speed/acceleration, lane changing and merging/weaving behavior, and the impact of incidents. Various algorithms tie the demand and supply components together to assign the dynamic demand to the network and determine the temporal propagation of flows. The resulting traffic conditions (including speeds, densities, travel times and delays) may be used for a variety of planning and real-time management applications.

DTA models differ in the mechanisms used to capture the time-varying nature of demand and supply processes and their interactions. A discussion of various DTA approaches follows.

1.2 Typology of DTA Models

While the mathematical properties and solution approaches of the static assignment problem are well understood, the associated modeling limitations yield unrealistic traffic characterizations that fail to capture driver behavior such as response to en-route information, and fundamental congestion phenomena such as queuing dynamics. The importance of modeling traffic in a dynamic setting has been stressed already,

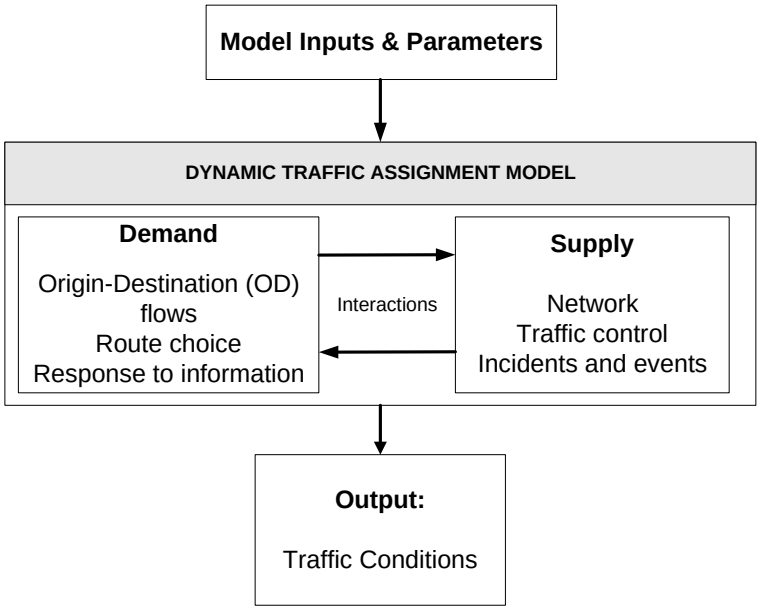

Figure 1-1: Structure of Generic DTA Model

particularly in the context of real-time incident management and en-route guidance. In this section, we review the literature on dynamic traffic assignment methods. The following discussion is intended to highlight the complexity of current DTA models, and the intermediate approaches leading to the state-of-the-art.

DTA model types can be broadly classified according to the nature of their modeling and solution frameworks. We divide the literature under three headings. The first describes some of the initial efforts to replicate observed dynamic congestion features using known static concepts. Next, two very different mainstream approaches, truly dynamic in their treatment of traffic, are outlined. One relies on analytical formulations (largely through optimization) and solution algorithms, while the other employs simulation techniques. We summarize with a note on the complexity of simulation-based DTA systems, leading to the motivation for this research.

1.2.1 Early developments

The development of computer programs capable of solving the static traffic assignment problem may have provided the impetus for the first non-static approaches to

modeling traffic phenomena. Peeta (1994) describes a few such efforts, classified as "quasi-dynamic" assignment. These approaches attempted to introduce dynamic considerations through the repeated application of static methods to sub-intervals of the period of interest. One such study (Halati and Boyce (1991), Halati et al. (1991)) focuses on evaluating the impact of route guidance during incidents. The authors divide the time horizon into "pre-incident", "incident" and "post-incident" regimes, and solve a static user equilibrium formulation to obtain steady-state conditions leading up to the incident. The subsequent analysis, targeting more dynamic treatment of congestion evolution, employs successive static user equilibrium problems over ten-minute intervals.

Quasi-dynamic methods were also the basis for some early simulation-based assignment models. An initial version of CONTRAM (CONtinuous TRaffic Assignment Model, Leonard et al. (1978), Leonard et al. (1989)) from the UK-based Transportation and Road Research Laboratory, captured within-day demand dynamics through the specification of different OD demand rates in each sub-interval. The model, however, allowed only one path per vehicle packet[1]. Also, the use of relatively long time sub-intervals, coupled with a network equilibrium objective, rendered it unsuitable for short-term planning and real-time/ATIS applications. The current CONTRAM software release addresses many of these limitations, and is discussed in Taylor (2003).

SATURN (Simulation and Assignment of Traffic to Urban Road Networks), developed at the University of Leeds (UK), is another quasi-dynamic system with features and properties similar to the first CONTRAM model. Details of this model are provided in Hall et al. (1980) and Vliet (1982). It improves over CONTRAM in terms of its detailed treatment of urban intersections and traffic signals.

In summary, quasi-dynamic approaches represent the earliest research that moves away from traditional static traffic modeling. While they began to harness the growing capabilities of computers and simulation, their applicability and modeling accuracy were still severely limited:

[1]Vehicles between an OD pair were aggregated into homogeneous groups called packets in order to reduce computational requirements.

- Time intervals were too long to replicate fast-evolving traffic conditions.

- Static assignment assumes the instantaneous propagation of the entire OD demand within each interval. The reliance on a series of static assignments thus ignored the impact of vehicle interactions on capacity and traffic dynamics.

- Solving a successive set of *independent* static assignments did not guarantee flow conservation and continuity across time interval boundaries.

The above limitations led to research into truly dynamic assignment techniques that could, directly or indirectly, capture vehicle interactions and traffic dynamics.

1.2.2 Analytical approaches

Several researchers worked on extending the analytical static assignment problem to capture some prominent features of time-varying traffic conditions. Analytical DTA models approximate the DTA problem for a specific objective (such as User Equilibrium (UE) or System Optimal (SO)), through mathematical formulations and explicit constraints. The corresponding solutions of interest are usually obtained by the application of a traditional (typically non-linear) optimization algorithm that solves for the unknown variables (network descriptors).

Analytical DTA models can be classified according to the basic assumptions underlying their formulations. We present here an outline of three major formulation ideas, focusing on the crucial limitations that motivate simulation-based DTA approaches (such as those described in Section 1.2.3). The three classes rely on concepts drawn respectively from the areas of mathematical programming, optimal control and variational inequalities. A more detailed discussion of these modeling concepts can be found in the review paper by Peeta and Ziliaskopoulos (2001) and the references contained therein. We present here a brief synopsis of the three approaches.

Mathematical programming formulations

DTA models based on mathematical programs employ a time-discretization scheme to model traffic dynamics. The basic formulation, after being adjusted and modified repeatedly, possesses desirable mathematical properties such as convexity and well-behaved functional forms. In particular, convexity allows for solution approaches that exploit standard non-linear programming packages.

The mathematical properties of academic problems consisting of a single OD pair have been thoroughly analyzed in the literature (Merchant and Nemhauser, 1978; Carey, 1987). However, extensions to the realistic scenario of multiple OD pairs has highlighted the crucial limitations of the mathematical programming approach. For instance, vehicles may "jump" over other vehicles, if such decisions result in a lower objective function value. A less serious effect concerns overtaking maneuvers, a regularly observed feature in real traffic, that is precluded through explicit First-In-First-Out (FIFO) requirements. Yet another limitation concerns the problem formulation with a System Optimal objective, with the solution algorithm potentially subjecting a subset of vehicles to unreasonable delays at certain nodes in order to reduce the travel times of other traffic streams (and consequently lowering system-wide costs). Both these issues are discussed in detail by Carey and Subrahmanian (2000).

In general, in spite of other extensions that aim to increase the realism of mathematical DTA approaches (Janson, 1991; Birge and Ho, 1993; Ziliaskopoulos, 2000), mathematical programming formulations lack the ability to accurately replicate real-world congestion patterns and driver behavior. Where more realistic approaches are attempted by incorporating additional constraints, the resulting problems are too complex to be applied to large-scale networks in real-time applications.

Optimal control formulations

Formulations based on optimal control theory differ from mathematical programming approaches in their representation of the temporal dimension. While the previous class of models sliced time into discrete intervals, optimal control formulations treat

OD demand and link flows as continuous functions of time (Friesz et al., 1989; Wie, 1991). The resulting approaches share many of the properties and limitations of mathematical programming models. The generalization of Wardrop's UE principle has been particularly difficult, when more than one OD pair is considered.

Several papers discuss methodologies and relaxations that may help enhance the realism of optimal control approaches (Ran and Shimazaki, 1989; Ran et al., 1993; Boyce et al., 1995). However, these efforts remain preliminary, with few practical examples or data to support their applicability. The lack of efficient solution algorithms has also severely hindered the progress of this class of DTA models.

Variational inequality formulations

Applications of variational inequalities (VI) to the traffic assignment problem have been well-documented for the static case. The underlying concepts were recently transferred to the dynamic case with some success, resulting in more general models. Further, mathematical analyses of VI problems have helped identify limitations of other analytical methods when dealing with asymmetric link costs. However, basic limitations of analytical approaches (such as lack of realistic representation of congestion and driver behavior) persist in this class of models.

Friesz et al. (1993) present a path-based DTA formulation which is one of the few analytical approaches to include driver behavior. The authors approximate drivers' route and departure time choices by utilizing link performance functions together with desired arrival times and early/late arrival penalties while computing path costs. However, the resulting system of simultaneous integral equations cannot be solved efficiently using existing algorithms.

Subsequent work using a link-based approach has shown improved traffic realism at the expense of computational overhead (Ran and Boyce, 1996; Chen and Hsueh, 1998). Path-based approaches are however better suited to route guidance and ATIS situations, since driver behavior realistically includes perceptions of entire paths or sub-paths (sequences of consecutive links) rather than individual links. VI formulations thus exhibit more flexibility in capturing real traffic phenomena, but their

solution remains prohibitively expensive even for moderately-sized networks.

Conclusions

Analytical DTA approaches have attempted rigorous mathematical formulations of ever-increasing complexity, in a bid to close the gap between the models' capabilities and observed reality. However, such efforts, while meeting with limited success on small networks with simplified behavioral assumptions, largely fail to capture the truly dynamic characteristics revealed in the real world. As the focus of traffic planning and operations shifts towards demand management and real-time route guidance, there is a need for DTA models capable of capturing the full complexity of individual drivers' decisions relating to route and departure time choice and response to en-route information and control messages. The capabilities of such DTA systems must go beyond traffic assignment, to estimating and predicting OD flows, travel times, delays and queues. Such detailed modeling abilities lie in the realm of simulation. Section 1.2.3 reviews the cutting edge of simulation-based dynamic traffic assignment, which is the focus of this thesis.

1.2.3 Simulation-based approaches

Traffic simulation models may be classified based on their level of abstraction of drivers and driving behavior. *Microscopic* models represent individual drivers, their decisions and interactions at a high level of detail. Interactions may include car following, lane changing, merging and yielding maneuvers that indirectly determine network capacity and traffic dynamics. *Macroscopic* models treat traffic as a uniform or homogeneous flow, and adapt physical concepts (such as fluid dynamics) to approximate their propagation through the network. Such approaches are unable to capture behavioral elements such as route and departure time choice, response to information or drivers' interactions with adjacent vehicles. *Mesoscopic* models combine some elements from both microscopic and macroscopic approaches, representing individual drivers and their travel decisions but replacing vehicle interactions with macroscopic

traffic relationships. These relationships typically reflect the inter-relationships between flows, speeds and densities on the various links of the network.

Given the interest to model dynamic traffic for both planning and real-time applications, numerous simulation models have been developed to date. Some early microscopic simulators include NETSIM, TRAF-NETSIM, INTRAS and FRESIM. Macroscopic tools have also been widely used for network modeling and signal timing optimization, including TRANSYT, FREQ, FREFLO, KRONOS and CORQ. INTEGRATION, along with later versions of CONTRAM and SATURN have been classified as mesoscopic systems in the literature.

While the list of available simulation tools is large, the tools themselves are often tailored for a specific type of application, such as a freeway corridor or an urban intersection. Further, the literature indicates many limitations (relating to critical aspects such as the replication of congestion patterns, driver behavior in weaving segments, and response to information) on the applicability of these early models. The handling of alternative paths between OD pairs, and the modeling of drivers' perceptions of the same, is another area that has received attention only recently. We now review some of the more sophisticated DTA models in use today, to provide a flavor of their complexity and realism.

Current DTA models

Microscopic models are widely employed today both by the research community and transportation professionals. Such models capture traffic dynamics through detailed representations of individual drivers and vehicular interactions. Ahmed (1999) presents a comprehensive discussion on microscopic traffic model components. Popular commercial microscopic software packages include CORSIM (FHWA, 2005), PARAMICS (Smith et al., 1995), AIMSUN2 (Barcelo and Casas, 2002), MITSIMLab (Yang and Koutsopoulos, 1996; Yang et al., 2000), VISSIM (PTV, 2006) and TransModeler (Caliper, 2006). Such tools have been applied in a wide range of planning and design contexts (Abdulhai et al., 1999; Mcdougall and Millar, 2001).

Macroscopic models achieve fast running times on very large networks, at the ex-

pense of individual driver behavior modeling. Flows are typically treated as fluids, and the speed of flow is captured through a macroscopic function such as a speed-density or speed-flow relationship. While such models may be used for long-range planning applications, the lack of behavioral detail (such as route choice) is a limitation in applications that involve driver response to information (such as the impact of variable message signs (VMS) and the evaluation of ATIS strategies). Several macroscopic models are reported in the literature, including METANET (Messmer and Papageorgiou, 2001), EMME/2 (INRO, 2006), VISUM (PTV, 2006) and the cell transmission model (CTM, Daganzo (1994)).

Mesoscopic models target the estimation and prediction of traffic conditions in real-time. Such systems are syntheses of microscopic and macroscopic modeling concepts, coupling the detailed behavior of individual drivers' route choice behaviors with more macroscopic models of traffic dynamics. Mesoscopic models have significantly faster run times than microscopic models, yet capture the individual decision processes that are required for evaluating drivers' response to information. Such models are therefore suitable for real-time or on-line applications such as incident management and route guidance generation. Examples of such systems include Dynamic Network Assignment for the Management of Information to Travelers (DynaMIT, Ben-Akiva et al. (2001, 2002)) and DYnamic Network Assignment-Simulation Model for Advanced Road Telematics (DYNASMART, Mahmassani (2002); UMD (2005)).

DynaMIT integrates detailed demand and supply simulators to estimate and predict network state (including flows, speeds, densities and queue lengths). The demand simulator models network demand at two resolutions: disaggregate drivers' route and departure time choices are simulated using sophisticated discrete choice models, while OD demand is modeled as aggregate flows. DynaMIT-R, developed for real-time applications, synthesizes estimates of current network conditions from historical information along with real-time surveillance data. OD predictions (based on current network state) are then assigned using a mesoscopic supply simulator to assess network performance in the near future. DynaMIT-R is flexible enough to allow the simulation of a wide range of ITS strategies, including variable message signs,

on-board traveler information devices, and HOV lanes.

DYNASMART-X performs traffic routing functions similar to the real-time DynaMIT system. It may operate in different modes, including predictive, decentralized reactive (when local network controllers route vehicles by reacting to events such as incidents), and hybrid (a combination of the centralized and decentralized approaches).

While sophisticated microscopic models have been applied to large, integrated, urban networks, the associated computational requirements limit their use to short- and medium-range planning. The repeated use of microscopic simulations for planning entire cities or regions may involve extremely costly computer runs, though these applications may not be required to run faster than real-time. Large run times can also occur in highly congested situations. DTA systems have therefore been developed for short-term planning applications on large networks. The DynaMIT-P system, for example, draws on detailed demand and supply simulators to estimate dynamic OD flows from link counts, simulate drivers' day-to-day travel time learning behavior and predict the impact of a variety of information provision strategies during work zones, special events and other pre-planned scenarios. The system can further simulate multiple user classes and HOV lane use. DYNASMART-P is a similar variant of the real-time DYNASMART-X system.

Dynameq (Dynamic Equilibrium), recently developed by INRO, is a commercial network planning tool equipped to perform iterative simulations towards a dynamic user equilibrium solution. Dynameq employs innovative algorithms to achieve significant run-time savings when compared to most existing microscopic models, while retaining some of the details such as car-following (Mahut et al., 2005).

1.2.4 Synthesis

State-of-the-art DTA models have been developed in the past decade, for a variety of traffic network design, planning and operations management situations. These models employ sophisticated algorithms and detailed microscopic, macroscopic and mesoscopic simulation techniques to estimate network performance, predict (short-term) future conditions and generate route guidance. Analytical approaches to route

guidance generation are also being explored. Such advanced systems are being actively pursued today in the context of ATMS, APTS and ATIS, with the on-line deployment of real-time, predictive guidance systems a distinct possibility in the next few years. DTA models are also being increasingly applied for short-term planning purposes with the aim of including dynamic congestion evolution in the analysis of network performance.

1.3 Motivation and scope

The value of DTA models (particularly large-scale simulation systems) depends on their ability to accurately replicate conditions for the specific network being studied. Indeed, the true impact of a new signal timing plan, for example, may be assessed only if simulations of current (base-case) traffic control measures and drivers' reaction and response to the associated control messages are realistic. While advanced DTA models provide realistic abstractions of actual demand and supply processes, their outputs are governed by a large set of inputs and parameters that must be estimated before the models are applied. Well-calibrated models are therefore critical to the success of any DTA application.

The goal of DTA model calibration is to obtain accurate depictions of the following aspects of a region's transportation and traffic patterns:

Travel demand: Time-varying matrices of OD demand are important inputs to DTA models. They capture local trip rates and travel patterns. OD demand can potentially vary according to changing activity patterns of the travelers, which may vary by day of the week, season, weather conditions, major work zones and special events. DTA models also capture a variety of drivers' travel behavior, such as route choices and response to information. DTA models rely on time-varying OD profiles and route choice models to capture demand-side effects.

Network supply: The capacities of network elements such as links and intersec-

tions (signalized and unsignalized) are determined by a host of factors related to both the network and the drivers. The number of lanes on a freeway section, for example, imposes constraints on the section's throughput. Traffic signals perform a similar function at arterial intersections, by allocating available capacity among competing streams of traffic. Complex vehicle interactions (such as passing, lane changing, merging, yielding and weaving) further influence evolving traffic dynamics and indirectly impact capacities. Incidents can also potentially impact the smooth flow of vehicles, especially when the network is operating with high traffic volumes. DTA models replicate the network's supply phenomena through detailed representations of capacities and traffic dynamics.

Demand-supply interactions Traffic patterns realized on the network are the result of complex interactions between travel demand and network supply. DTA models employ detailed algorithms to capture these interactions and ensure accurate estimates of queues, spillbacks and delays.

The topology of the traffic network, a critical input to all DTA models, will be treated as an exogenous input in this research. It is assumed that a node-link representation of the network at a resolution suitable to the proposed modeling task and chosen DTA model is available from sources such as GIS (geographic information systems) databases, off-line and on-line maps, satellite images, aerial photographs and prior network studies.

DTA models involve a large number of parameters and inputs that must be calibrated with actual traffic data to accurately predict traffic conditions. *Off-line* calibration typically results in the creation of a historical database that ensures the model's ability to replicate average conditions potentially covering a wide range of factors such as day of the week, month, season, weather conditions and special events. Such a calibration is expected to perform satisfactorily in planning studies, including the evaluation of alternative network configurations and traffic management strategies.

On-line DTA applications require accurate real-time predictions of traffic conditions on a given day. Traffic conditions are impacted by factors such as weather,

road surface conditions, traffic composition and incidents. The results of the off-line calibration must therefore be adjusted in real-time to be sensitive to the variability of traffic conditions from their average values. On-line calibration is performed using real-time surveillance data. The results of the off-line calibration are used as *a priori* estimates during the on-line calibration process. Figure 1-2 illustrates an integrated framework that captures the relationship between off-line and on-line calibration in the context of DTA model applications. The typical data and parameters involved in each step are also indicated. This thesis focuses on the off-line calibration of DTA models.

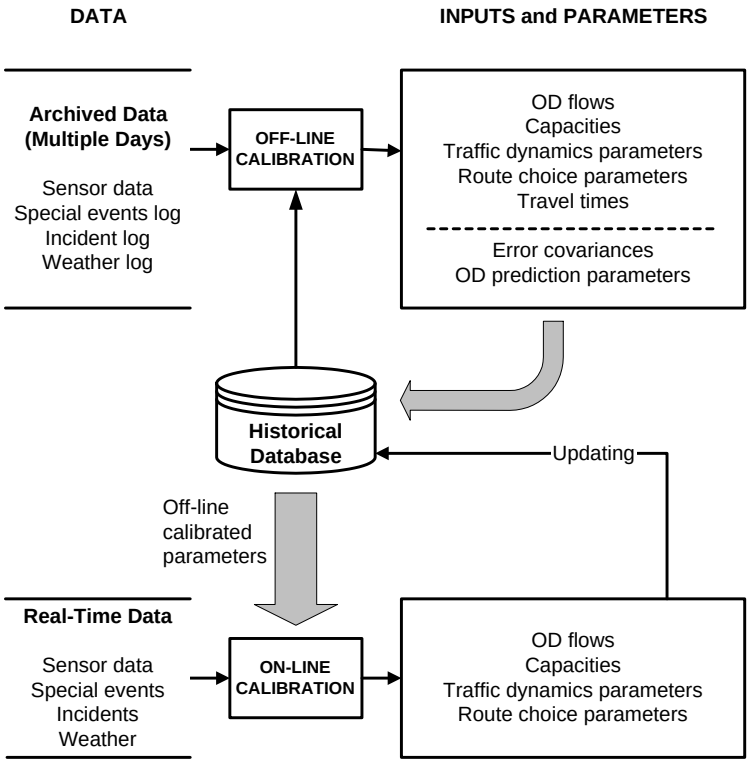

Figure 1-2: Off-line and On-line Model Calibration

Typically, models and the data needed to calibrate them fall in the same category. Route choice models are thus traditionally calibrated using disaggregate survey data. Similarly, modern OD matrix estimation methods rely on aggregate link sensor count observations[2]. Given the paucity of disaggregate (survey) datasets, however, the

[2]An exception to this is the use of disaggregate OD surveys to generate OD matrices. This

analyst will often be faced with the task of calibrating complicated DTA models using aggregate data alone. Alternatively, models estimated using traditional methods may have to be updated using recent sensor measurements. This thesis provides a rigorous treatment of these problems, and demonstrates how the disaggregate and aggregate models within a DTA system may be calibrated jointly using aggregate data.

The off-line calibration of a DTA model is summarized in Figure 1-3. We have a DTA model with a list of unknown inputs and parameters (dynamic OD flows, route choice model parameters, capacities, speed-density relationships, etc). We must obtain estimates of these inputs and parameters by using the information contained in available aggregate, time-dependent traffic measurements, so that the DTA model's outputs accurately mirror the collected data. This data includes, but is not limited to, counts and speeds from standard pavement loop detectors[3]. Different sets of parameters may be estimated to reflect any systematic variability in traffic patterns identified from several days of observed data. *A priori* estimates for some or all of the parameters, if available, may serve as starting values to be updated with the latest data.

Automated data collection technologies afford the measurement and storage of large amounts of traffic data. This data is expected to span many days, representing the various factors (demand patterns, supply phenomena, incidents, weather conditions and special events) characteristic of the region. In order to apply the DTA model in the future, a database of model inputs and parameters must be calibrated for each combination of factors observed in the data. The calibration task must therefore begin with data analysis that reveals these combinations, and partitions the sensor measurements accordingly.

Once calibrated, an important practical consideration is the maintenance of the historical database as sensor data from future days become available. The historical estimates of model inputs and parameters must be updated with every new day of measurements. In keeping with this requirement, we focus on the development

method has largely been replaced in recent times by the approach based on link counts.

[3]A more detailed description of aggregate data sources is provided in Section 3.2.

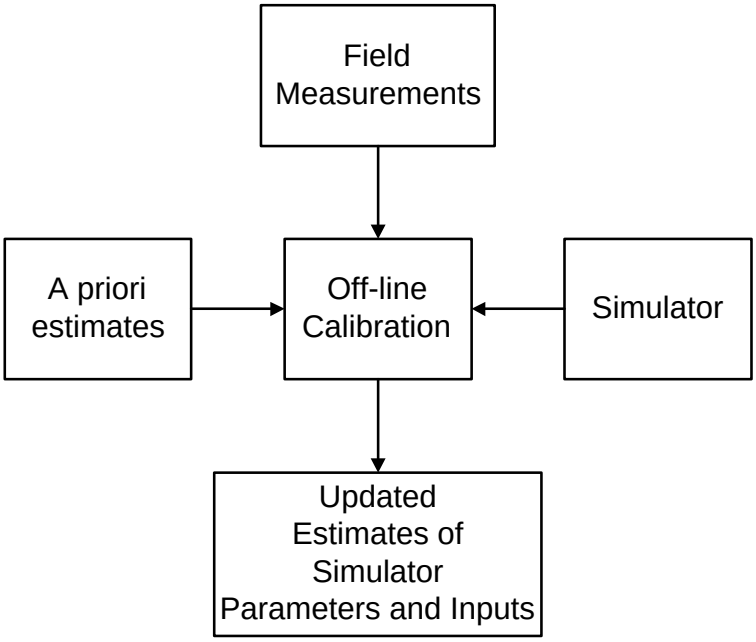

Figure 1-3: Calibration Framework

of a systematic calibration methodology that estimates all demand and supply variables using one day of data at a time. Methods to periodically update the database have been proposed elsewhere (see, for example, Ashok (1996) and Balakrishna et al. (2005a)), and are discussed in detail in Section 3.3.

1.4 Problem definition

Let the period of interest each day be denoted by $\mathcal{H}$. This period could include the entire day, or a specific portion of the day (such as the AM or PM peak). Let $\mathcal{H}$ be divided into H intervals of equal duration, represented by $h = \{1, 2, \ldots, H\}$. Let G denote the directed graph of nodes and links (and their associated characteristics, including records of major incidents and special events) corresponding to the physical transportation network input to the DTA model[4]. Let $\mathbf{x}$ represent the set of dynamic OD flows $\mathbf{x}_h$, $h \in \mathcal{H}$ prevalent on that day. Each n_{OD}-sized vector[5] $\mathbf{x}_h$ represents the OD flows departing from their origin nodes during interval h. Further, let $\boldsymbol{\beta}$ be the

[4]We refer here to a general DTA model, chosen to suit the application at hand.

[5]n_{OD} denotes the number of OD pairs on the network.

set of time-specific model parameters $\boldsymbol{\beta}_{\mathsf{h}}$ comprised of route choice model parameters and supply-side variables such as link/segment output capacities and speed-density function parameters. We denote the total set of unknown parameters $[\mathbf{x}\ \boldsymbol{\beta}]$ as $\boldsymbol{\theta}$. Let the travel time inputs to the route choice model be $\mathbf{TT}^{\mathrm{rc}}$.

The off-line DTA calibration problem can now be defined as the simultaneous estimation of all the demand and supply variables in $\boldsymbol{\theta}$, and a consistent set of route choice travel times, error covariances and OD prediction model parameters, using time-dependent counts and speeds recorded by traffic loop detectors. Traffic data $\mathbf{M}$ are assumed to be available over the H intervals in $\mathcal{H}$, so that

$$\mathbf{M} = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_{\mathsf{h}}, \ldots, \mathbf{M}_{\mathsf{H}}\}$$

The vector $\mathbf{M}_{\mathsf{h}}$ contains records of vehicle counts and speeds recorded during interval h.

A priori estimates $\mathbf{x}^{\mathsf{a}}$ and $\boldsymbol{\beta}^{\mathsf{a}}$, if available, can provide valuable structural information that must be exploited by the calibration methodology. The non-zero cells in $\mathbf{x}^{\mathsf{a}}$, for example, indicate the OD pairs that contribute to network flows (and hence must be included as optimization variables). Further, speed-density or speed-flow equations fitted to actual sensor data may serve as a good starting solution to be refined through systematic calibration. Information about the relative magnitudes of the various OD flows and model parameters may also be useful in speeding up the optimization through the use of appropriate lower and upper bounds.

The network G can vary from day to day. For example, a subset of links or lanes in the network might become unavailable for a few days due to severe incidents, weather conditions or scheduled maintenance activities. Such disruptions are treated as exogenous inputs to the calibration process. Details of planned special events that are expected to have a significant impact on the day's travel and traffic patterns are also included in G.

1.5 Thesis organization

The remainder of this thesis is organized as follows. Chapter 2 presents a detailed review of existing DTA model calibration approaches, and identifies the strengths and limitations of recent work in this area. Chapter 3 briefly describes typical DTA calibration variables and available sensor data, and outlines our formulation of the off-line DTA model calibration problem. Critical characteristics of the problem are analyzed, and an efficient calibration methodology is developed. Further, optimization algorithms are identified for solving the DTA calibration problem. The methodology is systematically tested in Chapter 4, using the DynaMIT DTA system, and results from a case study with a real dataset are presented and discussed in Chapter 5. Finally, we conclude with a synthesis of our major findings, contributions and directions for future research in Chapter 6.

Chapter 2

Literature review

Contents

Model calibration (or estimation) is the crucial step between the model development and application phases of any project or study. Indeed, the abundance of related literature underscores the importance of this problem. This chapter presents a review of recent model calibration efforts, while emphasizing their advantages and limitations. The chapter concludes by summarizing the current state-of-the-art, which will form the reference case for comparisons through case studies.

2.1 DTA calibration literature

Work relating to the rigorous and systematic *off-line* calibration of simulation-based DTA systems remains limited. The paucity of theoretical results on the calibration front is partly because DTA is a relatively new field, and much of the research effort thus far has focused on developing the theoretical foundations for various underlying model components and their interactions. The resulting systems are large-scale and complex, and employ detailed behavioral techniques and simulation approaches to achieve a high degree of congruity with real-life processes and phenomena.

The lack of significant quantities of real-world traffic data (typically caused by resource and technology constraints) has often limited the scope of DTA system calibration studies to date. The dynamic nature of DTA systems' model inputs and parameters requires the support of *dynamic* data recorded over extended periods, which depends on advanced surveillance/sensor technologies. The demonstration of the applicability and functionalities of DTA systems has thus been limited to small-scale networks and short time periods, for which sufficient data can be collected.

The extensive deployment of ITS infrastructure in recent years has substantially increased the amount of time-dependent surveillance data from large, complex networks. More cities are installing pavement loop detectors and road side sensors capable of measuring traffic characteristics (such as vehicle counts and speeds) over time intervals as low as 30 seconds. In addition, communications technologies have enabled the transfer and storage of large traffic databases for future processing, thereby facilitating the study of DTA calibration and validation methods on real networks.

The literature pertaining to the calibration of DTA models can be grouped based on the constituent model component(s) whose inputs and parameters are under scrutiny. For this purpose, it is convenient to divide DTA model components into two categories: (a) the **demand** models that capture time-dependent origin-destination (OD) flows and traveler behavior (including mode, route and departure time choice), and (b) the **supply** models that mimic traffic dynamics and the phenomena of queue formation, dissipation and spillback. We organize the literature review into the following three broad classes:

- Demand-supply calibration of DTA models
- Estimation of supply models
- Estimation of demand models

While the calibration of the demand or supply models individually may be viewed as only a part of the overall problem of DTA system calibration, the experiences from such analyses provide valuable insights and suggest directions for the simultaneous, efficient and systematic estimation of all relevant DTA model parameters. We now focus on each of the three classes of estimation problems individually.

2.2 Demand-supply calibration of DTA models

A majority of the research on the calibration of a DTA system's demand and supply model components treats the various components as independent entities whose parameters are calibrated through a combination of prior experience, engineering judgment and manual adjustment. Chen et al. (2004), for example, present preliminary results from a case study applying the DYNASMART-P simulator in Zwolle in the Netherlands. The approach lacks a consistent model estimation framework, and relies almost entirely on manual adjustments to individual model components based on prior experiences with the network and its traffic patterns. Chu et al. (2004) present a similar calibration approach using PARAMICS in Irvine, California that combines heuristics and static approaches to assist in the fine-tuning of various model

parameters. Gomes et al. (2004) evaluate the fit to observed speeds through visual inspection, and iterate after manual adjustment until a satisfactory fit is obtained. There are many drawbacks to such approaches. First, the process is tedious and time-consuming, with no guarantee of improving upon the initial (or default) solution. Second, the complex inter-dependencies between model components are ignored, resulting in potentially biased parameter estimates. Finally, the large-scale nature of most traffic simulation applications would render any manual procedures infeasible.

Literature on the demand-supply calibration of DTA systems is limited. He et al. (1999) attempt to list the major sources of error in a DTA system, and lay out frameworks for the off-line and on-line calibration of the system. The proposed frameworks treat the calibration of the dynamic travel time, route choice, flow propagation and OD estimation models sequentially. The authors consider a modified Greenshields model to explain dynamic travel time variations on freeway links, and split the travel times on arterials into a cruise time component and a delay component[1]. The proposed calibration approach aims to minimize the "distance" between the analytically computed travel times and those measured by detectors. Further, the maximum likelihood estimation procedure suggested for the calibration of the route choice model relies heavily on the availability of adequate survey data about travelers' route choices. This assumption would fail in many real cases, where only aggregate network performance measures (such as link counts) are available. A procedure similar to that adopted for the dynamic travel time models is applied for the flow propagation model, where link inflows and outflows are matched against detector data. While such a detailed level of model calibration might be preferable, the lack of a sufficiently rich dataset at the link level would often render the approach infeasible. In addition, the approach does not include OD estimation, which constitutes a critical part of demand calibration.

In a subsequent paper, He and Ran (2000) suggest a calibration and validation approach that focuses on the route choice and flow propagation components of a DTA system. This paper again assumes prior knowledge of time-dependent OD matrices,

[1]Delays on arterial links are attributed to queuing at intersections.

and further simplifies the demand process by imposing temporal independence of OD flows between all OD pairs. The assumption of disaggregate data availability to allow a Maximum Likelihood Estimation of the route choice model is still a restriction on the practical applicability of the proposed approach.

The approaches reviewed thus far fail to address the overall problem of jointly calibrating the OD estimation, route choice and supply models. Hawas (2002) attempts to study the propagation of calibration errors through a DTA system, by assigning integer ranks to processes based on their external data requirements and internal information flows. The author recommends that processes with lower rank (and hence fewer internal interactions) be calibrated first, in order to minimize an overall error statistic computed with simulator outputs. However, there are serious questions about the applicability of the proposed approach to real networks with large sets of parameters. For example, the effort involved in individually perturbing each variable to study its impact on the model's output, as described in the paper, is likely to be great. The computational overhead will potentially increase further if model outputs from multiple replications must be averaged to account for simulator stochasticity (a point not covered in the paper). Moreover, the magnitude of the perturbation may be hypothesized to vary across the parameters, given the non-linear nature of the objective function. A uniform perturbation for all variables is therefore not optimal. Lastly, the case study on a small network, with known OD flows, simulated sensor data and just two calibrated parameters does not capture the complexity of a real-world calibration.

Mahut et al. (2004) describe the calibration of a mesoscopic traffic simulator in Alberta, Canada. Their software combines a microscopic network loading model (with gap acceptance parameters, for example, that are closer to lane-changing models) and a mesoscopic routing engine that employs volume-delay functions. Iterations between the two components are used to establish dynamic equilibrium travel times on the network.

The described calibration approach is again based on heuristics, and the different model components are treated independently. Hourly OD flows are estimated

by matching turning movement counts at major intersections with their simulated values. It should be noted that such detailed counts, collected for this case study through an extensive survey, are rarely available. Capacities are estimated based on saturation flow rates and empirically derived downstream intersection capacities (the method for obtaining downstream capacities is not explained). Triangular volume-delay functions approximated based on posted speed limits and estimated capacities may not accurately reflect ground conditions, since drivers on average travel at speeds higher than the speed limit under uncongested conditions[2]. The gap acceptance and route choice parameters were adjusted manually to minimize the objective function. While valuable insights are provided to support these adjustments, they remain case-specific, and may not generally be transferable.

Mahmassani et al. (2004) report on the calibration of DYNASMART-X for a real-time application on the Irvine network. Modified Greenshields speed-density functions were calibrated using linear regression, after applying a logarithmic transformation. Dynamic OD flows were estimated using a least squares approach similar to that of Balakrishna et al. (2005a) (discussed in detail in Section 2.4.2), the required assignment being generated by the DTA model. The report, however, does not mention the calibration method for capacities and the route choice model.

Gupta (2005) demonstrates the calibration of the DynaMIT mesoscopic model by combining the demand approach of Balakrishna et al. (2005a) with the supply estimation approach of Kunde (2002) (see Section 2.3.1). In this thesis, supply and demand estimations are performed sequentially using real sensor count and speed data from the Los Angeles network. Although the application has many limitations (such as sequential OD estimation and local supply fitting) discussed in later sections, it contributes significantly through the development of an *observability* test that allows the modeler to ascertain if unique OD flows may be estimated from the given sensor configuration and coverage.

Several recent studies have focused on calibrating both demand and supply model

[2]This is particularly true on multi-lane freeway links, where speeds generally increase from one lane to the next.

inputs for micro-simulation. Mahanti (2004) calibrates the demand and select supply parameters for the MITSIMLab microscopic simulator by formulating the overall optimization problem in a Generalized Least Squares (GLS) framework. The approach divides the parameter set into two groups: the OD flows (which may be estimated efficiently using existing tools) and the remaining parameters (including a route choice coefficient, an acceleration/deceleration constant in the car-following model, and the mean and variance of the distribution of drivers' desired speeds relative to the speed limit). An iterative solution method is implemented, with the OD flows estimated using the classical GLS estimator, and the parameters estimated by Box-Complex iterations (see Box (1965) for a detailed outline of the solution algorithm).

Toledo et al. (2004) formulate the problem of jointly calibrating the OD flows, travel behavior and driving behavior components of microscopic models, using aggregate sensor data sources. The OD estimation step (utilizing a GLS formulation) is introduced as an explicit constraint, and a bi-level heuristic solution algorithm is used to solve for the three components iteratively. The use of the Box-Complex algorithm is reported for the estimation of select behavioral parameters. A more general formulation incorporating drivers' day-to-day travel time learning mechanisms was presented by Balakrishna et al. (2004).

Jha et al. (2004) calibrate MITSIMLab for a large-scale network in Des Moines, IA comprised of 20,953 OD pairs for four 15-minute intervals. They estimate driving behavior parameters independently on a single freeway section for which OD flows could be inferred easily (and without error) from sensor counts. Subsequently, OD flows, a route choice parameter and habitual travel times were obtained by iteratively calibrating each component individually until convergence. The authors discuss several practical issues relevant to large-scale model calibration, the most important being the effect of stochasticity and extremely low OD flows on the quality of the simulated assignment matrices used for the GLS-based OD estimation.

To summarize, the calibration of a DTA model's demand and supply components has generally been attempted through a sequential procedure that first estimates supply parameters (assuming known demand inputs), then estimates demand parameters

with fixed supply. The complex interactions between demand and supply are thus ignored. The remainder of this chapter focuses on methods that have been employed to individually calibrate the supply and demand parameters.

2.3 Estimation of supply models

Network supply is modeled in a variety of ways in DTA models. Macroscopic traffic models capture traffic dynamics through aggregate relationships derived by approximating vehicular flow as a fluid. Mesoscopic models use speed-density or link performance functions that are based on traffic variables such as flows, speeds and densities. In either case, a potentially large set of parameters needs to be calibrated for the DTA system to replicate field measurements.

Recent studies have employed systematic algorithms for the calibration of DTA supply models, with varying degrees of success. In this section we review the experience with sophisticated optimization algorithms applied to macroscopic, mesoscopic and microscopic supply model calibration.

2.3.1 Macroscopic and mesoscopic supply calibration

Supply calibration in the macroscopic and mesoscopic contexts generally involves the estimation of capacities and link performance functions. The typical data used for this task are sensor records of at least two of the three primary traffic descriptors: speeds, flows (or counts) and densities (or detector occupancies).

Leclercq (2005) estimates four parameters of a two-part flow-density function with data from arterial segments in Toulouse, France. An interior point, conjugate gradient method is employed to optimize the fit to observed sensor flows, with the fitted flows obtained from an aggregate relationship comprised of a parabolic "free-flow" part and a linear congested regime. Van Aerde and Rakha (1995) describe the calibration of speed-flow profiles by fitting data from loop detectors on I-4 near Orlando, Florida. Realistic sensor coverage levels, however, require that the links or segments on large networks be grouped based on the traffic characteristics observed at sensor locations,

and separate functions estimated for each group. Similar approaches have been widely applied on networks of realistic size and structure.

A major drawback of the above approach is one of localized fit. The estimated link performance functions reflect spot measurements at discrete sensor stations, and do not necessarily correspond to overall link dynamics (especially in the presence of congestion). The estimation procedure does not enforce consistency across contiguous links or segments, stressing the need for an expanded approach that considers larger sections of the network.

Most calibration approaches focus on the independent estimation of subsets of supply parameters. Muñoz et al. (2004) describe a calibration methodology for a modified cell transmission model (MCTM), applied to a 14-mile westbound stretch of the I-210 freeway. Free-flow speeds are obtained through least squares, by fitting a speed-flow plot through each detector's data. Free flow speeds for cells without detectors are computed by interpolating between the available speed estimates. In the case of bad or missing sensor data, a default of 60 mph was assumed. For the purpose of capacity estimation, the freeway is divided into congested and free-flow sections by studying speed and density contours from detector data. Capacities in the free-flow cells are set to be slightly higher than the maximum flow observed at the nearest detector. Bottleneck capacities are estimated to match the observed mainline and ramp flows just upstream of the free-flow part of the bottleneck. Speed-flow functions are obtained through constrained least squares on sensor data from congested cells. Demands are calculated from complete knowledge of all mainline and ramp flows.

Yue and Yu (2000) calibrate the EMME/2 and QRS II models for South Missouri City, a small suburban network outside the city of Houston, TX. While no systematic calibration approach is outlined, the authors "adjust" and "fine-tune" the free-flow travel times and turning fractions to match detector count data. Such ad-hoc procedures are unlikely to perform satisfactorily when applied to large-scale models and networks.

Many applications of macroscopic traffic models focus on freeway corridors or sec-

tions. The automated calibration of macroscopic supply models for such cases has been documented in the literature. Messmer and Papageorgiou (2001) use the Nelder-Mead method (a gradient-free algorithm working directly with objective function evaluations) to develop a parameter calibration approach for METANET. Several subsequent papers report on applications of the above method. Ngoduy and Hoogendoorn (2003) calibrate METANET for a section of the A1 freeway in The Netherlands, and use the calibrated model to study the prevention of traffic breakdown due to very high densities. The authors develop model predictive control (MPC) methods for setting dynamic freeway speed limits. Ngoduy et al. (2006) calibrate six parameters in the METANET speed-density function and shock wave propagation equation, for a freeway section with no ramps. An objective function measuring the fit to count and speed data is optimized.

A recent effort on the calibration of the supply models within a mesoscopic DTA system is described in Kunde (2002). A three-stage approach to supply calibration is outlined, in increasing order of complexity. At the *disaggregate* level, segment speed-density relationships are estimated similar to Van Aerde and Rakha (1995). In the second stage, a suitable *sub-network* is chosen, and the estimates from the previous stage are refined by accounting for interactions between the segments. The choice of a subnetwork depends on the structure of the network and the location of sensors. An ideal sub-network would allow one to deduce the true OD flows for the sub-network from the available sensor count information, so that the supply parameters may be inferred under known demand conditions. The final stage utilizes the entire network to incorporate demand-supply interactions into the calibration process.

The above thesis demonstrates the proposed approach using data from Irvine, California, by applying two simulation optimization algorithms to the supply calibration problem. SPSA (Simultaneous Perturbation Stochastic Approximation) approximates the gradient of the objective function through finite differences. Critically, the approach infers the components of the gradient vector from two function evaluations, after perturbing all components of the parameter vector simultaneously. The computational savings are thus significant when compared to traditional stochastic

approximation methods, though many replications may be required in order to obtain a more stable gradient through smoothing (Spall, 1994b). The Box-Complex algorithm (Box, 1965) was applied with better success, though the number of convergence iterations was still too few to study computational performance and the quality of the final solutions.

2.3.2 Microscopic supply calibration

Although microscopic traffic models are not within the scope of this thesis, we review a segment of literature on the calibration of such models. Some of the methods and algorithms employed in this context are relevant to the problem at hand, and may be appropriate for DTA model calibration after enhancement and modification.

The calibration of microscopic traffic simulation models has received serious attention in recent years, fueled by the widespread use of such models in professional and academic circles. Early studies often relied on manual adjustments and heuristics to reduce the discrepancy between observed and simulated quantities (see, for example, Daigle et al. (1998), Gabriel Gomes and Adolf May and Roberto Horowitz (2004) and Liu et al. (2004)). The time-consuming nature of this process, coupled with lack of a systematic approach capable of handling large parameter sets, have motivated research into the use of optimization algorithms to solve the calibration problem.

Kurian (2000) describes an early attempt to use sophisticated optimization packages for the calibration of MITSIMLab, using data from 16 sensor stations in the I-880 freeway corridor. He selects, through an experimental design, four parameters that control deceleration characteristics in the car-following model (the time-varying OD flows are obtained from a previous study, and are unchanged during calibration). In his approach, the BOSS Quattro package is used to optimize the chosen parameters using MITSIMLab as a black-box function evaluator. The algorithm, based on the steepest descent concept, progresses by moving a certain step size along a direction derived from the gradient at the current location. However, the inherent stochasticity in MITSIMLab, together with the optimizer's reliance on numerical gradients,

resulted in noisy derivatives that prevented stable convergence. Further, the highly nonlinear objective function resulted in the determination of different local optima based on the starting parameter values selected.

Subsequent experiences with MITSIMLab involved the use of the Box-Complex (Box, 1965) algorithm, a population-based approach that maintains a complex (or set) of parameter vectors (points) and their corresponding objective function values. The size of the complex was pre-determined based on the recommendations in Box (1965). The algorithm begins by initializing the complex with points generated at random, so as to cover the feasible region defined through lower and upper bounds on each individual parameter. At every iteration, a point with the "worst" objective function value is replaced by its reflection about the centroid of the remainder of the complex, thus driving the complex towards the optimal solution. Darda (2002) applies the Box-Complex method to estimate select car-following and lane-changing model parameters under the assumption of a fixed OD demand matrix. However, the convergence criterion on the maximum number of iterations was insufficient to ascertain convergence.

The gradient-free downhill simplex algorithm (adapted from the Nelder-Mead simplex procedure) was used by Brockfeld et al. (2005) to calibrate a small set of supply parameters in a wide range of microscopic and macroscopic traffic models. Others report on the successful application of genetic algorithms (GA) for the calibration of select parameters in various microscopic traffic simulation tools (Abdulhai et al., 1999; Lee et al., 2001; Kim, 2002; Kim and Rilett, 2003). The use of GA in transportation is illustrated by Kim and Rilett (2004), who describe the calibration of driving behavior parameters in the CORSIM and TRANSIMS microscopic models. The data for the research consisted of traffic volume data from the I-10 and US-290 freeway corridors around Houston, TX. Apart from the simple structure (and the corresponding lack of route choice) inherent to the test networks, the broader applicability of their work is also limited by the small number of parameters estimated. Indeed, the paper reports computational constraints even on such small examples! The numerical values of several algorithmic constants were also assumed from prior

analyses without sufficient elaboration. However, the authors do state the impact of the OD matrix on the final results, though they do not include OD flows as variables in the optimization.

Henderson and Fu (2004) provide a concise review of the transportation applications of GA to date. Indeed, the range of studies reported therein share several common characteristics that limit the scope of their conclusions:

- All applications are in the domain of traffic micro-simulation, and focus entirely on a subset of car-following and lane-changing parameters. Apart from being few in number (the biggest problem instance involved 19 parameters), OD flows were treated as exogenous to the GA application. This is a critical limitation, as the estimation of OD flows will significantly increase the scale of the optimization problem.

- The studies rarely compare the performance of GA against other well-established non-linear optimization methods. Often, the primary measure of performance is the improvement in the objective function value over the starting point, as employed by Yu et al. (2005). The claimed superiority of GA is thus not clearly established.

- GA involves a large set of highly sensitive tuning parameters and strategies such as variable encoding schemes and crossover and mutation probabilities. Most existing studies use default settings from earlier approaches, without any analysis or justification for their use.

Systematic optimization techniques are thus being increasingly applied for the calibration of supply model parameters. However, these experiences have been limited to simple networks and small parameter sets. While some of the algorithms have shown promise, tests on larger networks and variable sets should be performed to ascertain their suitability for overall DTA model calibration.

2.4 Estimation of demand models

Demand calibration involves the estimation of (a) travel behavior model parameters, (b) time-varying flows for each OD pair on the network, and (c) other parameters that the DTA model might use in the estimation and prediction of OD flows. We begin the review of demand model estimation techniques with a discussion of travel behavior model estimation.

The literature provides a rich spectrum of mode, departure time and route choice models that capture driver behavior at both *pre-trip* and *en-route* levels. Pre-trip decisions could include the choice of travel mode, departure time and route based on perceptions of expected traffic conditions for the trip under consideration. Trip chaining decisions (making one (or more) stop(s) before the final destination) and multi-modal route selections (including park-and-ride transit options) may also be made at the pre-trip stage. En-route decisions are made by drivers in response to evolving trip conditions. The most common example of an en-route choice is to change route due to unexpected traffic congestion, or in response to traveler information obtained through a Variable Message Sign (VMS) or an on-board device (such as a cell phone or radio).

The class of traveler behavior models considered here are *disaggregate*, in that they predict the choices made by individual drivers (trip makers). In Section 2.4.1, we briefly outline the standard discrete choice approach to estimating such models using disaggregate survey data.

Off-line demand calibration at the aggregate level has primarily focused on the estimation of OD flows from archived field measurements such as surveys, manual traffic counts or automated loop detector counts. The **OD estimation problem** has attracted substantial interest in the last few decades, and represents the calibration of demand parameters (OD flows) that form critical inputs to any DTA system. While OD estimation research covers both on-line (real-time) and off-line methods, our review of the relevant literature (Section 2.4.2) will be limited to the off-line case.

Section 2.4.3 reviews work on the joint calibration of a DTA model's demand

models in an off-line setting. Such approaches capture the role of travel behavior on the OD estimation problem, and attempt to incorporate their inter-relationships while obtaining consistent estimates of the various OD flows and travel behavior model parameters.

2.4.1 Travel behavior modeling

Discrete choice theory forms the backbone of most travel behavior analyses, and is best illustrated in the context of DTA through route choice. The route choice model contains the following dimensions:

- An individual driver $n \in \{1, 2, \ldots, N\}$ chooses from a set of alternatives (routes) C_n.

- Driver n is described by a vector of *characteristics*. Each route i in the choice set is similarly described by a vector of *attributes*. The combination of the characteristics for driver n, along with the corresponding attributes for route i, is represented by the vector X_{in}.

- Each driver n is assumed to perceive a "utility" associated with every route i in his/her choice set. The utilities map the attributes and characteristics into a real number for comparison.

- A decision rule is employed to determine the chosen route for each driver.

The principle underlying discrete choice theory is that of utility maximization: each individual n will pick the route with the maximum perceived utility $U_{jn},\ j \in C_n$. From a modeling perspective, however, the utilities U_{jn} are not directly observed. This discrepancy between the "true" utilities and their systematic model equivalents is captured through Random Utility Theory:

$$U_{in} = V_{in} + \epsilon_{in} \tag{2.1}$$

where V_{in} is a "systematic" utility computed as a linear function of the variables in X_{in}. For example, $V_{in} = \beta' X_{in}$, where β is a vector of coefficients to be estimated. The term ϵ_{in} represents the error between the model and the true utilities. Utility maximization then yields the probability of driver n selecting route i as:

$$P(i) = Pr(U_{in} \geq U_{jn} \; \forall j \in C_n) \tag{2.2}$$

Combining Equations (2.1) and (2.2), we get:

$$P(i) = Pr(\epsilon_{in} - \epsilon_{jn} \geq V_{jn} - V_{in} \; \forall j \in C_n) \tag{2.3}$$

Assumptions on the distribution of the error terms ϵ_{in} (or the difference $\epsilon_{in} - \epsilon_{jn}$) dictate the structure, richness and complexity of the resulting model. The assumption of normally distributed errors, for example, results in the Multinomial Probit (MNP) model, while Gumbel errors yield the popular Multinomial Logit (MNL) model.

The Probit model can potentially capture complex correlations among the alternative paths. However, its use involves the evaluation of high-dimension integrals that do not possess closed-form solutions. The Logit model, with its attractive closed-form expression, has thus been the most popular approach to capturing individual drivers' travel decisions. Several variants and extensions to the above model classes have been postulated, analyzed and tested, in view of the Logit model's inability to handle perceived correlations arising from the physical overlapping of alternative paths (the well-known IIA property). These include the C-Logit and Path-Size Logit models, and the flexible Logit Kernel approach. An overview of the various route choice model structures is presented in Ramming (2001).

Traditional route choice model estimation (or the calibration of the vector β) requires data from an individual (disaggregate) route choice survey. Each sampled driver n responds with his/her characteristics (including both socio-economic variables as well as descriptors such as trip purpose), a set of alternative routes in his/her choice set C_n, perceived route attributes and the chosen route.

The vector β of unknown systematic utility coefficients is estimated using standard concepts from Maximum Likelihood theory, by maximizing the joint probability of the chosen paths in the dataset. A detailed mathematical treatment of the mechanics of maximum likelihood estimation can be found in Ben-Akiva and Lerman (1985).

Disaggregate route choice models estimated from survey data possess several advantages. They provide a way of incorporating individual-specific characteristics and tastes into the systematic utilities. The resulting estimates also largely reflect actually observed choices made by individual drivers. Sampling issues, however, impose limitations on the choice-based model. A restricted sample size due to resource constraints and non-response may introduce bias in the estimated parameters, as the resulting datasets may not be representative of the general driver population. Justification bias (or a respondent's tendency to provide data that validates his/her choice) may further skew the resulting parameter estimates. Moreover, the high costs associated with administering surveys introduces significant lag times that could date the parameter estimates obtained. The use of aggregate data for enhancing route choice model estimation has only just begun to receive attention (examples include Ashok (1996), Tsavachidis (2000), Toledo et al. (2004), Jha et al. (2004) and Balakrishna et al. (2005a)).

2.4.2 The OD estimation problem

The OD estimation problem has received significant attention in the fields of transportation, computer network routing and general estimation theory. The problem focuses on the inference of the elements of an unobserved matrix $\mathbf{x}$ of point-to-point network trip demand[3], based on aggregate traffic flow measurements $\mathbf{y}$ collected at specific links on the network. The matrix $\mathbf{x}$ would have as many rows as there are potential trip origin nodes, and as many columns as there are destinations. Each cell in $\mathbf{x}$ thus represents the number of trips between a specific origin-destination pair.

Much of the OD estimation literature concentrates on the *static* problem. A single

[3]In the context of computer networks, demand may be defined in terms of data packets rather than vehicle trips.

OD demand matrix is estimated across a relatively large time period, such as an entire day or the morning peak (see, for example, Cascetta and Nguyen (1988)). Static approaches work with average flows across the entire study period. A critical limitation of such methods is their inability to capture within-period temporal patterns in OD demand, such as peaking. The demand inputs to DTA models must be *dynamic*, to facilitate the modeling of time-dependent phenomena such as the formation and dissipation of queues and spillback.

The following sections summarize the different approaches proposed for the estimation of dynamic OD demand from observed sensor count measurements[4]. Common to all methods is the assumption that the period of interest $\mathcal{H}$ is divided into intervals $h = 1, 2, \ldots, H$ of equal duration. Let $\mathbf{x}_h$ represent the matrix of OD flows departing their origins during interval h, and $\mathbf{y}_h$ the vehicle counts observed on various network links at the end of h. The objective of the dynamic OD estimation problem is to estimate the flows $\hat{\mathbf{x}}_h$ that replicate observed counts $\mathbf{y}_h$, $\forall h \in \mathcal{H}$.

Least squares approach

The most widely employed dynamic OD estimation technique is based on extensions to the least squares technique proposed by Cascetta and Nguyen (1988) in the static context. Cascetta et al. (1993) propose a generalized least squares (GLS) framework that fuses data from two sources to efficiently estimate dynamic OD flows. The authors present two alternative estimators that work within this framework. The *sequential* estimator optimizes for the unknown OD flows one interval at a time:

$$\hat{\mathbf{x}}_h = \arg\min_{\mathbf{x}_h} \left[f_1(\mathbf{x}_h, \mathbf{x}_h^a) + f_2(\mathbf{y}_h, \hat{\mathbf{y}}_h) \right] \tag{2.4}$$

where $\mathbf{x}_h$ is the current best solution; $\mathbf{x}_h^a$ are *a priori* flows (extracted from other studies, or set to $\hat{\mathbf{x}}_{h-1}$); $\hat{\mathbf{y}}_h$ are the fitted counts obtained by assigning $\mathbf{x}_h$ to the network; $f_1(\bullet)$ and $f_2(\bullet)$ are functions that measure the "distance" between the estimated or fitted quantities from their *a priori* or observed values. It is generally expected that

[4]Time-varying vehicle counts are currently the most common source of field traffic data.

the number of link count observations (the dimension of $\mathbf{y}_h$) is much smaller than the number of unknowns (the number of non-zero cells in $\mathbf{x}_h$). The *a priori* flows $\mathbf{x}_h^a$ thus provide valuable structural information that renders feasible a problem that is otherwise indeterminate.

A measurement equation maps the OD flows $\mathbf{x}_h$ to the counts $\mathbf{y}_h$ through a linear *assignment matrix* mapping:

$$\mathbf{y}_h = \sum_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p + \mathbf{v}_h \tag{2.5}$$

where the elements of $\mathbf{a}_h^p$ specify the fractions of each OD flow in $\mathbf{x}_p$ (departing during interval p) that arrive at every sensor location during interval h. $\mathbf{v}_h$ is an error term. p' indicates the number of intervals spanning the longest trip on the network, and is a function of network topology as well as congestion levels. Since the sequential estimator constrains the flows in prior intervals to their best estimates, the measurement equation may be re-written as:

$$\tilde{\mathbf{y}}_h = \mathbf{y}_h - \sum_{p=h-p'}^{h-1} \mathbf{a}_h^p \hat{\mathbf{x}}_p = \mathbf{a}_h^h \mathbf{x}_h + \mathbf{v}_h \tag{2.6}$$

Consistent with the GLS formulation, Equations 2.4 and 2.6 yield the following estimator:

$$\hat{\mathbf{x}}_h = \arg\min_{\mathbf{x}_h} \left[(\mathbf{x}_h - \mathbf{x}_h^a)' \mathbf{W}_h^{-1} (\mathbf{x}_h - \mathbf{x}_h^a) + (\tilde{\mathbf{y}}_h - \mathbf{a}_h^h \mathbf{x}_h)' \mathbf{R}_h^{-1} (\tilde{\mathbf{y}}_h - \mathbf{a}_h^h \mathbf{x}_h) \right] \tag{2.7}$$

The above optimization is constrained so that $\mathbf{x}_h \geq 0$. $\mathbf{W}_h$ and $\mathbf{R}_h$ are error variance-covariance matrices that may be used to reflect the reliability of the different measurements. Cascetta et al. propose setting them to identity matrices of appropriate dimensions, in the absence of reliable estimates for the same.

The authors propose a second estimator that solves for the OD flows in multiple

intervals *simultaneously*:

$$(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_H) = \underset{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_H}{\arg\min} \quad [f_1(\mathbf{x}_1, \hat{\mathbf{x}}_2, \ldots, \mathbf{x}_H; \mathbf{x}_1^a, \mathbf{x}_2^a, \ldots, \mathbf{x}_H^a) + f_2(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_H; \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_H] \tag{2.8}$$

with Equation 2.5 serving as the measurement equation for counts.

A qualitative comparison of the two estimators is in order. The sequential approach estimates OD flows $\mathbf{x}_h$ based only on $\mathbf{y}_h$, the first set of counts it contributes to (these flows are fixed while estimating OD flows for subsequent intervals). Future count measurements are thus not used to refine past estimates. The simultaneous estimator is more efficient in this regard, since it captures the contribution of $\mathbf{x}_h$ to the counts measured in all subsequent intervals. However, the approach involves the calculation, storage and inversion of a large augmented assignment matrix, which has been found to be too computationally intensive on large networks (see Cascetta and Russo (1997), Toledo et al. (2003) and Bierlaire and Crittin (2004)). The sequential approximation is therefore an attractive option for many applications.

Assignment matrices themselves are linear approximations of the relationship between OD flows and sensor counts. They are typically obtained through a network loading model that mimics the progression of candidate OD flows along a set of paths between each OD pair. Knowledge of drivers' route choice behavior and network travel times are thus essential to the computation of the assignment matrix, due to their roles in splitting OD flows into path flows that can subsequently be propagated along each path. Travel times, in turn, depend on the OD flows, whose "true" values are as yet unknown. OD estimation in the presence of congestion is therefore a *fixed-point* problem requiring an iterative solution methodology that captures the complex dependencies between the OD flows, the route choice model and the network loading model.

Cascetta and Postorino (2001) apply iterative schemes based on the method of successive averages (MSA) to solve the fixed-point OD estimation problem and obtain consistent OD flows and assignment matrices on congested networks. A GLS estima-

tor is used to generate updated flows in each iteration, which are then "smoothed" by the MSA technique. The authors provide empirical evidence in support of a modified OD smoothing algorithm (which they term MSADR[5]), that provides faster convergence by re-initializing MSA's iteration counter as the algorithm progresses. The re-setting of the counter is performed with decreasing frequency. Intuitive arguments are provided to show the equivalence of the final MSA and MSADR solutions. However, the approach pertains to the static case.

Kalman Filter approach

Ashok (1996) develops a sequential off-line OD smoothing scheme based on state-space modeling concepts. This approach uses *transition* equations to capture the evolution of system *state*, and *measurement* equations to incorporate the sensor count measurements. The authors provide a key innovation over previous state-space approaches, by defining the state in terms of *deviations*: the difference of OD flows $\mathbf{x}_h$ from their historical or expected values $\mathbf{x}_h^H$. The above transformation allows the state to be represented through symmetrical distributions (such as normal) that possess desirable estimation properties, which would not be appropriate for OD flows directly.

A transition equation based on an autoregressive process describes the interval-to-interval evolution structure for network state:

$$\mathbf{x}_{h+1} - \mathbf{x}_{h+1}^H = \sum_{p=h-q'+1}^{h} \mathbf{f}_{h+1}^p (\mathbf{x}_p - \mathbf{x}_p^H) + \mathbf{w}_{h+1} \tag{2.9}$$

where $\mathbf{f}_{h+1}^p$ is a matrix relating spatial and temporal OD relationships between intervals p and $h+1$. The parameter q' is the degree of the autoregressive process, representing the length of past history affecting the current interval.

The measurement equation is obtained by adapting Equation 2.5 to work with deviations:

$$\mathbf{y}_h - \mathbf{y}_h^H = \sum_{p=h-p'}^{h} \mathbf{a}_h^p (\mathbf{x}_p - \mathbf{x}_p^H) + \mathbf{v}_h \tag{2.10}$$

[5]MSADR stands for MSA with Decreasing Reinitializations.

with $\mathbf{y}_h^H = \sum_{p=h-p'}^{h} \mathbf{a}_h^p \mathbf{x}_p^H$ serving as historical count estimates. $\mathbf{R}_h$ and $\mathbf{Q}_h$ denote the error covariance matrices for $\mathbf{v}_h$ and $\mathbf{w}_h$ respectively. The author provides a Kalman Filter solution approach for estimating $\mathbf{x}_h$: In a forward pass, the flows $\mathbf{x}_h$ are estimated sequentially for $h = 1, \ldots, H$ (ignoring the contribution of $\mathbf{x}_h$ to counts $\mathbf{y}_{h+1}, \mathbf{y}_{h+2}, \ldots, \mathbf{y}_H$). Each OD matrix is then re-estimated (updated) while backtracking from $h = H$ to $h = 1$. The information contained in $\mathbf{y}_H, \mathbf{y}_{H-1}, \ldots, \mathbf{y}_{h+1}$ is thus completely used in identifying the flows for interval h.

This work also contains modeling enhancements for practical applications. First, state augmentation is proposed as a way of improving the efficiency of the estimated OD flows by exploiting the information about prior OD departure intervals $(\mathbf{x}_{h-1}, \mathbf{x}_{h-2}, \ldots)$ contained in sensor measurements $\mathbf{y}_h$. In this approach, OD deviations from a pre-defined number of past intervals are added to the state vector, and are re-estimated periodically as future intervals are processed. State augmentation may be perceived as a compromise between the computationally attractive sequential estimator, and its more efficient simultaneous adaptation.

Further, the author briefly discusses methods to estimate the initial inputs required by the Kalman Filter algorithm: the historical OD flows $\mathbf{x}_h^H$, error covariance matrices $\mathbf{Q}_h$ and $\mathbf{R}_h$, and autoregressive matrices $\mathbf{f}_h^p$, which are an important part of off-line demand model calibration.

Maximum Likelihood (ML) approach

Hazelton (2000) presents an OD estimation methodology that uses traffic counts and *a priori* OD flow estimates in a maximum likelihood framework. A theoretical estimator is developed by assuming a general distribution for the OD flows and sensor counts. Indeed, they show that the GLS formulations by Cascetta et al. (1993) may be obtained by selecting the normal distribution to model all error terms. While this approach presents an elegant generalization of some prior results, its applicability to the DTA calibration context is limited. First, the analysis focuses on static OD estimation. The further assumption of uncongested network conditions, while allowing the author to ignore temporal dynamics in route choice fractions, renders the

approach unrealistic in the DTA context. Lastly, the general framework still requires the assumption of a realistic distribution if the method is to become operational.

van der Zijpp's approach

A method for estimating OD flows on freeway networks is developed by van der Zijpp (1996), in which the time interval boundaries are determined by analyzing space-time trajectories. Assuming that vehicle speeds are known, the trajectories of the first and last vehicles in each departure interval are calculated. Trajectories for all other vehicles departing during the interval are determined based on first-in, first-out (FIFO) rules. The set of trajectories is then used to estimate split fractions that allocate sensor counts to OD flows from the current and previous intervals. The split fractions are modeled by a truncated multivariate normal (TMVN) distribution, and are updated at each step through a Bayesian formula.

The above approach has been packaged into the DelftOD software (van der Zijpp, 2002), which has been applied in many freeway situations (see, for example, Hegyi et al. (2003), Ngoduy and Hoogendoorn (2003) and Park et al. (2005)). However, the lack of a closed-form expression for the TMVN distribution poses practical difficulties when determining its mean. Further, the calculation of complete vehicle trajectories requires knowledge of speeds during the entire trip. An accurate predictor of future speeds or travel times is thus essential for real-world applications.

A note on assignment matrices

In his thesis, Ashok (1996) outlines two ways of obtaining an assignment matrix for OD estimation. The simpler approach involves the use of a traffic simulator, say the DTA model being calibrated, to load the current best OD flows onto the network. The required fractions in the assignment matrix may then be calculated through a simple book-keeping of vehicle records at sensors. However, recent experiences with a network from Los Angeles (Gupta, 2005; Balakrishna et al., 2006) have indicated that simulated assignment matrices may be sub-optimal for OD estimation. An issue of particular concern centers around the stability of the calculated assignment fractions.

In the absence of a good starting OD matrix, artificial bottlenecks may result due to spatial and temporal OD patterns that are far from optimal, yielding incorrect assignment fractions. The use of small starting flows to offset this problem typically results in highly stochastic and small fractions, since few vehicles will be assigned to each path.

An alternative approach for calculating the assignment fractions from link travel times has been discussed by Ashok (1996). Under certain assumptions regarding vehicles' within-interval departure times[6], and with knowledge of time-dependent link traversal times, one may calculate *crossing fractions* that represent the percentage of each path flow departing during interval $\mathtt{p}$ that reaches every sensor during interval $\mathtt{h}$. The assignment fraction for a given sensor and OD pair may be computed by summing across all paths (for the OD pair) the product of crossing fractions (to the sensor in question) and the corresponding path choice fractions (obtained through the application of a route choice model). Such analytically calculated assignment matrices possess many advantages. First, the link travel times obtained from a traffic simulator are average values computed from several vehicles (across different OD pairs). The resulting assignment fractions are therefore less stochastic than those obtained directly from the simulator. Secondly, uncongested (free-flow) travel times are generally known with a high degree of accuracy, from observed sensor speed data. The assignment fractions (and estimated OD flows) for the intervals leading up to the congested regime are therefore accurate, and may be expected to yield accurate travel times even in subsequent intervals when coupled with sequential OD estimation. Further, the contributions of current OD departures on future intervals accurately account for congestion (through the travel times), minimizing the effects of the starting OD matrix. Finally, the calculated fractions capture all possible paths, including those that may be assigned few or no vehicles during the simulation. The last two points, however, lead to assignment matrices that are not as sparse as their simulated counterparts, and significantly increase the time taken to solve the OD

[6]Vehicle departure times within an interval are assumed to be uniformly spaced, with the first and last vehicles departing at the beginning and end of the interval.

estimation problem. Gupta (2005) reports on the improved convergence and fit to counts when the assignment matrices are calculated from travel times.

The role of the assignment matrix in OD estimation underscores the importance of route choice in demand model calibration. Some prior studies have strived to estimate accurate network travel times (used by the route choice model) that are consistent with the estimated OD flows. However, few have focused on the parameters of the route choice model itself. Often, these parameters are assigned *ad hoc* or convenient values that are then fixed for the remainder of the OD estimation procedure. Mahmassani et al. (2003), for example, describe the calibration of dynamic OD flows for DYNASMART using traffic sensor data and *with assumed route choice model parameters*. The authors cite the general lack of calibration data as the reason for assuming known route choice splits. These splits are hypothesized as outputs of some other procedure, and are hence exogenous to the functioning of the DTA system.

Ignoring the role of the route choice model can lead to biased and inconsistent estimates of travel demand. We now look at some literature related to the joint estimation of OD flows and the route choice model parameters.

2.4.3 Joint estimation of OD demand and travel behavior models

The demand simulator of a DTA model relies on estimates of OD demand, route choice model parameters and network travel times in order to accurately model the network and its underlying demand patterns. Demand calibration therefore involves solving a fixed-point problem that explicitly includes the route choice model parameters as variables. Initial research calibrating OD flows treated route choice as external to the OD estimation problem, potentially leading to biased OD flow estimates. Cascetta and Nguyen (1988), for example, assume an *ad hoc* travel time coefficient in the route choice model while estimating OD flows using a GLS approach.

Ashok (1996), while demonstrating his Kalman Filter based OD estimation methodology, estimates a route choice model using traffic counts data from the A10

beltway in Amsterdam. The Logit route choice model contained a single coefficient, that of travel time, whose optimal value (resulting in the best fit to the observed counts) was identified through a line search that was conducted independent of the OD estimation process. The final estimated parameter was very low, corresponding to an all-or-nothing assignment to the shortest path. It should be noted that the beltway provides exactly two paths between each OD pair, with little or no overlap between them. An all-or-nothing assignment would therefore be reasonable for this network, where one of the two routes is often much longer than the other.

Balakrishna (2002) uses multiple days of sensor counts to jointly calibrate dynamic OD flows and a route choice model within the DynaMIT traffic estimation and prediction system. The study focused on an urban network from Irvine, CA consisting of both arterial and freeway links. A static OD matrix for the AM peak (available from the Orange County MPO through a previous planning exercise) was adjusted systematically using a sequential GLS estimator to obtain dynamic OD matrices for the entire AM peak. A Path-Size Logit based route choice model (Ramming, 2001; Ben-Akiva and Bierlaire, 2003) with three parameters was estimated using an approach similar to the one outlined by Ashok (1996).

The joint calibration of DynaMIT's route choice model and OD estimation and prediction model using three days of sensor count data was carried out iteratively (a detailed presentation of the algorithm may be found in Balakrishna et al. (2005a)). Other estimated parameters included the error covariance matrices and autoregressive factors used by DynaMIT's OD estimation and prediction module. The performance of the calibrated DynaMIT system was validated using two independent days of data not used during calibration.

A related effort (Sundaram, 2002) develops a simulation-based short-term transportation planning framework that jointly estimates dynamic OD flows and network equilibrium travel times. While the coefficients of the route choice model are not estimated, a consistent set of OD flows and travel times are obtained (for the given route choice model) by iterating between an OD estimation module and a day-to-day travel time updating model. The basis for the travel time iterations is a time-smoothing

procedure:

$$\mathbf{TT}_k^{rc} = \lambda\, \mathbf{TT}_{k-1} + (1-\lambda)\, \mathbf{TT}_{k-1}^{rc} \tag{2.11}$$

where drivers' perceived route choice travel times $\mathbf{TT}_k^{rc}$ on day k are modeled as a function of the perceived estimates $\mathbf{TT}_{k-1}^{rc}$ from the previous day, and the latest *experienced* (simulated) quantities $\mathbf{TT}_{k-1}$. The parameter λ captures a learning rate (a value between 0 and 1) whose magnitude would be affected, among other factors, by drivers' familiarity with the network and its traffic patterns, and the prevalence of traveler information services.

Sundaram's approach operates in two steps. Travel times are established for a given set of dynamic OD demands. The resulting equilibrium travel time estimates are used to re-calculate assignment matrices for OD estimation. Travel times may then be computed again based on the new OD estimates if convergence has not been attained.

2.5 Conclusions: state-of-the-art (reference case)

The discussion in this chapter leads to a definition of standard procedures adopted when DTA models are currently calibrated using aggregate sensor data. The state-of-the-art, as defined here, will form the reference (henceforth known as the reference case, or Ref) against which the calibration methodology developed in this thesis will be compared:

- Demand and supply models are calibrated independently (sequentially), ignoring the effect of their interactions. Supply parameters are estimated first, then demand parameters are calibrated with fixed supply models.
- Supply calibration:
  - Capacities are estimated from sensor data and network geometry (primarily the number of lanes), based on the Highway Capacity Manual (HCM). Capacities during incidents are approximated from the HCM, based on the number of affected lanes, total number of lanes and incident severity.

  - Speed-density functions (or volume-delay relationships) are identified locally (ignoring network effects) by fitting appropriate curves to sensor data. Links are grouped according to physical network features (such as the number of lanes and the position relative to diverge and merge points, on- and off-ramps), and the most representative function is assigned to each group.

- Demand calibration:

  - OD flows and route choice model parameters are estimated iteratively (sequentially).

  - Time-dependent OD flows are estimated sequentially[7] using one of several methods (GLS or the Kalman Filter, for example) that rely on a set of assignment matrices. The assignment matrices may be simulated, or computed analytically based on the latest known travel times and route choice parameters.

  - Route choice parameters are estimated through manual line or grid searches. A limited number of parameters might be handled in this way.

2.6 Summary

A review of the literature indicates several shortcomings in the state-of-the-art of DTA model calibration, primary being the sequential treatment of demand and supply parameters. Most prevalent practices rely on heuristics and manual parameter adjustment approaches that are largely based on judgment. Applications of systematic optimization algorithms for model calibration have been few, and focus primarily on DTA model *components*. Moreover, these studies typically estimate a small subset of parameters deemed important in explaining observed data for specific networks and datasets, and typically do not perform sufficient iterations to ensure a high degree of accuracy.

[7] As discussed earlier in Section 2.4.2, the sequential approach, while computationally attractive, may not accurately capture long trips encountered on large or highly congested networks.

Limited experience with simulation optimization in the realm of transportation indicate the promise of select algorithms (such as the Box-Complex algorithm), but not others (such as stochastic approximation or gradient-based methods). There is need to explore this topic in depth, and develop a robust calibration methodology that can simultaneously estimate both demand and supply model parameters in a simulation-based DTA system. Chapter 3 presents a rigorous treatment of the off-line DTA calibration problem, analyzes its dimensions and characteristics, and proposes a robust and systematic estimator for its solution.

Chapter 3

Methodology

Contents

This chapter begins by outlining the various model inputs and parameters typically encountered in a generic simulation-based DTA model, and collects the variables that may need to be calibrated. The sensor data to be used for calibration are described next, followed by a mathematical statement of the problem. The statement is subsequently expanded to provide a detailed formulation and framework that accounts for the various issues and complexities associated with identifying model parameters through simulation. Finally, solution approaches are outlined in order to make the overall calibration framework operational.

3.1 Calibration variables

The set of critical DTA model parameters that must be calibrated for a specific network may be grouped into demand- and supply-side variables. Demand variables are typically common to all classes of DTA models (e.g. microscopic and mesoscopic), and include time-dependent OD demand for the period of interest as well as travel behavior model inputs and parameters. The vector $\mathbf{x}_{\mathfrak{h}}$ represents the OD demand rates between the $\mathfrak{n}_{\mathsf{OD}}$ OD pairs on the network, departing their origins during interval $\mathfrak{h}$. Demand is assigned to the network using drivers' perceived network travel times, through a probabilistic route choice model. The basic parameters in such a model include coefficients for various path attributes such as travel time, fraction of freeway links, number of left turns, number of signalized intersections and the frequency of freeway-arterial changes.

The number and nature of supply variables may vary depending on the level of detail employed while capturing traffic dynamics and queuing phenomena. Microscopic models generally possess a much wider set of models and parameters that operate under different traffic regimes and explain a complex set of individual driver decisions and maneuvers. These include car-following (acceleration, deceleration and desired speed), lane-changing (gap acceptance, merging, yielding and look-ahead) and compliance (response to arterial signals, ramp meters and toll plazas). Mesoscopic models capture network performance through aggregate (macroscopic) methods in-

volving parameters such as link/segment capacities and speed-density relationships. While the number of speed-density functions (and hence parameters) can technically be very large, it is limited in practice through the classification of segments into a few representative groups (see, for example, Van Aerde and Rakha (1995)). It should be noted that the vector $\boldsymbol{\beta}$ is used to denote the combined parameters in the route choice and supply models.

A more detailed list of parameters in the context of mesoscopic DTA may be found in Appendix A, which describes the modeling concepts behind the DynaMIT prediction and guidance generation system in depth.

3.2 Sensor data for calibration

Figure 3-1 provides a high-level overview of the traffic data collection process. Traffic conditions experienced on a road network are a result of interactions between the network itself (with its topology of alternative routes and their capacities), individual drivers (who constitute the demand, and who make a multitude of decisions based on their knowledge and perceptions about the network), and external factors (both unexpected, such as incidents, and anticipated, such as weather conditions or special events) that perturb the system away from "normal" or "regular" operating conditions. A subset of vehicles moving on the network are intercepted by instruments that form the *surveillance system*: a collection of modern technologies such as video cameras, closed-circuit televisions, loop detectors and electronic tag readers. that sense individual vehicles, and record traffic measurements.

The type of traffic data available through the surveillance system (or obtained by post-processing the recorded data) depends on the sensing technology deployed on the network. Footage from a video camera can be processed to yield vehicle trajectories that show individual driver behavior such as lane changing and speed adjustments. Data at such a fine level of detail are termed **disaggregate**, since they provide information at the resolution of the individual driver. However, the collection and processing of video data is extremely time-consuming, and such data

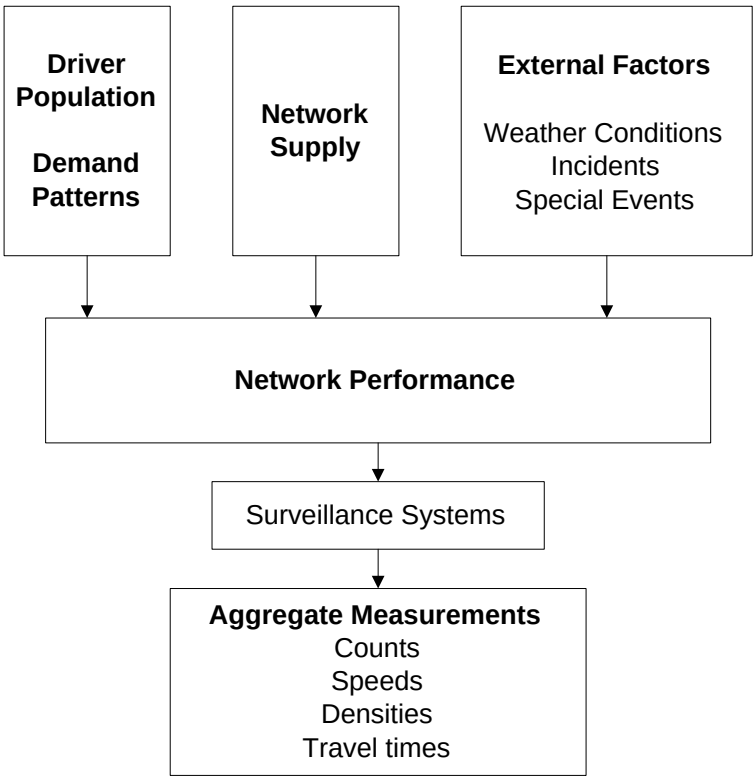

Figure 3-1: The Aggregate Measurement Process

is currently available only for short periods of time and on small sections of roads. Other disaggregate sources of traffic data include surveys for OD demand and route choice model estimation. It should be noted that the effort and costs involved in designing and administering surveys precludes their frequent application. Moreover, survey data inevitably suffers from response bias and sampling inefficiencies.

By contrast, **aggregate** data is collected at a lower resolution, representing the sum of multiple vehicles in each measurement. The most common source of aggregate measurements are inductance loop detectors buried beneath the pavement surface. These devices have the capability to record traffic counts, detector occupancy rates and average vehicle speeds, which can be collected in a central database for model calibration purposes. Since the data collection process is automatic, the most recent data may be obtained with minimum effort (once the surveillance system has been installed). Such sensor data also captures a larger proportion of the population, and reflects actual driver behavior (thus eliminating response bias).

A DTA system's component models, like the various sources of traffic data, can be disaggregate or aggregate. Disaggregate models predict individual driver decisions, such as route, departure time and mode choice, speed and acceleration choice and lane selection. On the other hand, the estimation and prediction of OD flows is performed at the aggregate level: the total flow between each OD pair, across all individual drivers. In this thesis, we develop a methodology for estimating all DTA model components (both disaggregate and aggregate) using aggregate data.

The aggregate traffic dataset used in this research consists of information collected using inductive loop detectors. Such sensors consist of wire loops embedded beneath the pavement surface. Metallic vehicles crossing the loops cause changes in their electromagnetic fields, which indicate the presence of vehicles above the loops. Individual loop detector units are designed and calibrated to translate these magnetic data into traffic descriptors. The traffic flow at a sensing location during time interval is typically recorded as the count of vehicles crossing the loop during the interval. This measure may then be converted to hourly flow (volume) estimates q. Loops also measure occupancy occ, the percentage of a time interval for which vehicles were sensed over the loops. Sensor occupancy is a reflection of the spatial density (sometimes known as *concentration*) of vehicles in the vicinity of the loop.

The space mean speed v may be inferred from flow and occupancy data, through the fundamental equation

$$v = \frac{q}{k} \tag{3.1}$$

together with the following relationship (May, 1990):

$$k = \frac{52.8}{\overline{L_V} + L_D}\,(occ) \tag{3.2}$$

where k is the density (vehicles/lane-mile); $\overline{L_V}$ is the mean vehicle length (feet); L_D is the detector length (feet). In some situations, accurate speed measurements may be obtained directly from the detector. When loops are deployed in pairs, vehicles' travel times between the loops may be measured and the corresponding speeds inferred.

In this thesis, we employ loop detector count and speed data to illustrate the

several dimensions and the solution of the calibration problem. Sensor data for calibration must be recorded over consecutive, uniform measurement intervals that span the period of interest.

3.3 The historical database

Off-line calibration serves two primary goals. The first is to provide a complete set of DTA model inputs and parameters (called a *historical database*) that may be used by planners and modelers for future applications of the model. OD demand patterns, route choice behavior and supply processes vary in the real world according to several factors. The OD flows for the morning peak period, for example, could depend on the day of the week. Further, route choice and traffic dynamics may be different on dry and rainy/snowy days. Traffic data covering such a wide range of conditions will be available through the archived sensor dataset. The historical database must therefore reflect this diversity in a practical way.

In theory, one can estimate a separate historical database for each day of observed sensor measurements, the underlying assumption being that every day is unique. However, such an approach raises the issue of selecting a database as DTA inputs on a new (future) day. A more useful approach is the classification of the historical database along criteria identified after closely analyzing the available sensor data from a large number of days. For example, the database might be stratified based on day of the week, weather conditions and special events. On a new day, the profile that best fits the criteria (both known, such as day of the week, and forecast, such as weather) can then be used to pick the most appropriate database for application. Obviously, the criteria used to set up the database may differ with geographic location, season, etc.

The second goal of calibration involves the maintenance of the historical database on a day-to-day basis, as new traffic measurements are recorded by the surveillance system. Travel patterns and traffic conditions evolve over time. While systematic day-to-day changes in OD patterns and network travel times may not be significant

in the short term, their impacts in the long term could be substantial. There is thus a need for a calibration methodology that can be re-applied at regular intervals in order to update the historical database. For example, the historical estimates (for the relevant class of day) could be updated at the end of each day.

The two goals above motivate a calibration framework based on a day-to-day updating mechanism, so that the historical database may be adjusted with the new estimates at the end of each day. We will now outline prior work in this area before discussing the scope of the current research on the updating framework.

The day-to-day updating framework proposed by Balakrishna et al. (2005a) is reproduced in Figure 3-2. This approach assumes the availability of sensor count data from M days of the same type, and estimates a historical database of *demand* parameters using a sequential process. OD flows and error covariances are estimated for the first day ($m = 1$) through the sequential estimator outlined in Section 2.4.2. Sequential OD estimation for the departure intervals $h = 1, 2, \ldots, T$ are iterated with error covariance estimation until convergence. The OD estimates from interval h form the *a priori* estimates for interval $h + 1$ for the sequential OD estimation.

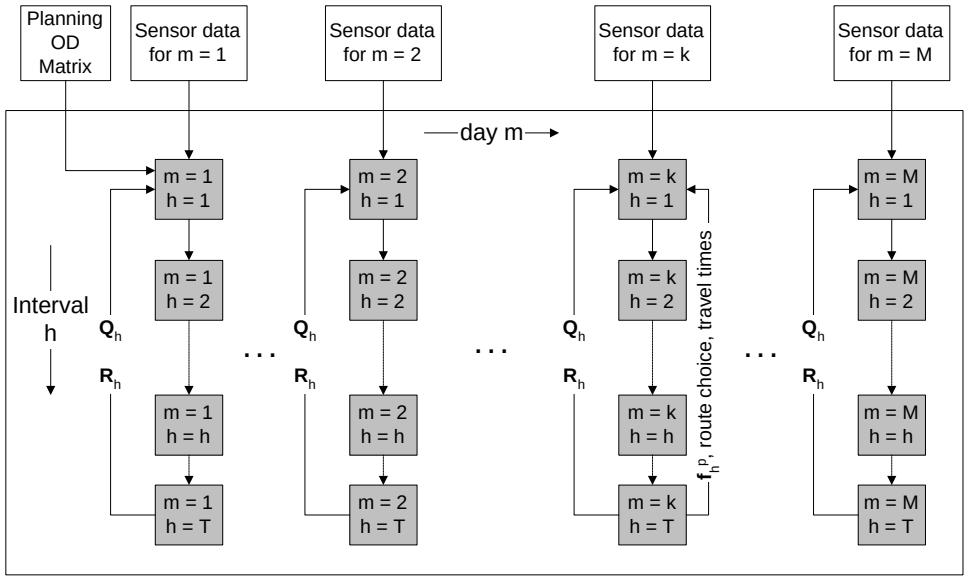

Figure 3-2: Day-to-Day Updating (Balakrishna et al., 2005a)

The estimated demand parameters from day $m = 1$ are used to calculate the

a priori inputs for the next day ($\mathfrak{m} = 2$). Demand parameters for this new day are obtained, and the procedure progresses until all M days are utilized. It should be noted that the framework allows for the periodic calibration of auto-regressive factors, route choice parameters and travel times. At the end of this calibration process, the historical database may be defined by one of several methods (Ashok, 1996; Balakrishna et al., 2005a). One option is to define the latest estimated parameters (OD flows, error covariances, auto-regressive factors, route choice parameters and travel times) as the current historical values that encapsulate the entire prior sequence of estimations. Alternatively, a moving average of all parameters could be maintained, and used as input for the next day of this type encountered in practice. Other updating schemes may also be developed. However, the evaluation of different schemes is beyond the scope of this thesis, and is a suitable topic for future research.

The focus of the preceding description is the calibration of demand variables using sensor count data. Supply model parameters are therefore treated as exogenous inputs. In this research, we replace the within-day sequential OD estimator described in the above framework, with an estimator that:

- calibrates the demand and supply model parameters (including route choice parameters) simultaneously,
- estimates OD flows across multiple intervals $\mathfrak{h}$ simultaneously, and without using the assignment matrix, and
- utilizes count and speed data.

The rest of the methodology (pertaining to the estimation of error covariances, OD prediction model parameters and travel times $\mathbf{TT}^{\mathrm{rc}}$) thus remains the same. A note on the application of our proposed methodology for the first few days is in order. While the approach can simultaneously estimate parameters across multiple intervals, this would require time-varying *a priori* parameter values. Such values are typically not available until a reliable historical database has been established. Under these circumstances, a sequential within-day estimation procedure across time intervals may

be used to initialize the process, noting that the demand and supply parameters may still be estimated simultaneously for each interval. The advantage of this approach is that the estimates for h may serve as *a priori* values for $h+1$ for the first few days, to be replaced by time-varying historical values once stable estimates are obtained.

The above initialization method is related to the concept of system observability, discussed in Gupta (2005) and Balakrishna et al. (2006) (and briefly in Ashok (1996) and Section 2.2). When sensor coverage is low, traditional OD estimation methods rely on *a priori* OD flows in order to solve for the unknown OD flows. However, the choice of *a priori* flows might impact the final outcome of the estimation. Observability ensures that sequential OD estimation from a nearly empty network (potentially encountered in the early hours of the day) yields reasonably stable OD estimates after many intervals, independent of the *a priori* flows selected for $h = 1$. The number of intervals may be roughly determined by comparing the number of OD flows to sensors (and is verified empirically in Gupta (2005)).

The methodology in this chapter focuses on simultaneously estimating the best model inputs (OD flows, route choice parameters and supply parameters) for a single day. We now formulate the resulting problem, dropping the day-index m without loss of generality.

3.4 General problem formulation

The off-line calibration problem is formulated using the following notation:

$\mathbf{x}$: OD flows, $\mathbf{x} = \{\mathbf{x}_h\},\ \forall\ h \in \mathcal{H}$

$\boldsymbol{\beta}$: Model parameters, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_h\}$

$\mathbf{M}$: Observed aggregate sensor count and speed measurements, $\mathbf{M} = \{\mathbf{M}_h\}$

$\mathbf{x}^a$: *A priori* OD flows, $\mathbf{x}^a = \{\mathbf{x}_h^a\}$

$\boldsymbol{\beta}^a$: *A priori* model parameter values, $\boldsymbol{\beta}^a = \{\boldsymbol{\beta}_h^a\}$

G : Road network (given), $G = \{G_h\}$

Let $\mathcal{M}$ be the set of all *fitted* sensor measurements. The off-line calibration problem is mathematically stated as the minimization of an objective function over the parameter space:

$$\underset{\mathbf{x},\,\boldsymbol{\beta}}{\text{Minimize}}\ z\left(\mathbf{M}, \mathcal{M}, \mathbf{x}, \boldsymbol{\beta}, \mathbf{x}^a, \boldsymbol{\beta}^a\right) \tag{3.3}$$

subject to the following constraints:

$$\mathcal{M} = f(\mathbf{x}, \boldsymbol{\beta}, G) \tag{3.4}$$

$$l_x \leq \mathbf{x} \leq u_x \tag{3.5}$$

$$l_\beta \leq \boldsymbol{\beta} \leq u_\beta \tag{3.6}$$

Equation (3.3) denotes the objective function for the minimization problem, and is interpreted as some goodness-of-fit measure between observed (or *a priori*) quantities and the corresponding fitted values. $\mathcal{M}$, the output of the model, is represented explicitly through constraint 3.4. The output for a particular interval h may be

stated as:

$$\mathcal{M}_h = f(\mathbf{x}_1, \ldots, \mathbf{x}_h; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_h; G_1, \ldots, G_h) \tag{3.7}$$

i.e. $\mathcal{M}_h$ is a function of all OD flows, model parameters and network characteristics up to (and including) the current interval. Further, any stochasticity in the model's output is controlled through the function $f(\bullet)$, by defining $\mathcal{M}_h$ as the mean output from multiple replications (a more detailed discussion on this aspect is provided in Section 3.5.4).

The optimization is performed over the vector of unknown quantities $\mathbf{x}$ and $\boldsymbol{\beta}$, whose lower and upper bounds are captured through Equations 3.5 and 3.6. Parameter bounds define the search space to be explored by the solution algorithm. The lower bounds l_x are particularly useful, as they may be used to enforce non-negativity of relevant parameters (such as capacities and OD flows). Route choice model parameters may employ the upper bound to restrict the search space. The coefficient of travel time, for example, must be negative, and hence has an upper bound of zero.

We expand the general objective function (Equation 3.3) into a three-part expression that measures the distances of measurements, OD flows and model parameters:

$$\underset{\mathbf{x},\,\boldsymbol{\beta}}{\text{Minimize}} \;\; [z_1(\mathbf{M}, \mathcal{M}) + z_2(\mathbf{x}, \mathbf{x}^a) + z_3(\boldsymbol{\beta}, \boldsymbol{\beta}^a)] \tag{3.8}$$

The functional forms of $z_1(\bullet)$, $z_2(\bullet)$ and $z_3(\bullet)$ depend on the specific goodness-of-fit measures chosen. Under the least squares framework, for example, $z_i(\epsilon_i) = \epsilon_i' \Omega_i^{-1} \epsilon_i$. The index $i = \{M, x, \beta\}$ denote the sensor data, OD flows and model parameters respectively. The term Ω_i represents a variance-covariance matrix. The deviations are computed as

$$\begin{aligned} \epsilon_M &= \mathbf{M} - \mathcal{M} \\ \epsilon_x &= \mathbf{x} - \mathbf{x}^a \\ \epsilon_\beta &= \boldsymbol{\beta} - \boldsymbol{\beta}^a \end{aligned}$$

The general formulation assumes that all calibration variables are dynamic. Specif-

ically, the model parameters $\boldsymbol{\beta}_h$ depend on h, reflecting the impact of unobserved time-varying factors on travel and driving behavior. These effects include vehicle mix, weather conditions and drivers' response to major and minor incidents. Modifications of the general approach may be constructed to fit specific modeling hypotheses that impose additional constraints on the optimization problem. We present one such special formulation before outlining the algorithmic component of our methodology.

A special case imposes constraints on the values of model parameters across the intervals in $\mathcal{H}$. It may be reasonable to envision that the parameters $\boldsymbol{\beta}$ remain constant across the period of interest. For example, the route choice model parameters are likely to be unchanged across the entire day. It may also be reasonable to suppose that capacities and speed-density functions are constant for given weather conditions. Note that the OD demands $\mathbf{x}_h$ are still modeled as dynamic variables.

The problem formulation under this hypothesis thus becomes:

$$\underset{\mathbf{x},\,\boldsymbol{\beta}}{\text{Minimize}} \sum_{h=1}^{H} [z_1(\mathbf{M}_h, \mathcal{M}_h) + z_2(\mathbf{x}_h, \mathbf{x}_h^a)] + z_3(\boldsymbol{\beta}, \boldsymbol{\beta}^a) \tag{3.9}$$

subject to the following constraints:

$$\mathcal{M}_h = f(\mathbf{x}_1, \ldots, \mathbf{x}_h; \boldsymbol{\beta}; G_1, \ldots, G_h)$$

In the context of generalized least squares, the objective function in 3.9 becomes:

$$\underset{\mathbf{x},\,\boldsymbol{\beta}}{\text{Minimize}} \sum_{h=1}^{H} \left[\epsilon'_{Mh}\, \Omega_{Mh}^{-1}\, \epsilon_{Mh} + \epsilon'_{xh}\, \Omega_{xh}^{-1}\, \epsilon_{xh}\right] + \epsilon'_{\beta}\, \Omega_{\beta}^{-1}\, \epsilon_{\beta} \tag{3.10}$$

with $\epsilon_{Mh} = M_h - \mathcal{M}_h$ and $\epsilon_{xh} = \mathbf{x}_h - \mathbf{x}_h^a$ (ϵ_β is as defined earlier). It may be possible to relax the dependence of covariance structures on time interval h in some situations.

The above approach assumes the absense of serial correlation across time intervals h. While this may at first be perceived as a strong assumption, it should be noted that spatial and temporal correlation effects are at least partially captured implicitly

through the model output $\mathcal{M}_h$ (the additional equations being embedded within the DTA model).

A comparison of the proposed methodology with existing OD estimation approaches is relevant here, for two reasons:

1. OD flow variables typically dominate any calibration problem, since their number increases rapidly with network size and the length of the study period.

2. Many prior studies have focused on approximations to simplify OD estimation, while treating the route choice and supply parameters as exogenous inputs.

The most popular OD estimation method is one that replaces Equation 3.7 with a linear approximation that maps the OD flow variables to the sensor data. The corresponding mapping, termed an assignment matrix, is essentially a set of linear equations that list the contributions of each OD flow to the vehicle counts observed at each sensor location (refer to Equation 2.5):

$$\hat{\mathbf{y}}_h = \sum_{p=h-p'}^{h} \mathbf{a}_h^p \, \mathbf{x}_p$$

Apart from being an approximation of the true relationship, the above equation also restricts the estimation to the use of sensor counts alone. The relationship between OD flows and other data (such as speeds or travel times) is clearly non-linear, and cannot be modeled as a linear equation. Thus, OD estimation, until recently, has been solved using counts alone.

Recent experiences have indicated the existence of multiple sets of OD flows that yield the same fit to link counts (see, for example, Lee et al. (2006)). This behavior is primarily because of the dependence on *a priori* flows to help resolve the underdetermined nature of the problem due to limited sensor count data coverage. The typical estimation method picks one of these many solutions, which often does not replicate traffic dynamics very well (the fit to speeds or probe vehicle travel times, for example, may be poor). It is therefore critical to constrain the calibration using additional traffic descriptors. While this procedure is difficult with the prior formu-

lation, our proposed formulation can easily include any type of field measurements. The non-linear relationship between OD flows and the observed field measurements are captured implicitly through the simulator (Equation 3.7). The generality of this approach also allows an expansion of the parameter set, to include other model parameters in addition to the OD flow variables. Simultaneous demand-supply calibration thus becomes possible.

The proposed approach does not rely on the costly computation, storage and inversion of assignment matrices. Simultaneous calibration of OD flows across multiple time intervals thus becomes practical, and allows the analyst to move away from the limitations of the sequential approach.

The remaining sections in this chapter discuss the solution of the off-line calibration problem formulated in previous sections.

3.5 Problem characteristics

The off-line DTA model calibration problem formulated in earlier sections possesses several characteristics that affect the effectiveness of different solution approaches. We discuss each of these factors, and draw conclusions that will subsequently assist in comparing existing optimization algorithms and choosing suitable approaches.

3.5.1 Large scale

A critical characteristic of the calibration problem is the size of $\boldsymbol{\theta}$, the vector of parameters over which the optimization must be carried out. The number of variables to be identified is strongly correlated with the physical extent and structure of the network being studied, as well as the desired temporal resolution for traffic dynamics modeling. Both aspects influence the number of OD pairs and links on the network, which directly impact the number of demand and supply parameters to be calibrated. While running time is not a primary concern given the off-line nature of the problem, the solution algorithm chosen to execute the calibration must still terminate in reasonable time.

3.5.2 Non-linearity

The off-line calibration problem is highly non-linear. The source of this non-linearity is embedded within the DTA simulator. The parameters in $\boldsymbol{\theta}$ pass through complex and non-linear transforms comprised of the various traffic dynamics models and algorithms embedded within the DTA system. For example, simulated vehicle speeds may be computed as

$$v = v_{max}\left[1 - \left(\frac{k - k_{min}}{k_{jam}}\right)^{\beta}\right]^{\alpha}$$

where v is the speed of the vehicle, v_{max} is the speed on the segment under free-flow traffic conditions, k is the current segment density, k_{min} is the minimum density beyond which free-flow conditions begin to break down, k_{jam} is the jam density, and α and β are segment-specific coefficients. In addition, the calculated speeds may be subject to a segment-specific minimum speed v_{min}.

The terms v_{max}, k_{min}, k_{jam}, α, β and v_{min} for each segment could be included in $\boldsymbol{\theta}$ as supply-side variables. Further, the density k is itself indirectly related to demand variables (OD flows, route choice model parameters) and non-linear route choice models. The functions $z_1(\bullet)$, $z_2(\bullet)$ and $z_3(\bullet)$ in the objective function are also typically non-linear, thus adding to the complexity.

The highly non-linear nature of the problem results in an objective function that potentially contains many local minima. Studying the true shape of the objective function is impractical, owing to the large number of parameters that must be analyzed. A further complication is introduced when the simulation model is stochastic, as function values will then be measured with error (the implications of stochasticity on the problem formulation and solution approach are discussed in Section 3.5.4).

3.5.3 Non-analytical simulator output

The dependence of the objective function on the output of a large-scale, non-linear simulation model implies that it is infeasible to obtain an explicit analytical expression for $f(\bullet)$ as a function of $\boldsymbol{\theta}$. Consequently, analytical derivatives with respect to

the parameters of interest are difficult to obtain. The implicit connection between the simulator output and the vector of unknowns suggests the appropriateness of optimization approaches that work with function evaluations (methods that essentially treat the simulation model as a black box).

3.5.4 Stochasticity

Traffic network quantities are stochastic, due to the presence of large numbers of individuals with a wide spectrum of unobserved characteristics, who are constantly making largely *independent* decisions that impact both local traffic conditions and the reactions and responses of surrounding drivers. A certain level of stochasticity is thus inevitable.

Sophisticated DTA systems whose constituent models reflect the complexity of the corresponding real-world processes, thus contain several sources of stochasticity. Simulated drivers' behavioral characteristics, specific departure times and habitual routes are sampled from pre-specified distributions. In addition, drivers' route choice decisions are modeled through probabilistic discrete choice models. Other sources of randomness in traffic simulation models are embedded within the algorithms that process vehicles and network elements, in order to replicate noisy system behavior. For example, stochasticity in trip travel times and network delays may be approximated through a randomization of the order in which incoming links at a node (or queued vehicles at an intersection) are processed in each simulation time step.

A stochastic simulator has implications on the problem formulation. In the absence of noise, replications of a simulation run yield identical outputs. It therefore suffices to directly compare the observed data to the output from a single simulation model run. With Monte-Carlo simulation, the comparison must account for the *distribution* of model output for a given set of input parameters. The function $\mathsf{f}(\bullet)$ in our case therefore represents the *mean* model output, obtained by averaging over

many simulation replications:

$$\mathcal{M}_h = \frac{1}{W} \sum_{i=1}^{W} f_i(\mathbf{x}_1, \ldots, \mathbf{x}_h; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_h; G_1, \ldots, G_h)$$

where W must be selected so as to obtain a pre-determined maximum allowable error d. The required sample size can be calculated using:

$$W = \left[z_{\alpha/2} \frac{\sigma}{d} \right]^2$$

where $z_{\alpha/2}$ is the critical value from the standard cumulative normal distribution, corresponding to a confidence level α, and σ is the standard deviation of the population. Since σ is usually not known, an iterative sample size determination based on the *sample* standard deviation may be adopted.

3.6 Review of optimization methods

Commonly adopted optimization methods rely on the availability of an explicit objective function. The use of a stochastic simulation model, however, renders such a function intractable. Most of the optimization literature is therefore not directly suited to the solution of the DTA model calibration problem, and the modeler must turn to simulation optimization methods (see Fu (2001) for a review of such techniques).

Optimization methods for large-scale, stochastic, non-linear problems can be classified into path search, pattern search and random search techniques. We review each class of methods next, and compare their advantages and limitations. For the purpose of this discussion, we adopt the following notation. An optimization problem of the following form is considered:

$$\underset{\boldsymbol{\theta}}{\text{Minimize}}\ z(\boldsymbol{\theta})$$

subject to the following constraints:

$$l \leq \boldsymbol{\theta} \leq u$$

In the above formulation, $\boldsymbol{\theta}$ is a K-dimensional parameter vector (to be calibrated), and is comprised of components θ_k. Lower and upper bounds on the parameters are specified through l and u, with $l = [l_1 l_2 \dots l_k \dots l_K]$ and $u = [u_1 u_2 \dots u_k \dots u_K]$.

3.6.1 Path search methods

Path search methods start at an initial parameter vector, and move in a certain direction with the aim of improving the value of the objective function. Often, the gradient of the function is used to determine the direction. Here, we review response surface methodology and stochastic approximation as two promising types of path search methods.

Response surface methodology

Response surface methodology (RSM) is a widely applied approach that (a) fits a local polynomial approximation based on the objective function values evaluated in the vicinity of the current parameter vector, (b) calculates the gradient of the resulting function, and (c) moves along the corresponding direction by some step size. The points at which the function must be evaluated are determined systematically, such as through an experimental design. Typically, a linear or quadratic response surface is chosen.

RSM can be implemented in two ways. In a sequential search, linear response surfaces are repeatedly applied until the objective function does not improve further. A quadratic surface is then fitted iteratively (and the parameters updated along the gradient-based direction) until the gradient estimate converges to zero. Higher-order approximations may be performed in a similar way, until the desired level of accuracy is achieved.

The second implementation of RSM involves meta-models: once the set of obser-

vation points around the current solution have been identified, the function values at these points are used to fit a response curve or meta-model. Deterministic optimization methods are then employed to generate the meta-model's gradient and update the parameter.

A detailed discussion of RSM may be found in Kleijnen (1987). While the statistical theory behind these algorithms is easily understood, they may be limited by the large number of function evaluations (model runs) needed for convergence. They have also been shown to be ineffective when the objective function contains sharp ridges and flat valleys.

A recently developed RSM-based algorithm that can potentially overcome these drawbacks, is SNOBFIT[1](Huyer and Neumaier, 2004). While this algorithm does not search for a path to the optimum in the strictest sense, it does fit a quadratic response surface at multiple points in each iteration. SNOBFIT maintains a population of points around which local quadratic surfaces are generated based on the function values of its nearest neighbors. The initial set of points is determined through a smaller optimization step that ensures a more uniform search of the feasible space (as defined by the lower and upper bounds). The algorithm is thus less likely to be affected by the choice of a starting point, and can also search over high ridges. The ability to control the relative local vs. global nature of the search (through a single tuning parameter) further increases the appeal of this approach and its ability to climb hills. At the end of each SNOBFIT iteration, a new set of points is recommended based on minimizations of the quadratic approximations. The function values are re-evaluated before proceeding to the next iteration.

The following steps broadly illustrate the underlying principles:

1. A set $\mathcal{S}$ of initial points $\boldsymbol{\theta}^s \ \forall \ s \in \{1, 2, \ldots, S\}$, their bounds $[l, u]$ and the function values $z(\boldsymbol{\theta}^s)$ are selected. The set dimension C denotes the number of points the modeler desires SNOBFIT to recommend in each pass. Typically, the points output by SNOBFIT in one iteration will form the input for the next iteration. At the start of the process, however, it is likely that only a single

[1]Stable Noisy Optimization using Branch and FIT

point $\boldsymbol{\theta}^0$ is available. In such an event, the algorithm will still recommend a set of S points that optimally span the feasible domain as defined by the bounds. The associated procedure is outlined in a later step of the algorithm.

2. The uncertainty $\Delta z(\boldsymbol{\theta}^s)$ in the function evaluated at $\boldsymbol{\theta}^s$, is determined. The corresponding vector Δs is used by SNOBFIT to capture model stochasticity. The elements of the uncertainty vector can easily be computed from multiple replications of the function at each point $\boldsymbol{\theta}^s$. A *resolution vector* $\Delta\theta$ with $\Delta\theta_k > 0 \; \forall \; k \in \{1, 2, \ldots, K\}$ is defined. Two points are considered different if they differ by more than $\Delta\theta_k$ in at least one of their K components.

3. The initial parameter bounds box is split into sub-boxes such that each sub-box contains exactly one point from $\mathcal{S}$. During subsequent iterations, every sub-box with more than one point is further divided through a branching procedure with the following steps:

   (a) Identify the parameter component k for which the variance of $\theta_k/(u_k - l_k)$ across all points in the sub-box is the maximum.

   (b) Sort the points such that $\theta_k^1 \leq \theta_k^2 \leq \ldots$

   (c) Split the k^{th} component at $\tilde{\theta}_k = \lambda\theta_k^{s^*} + (1-\lambda)\theta_k^{s^*+1}$, where $s^* = \arg\max(\theta_k^{s+1} - \theta_k^s)$ and λ is a function of the golden section number $\rho = \frac{1}{2}(\sqrt{5} - 1)$.

   Each sub-box is assigned a "smallness" measure that indicates the number of bisections of the full box $[l, u]$ required in order to obtain the current sub-box. Multiple sub-boxes may have the same smallness.

4. If the number of points currently in $\mathcal{C}$ is less than $(K + 6)$, there is insufficient information to attempt quadratic approximation. We therefore go to step 8.

5. A list of the $(K + 5)$ nearest neighbors is identified for each point $\boldsymbol{\theta}^s$. For each $k \in \{1, 2, \ldots, K\}$, the point closest to $\boldsymbol{\theta}^s$ among the points not yet in the list, and that also satisfy $|\theta_k^s - \theta_k^{s'}| \geq \Delta\theta_k$ is added to the list. The five closest points not yet in the list are added to make up $(K + 5)$ neighbors.

6. A local quadratic (Hessian) approximation model is fit around the current best point $\boldsymbol{\theta}^*$ and each of the remaining points, if their lists of nearest neighbors have changed.

7. A set of size S, comprised of five types of points, is now recommended:

   (a) A single type 1 point is identified by minimizing the quadratic approximation around $\boldsymbol{\theta}^*$, using a trust region of radius $\mathtt{d}$. A single type 2 point is generated by solving an identical problem with a radius of $\rho\mathtt{d}$. The radius is initialized to $\mathtt{d} = \frac{1}{4}(\mathtt{u} - \mathtt{l})$ for the first iteration, and is expanded or contracted in subsequent iterations based on comparisons of the realized function value at the current best point with those at the type 1 and type 2 points recommended at the end of the previous iteration. It should be noted that there is a possibility of a type 1 or type 2 point not being generated, if the calculated points differ from existing points by less than $\Delta\theta$.

   (b) Type 3 points are chosen from the minima (or valleys) of all remaining points (excluding the best point $\boldsymbol{\theta}^*$). Type 4 points are generated from as yet unexplored regions, by selecting points from sub-boxes of increasing smallness (denoting larger volumes) that have not already been used for type 3 selection. The split between types 3 and 4 is user-specified through $\mathtt{p}$, the probability of a type 3 point. If required, additional type 4 points may be generated at the end of the procedure (conditional on the availability of sub-boxes that have not yet been considered).

8. If the number of recommended points is still less than S, the necessary number of type 5 points are generated. A large set of random points are created, and a subset with the maximum distance from existing points is selected.

While a convergence rate result for SNOBFIT is currently unavailable, Huyer and Neumaier demonstrate that the algorithm is guaranteed to converge to the global minimizer $\boldsymbol{\theta}^*$ under the following set of assumptions: (a) the objective function $z(\boldsymbol{\theta})$

is continuous, at least in the neighborhood of $\boldsymbol{\theta}^*$, (b) $\Delta\theta = 0$, (c) the exploratory space $[l, u]$ is scaled to $[0, 1]^K$, (d) parameters are not rounded, implying (together with (b), that all points of types 3 and 4 are accepted, (e) at least one point of type 4 is generated in each iteration, and (f) the objective function is evaluated at the set of points recommended by SNOBFIT.

The authors list several merits of their approach, some of which are relevant for our problem setting. The algorithm works well with fewer function evaluations than existing methods, due mainly to its global search capability and intelligent local analysis. Moreover, the method explicitly accounts for noise in the function evaluations, which is attractive in the context of a stochastic simulator. The type 3 probability p may also be used to control the trade-off between local and global search. For example, requesting more type 4 points (with a probability of 1-p) allows SNOBFIT to spread out into unexplored regions of the domain.

The rigorous quadratic search employed by SNOBFIT is an obvious advantage, since the search intelligently handles the non-linearity in the objective function. An added advantage is the algorithm's memory: it "remembers" the points it already considered, and avoids recommending new points that are too close to them (the component-wise cut-off distance being a user-defined threshold).

A concern with applying SNOBFIT in situations with very wide parameter bounds is the initial effort spent on quadratic local fits on a randomly selected set of points. Indeed, if S is relatively small, the first complete set $\mathcal{S}$ recommended by SNOBFIT is likely to provide sub-optimal coverage of the feasible space. While a decision to increase S in such situations is justified, it might not be a practical option from a run-time perspective. This concern has potential implications on the applicability of this calibration solution method to large networks.

Stochastic approximation

Stochastic approximation (SA) traces a sequence of parameter estimates that converges to the zero of the objective function's gradient. The parameter updates at

each iteration i are generalized as:

$$\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^{i} - a^{i}\, \hat{g}(\boldsymbol{\theta}^{i}) \tag{3.11}$$

where $\boldsymbol{\theta}^{i}$ is the parameter vector at the beginning of iteration i; $\hat{g}(\boldsymbol{\theta}^{i})$ is a current estimate of the gradient; $\{a^{i}\}$ is a *gain sequence* of step sizes. The Robbins-Monro algorithm (Robbins and Monro, 1951) results when an unbiased estimator of the gradient is used to perform the parameter updates. However, such an estimator typically requires detailed knowledge of the model being calibrated, often generated through a perturbation analysis. Given the large-scale and stochastic nature of DTA models (and the variety of models in existence), such an exercise is expected to be unreliable and time-consuming.

When a finite-differences (FD) approach is used to approximate the gradient, the Kiefer-Wolfowitz algorithm (Kiefer and Wolfowitz, 1952) is obtained. Two popular FD schemes have been reported in practice:

$$\hat{g}_{k}(\boldsymbol{\theta}^{i}) = \frac{z(\boldsymbol{\theta}^{i} + c^{i}\, e_{k}) - z(\boldsymbol{\theta}^{i} - c^{i}\, e_{k})}{2c^{i}} \tag{3.12}$$

$$\hat{g}_{k}(\boldsymbol{\theta}^{i}) = \frac{z(\boldsymbol{\theta}^{i} + c^{i}\, e_{k}) - z(\boldsymbol{\theta}^{i})}{c^{i}} \tag{3.13}$$

where e_{k} is a K-vector with a one in the k^{th} location (and zeroes elsewhere), and $\{c^{i}\}$ is a decreasing sequence of small positive numbers. Equation 3.12 is known as a symmetric design, while Equation 3.13 is a one-sided or asymmetric design. The optimal asymptotic convergence rates have been shown to be $i^{-1/2}$ for the Robbins-Monro algorithm, and $i^{-1/3}$ or $i^{-1/4}$ for the Keifer-Wolfowitz algorithm (depending on the choice of one-sided or symmetric differences respectively)[2]. The faster convergence rate of Robbins-Monro has been attributed to its use of increased information through an estimate of the gradient, rather than a stochastic approximation based on noisy function evaluations (Spall, 1999).

It should be noted that a total of $2K$ function evaluations are required per sym-

[2]see Pflug (1996) for details.

metric gradient computation. The less-expensive asymmetric computation requires only $K+1$ function evaluations. In spite of the computational savings in using Equation 3.13 (relative to Equation 3.12), the effort per iteration increases linearly with the problem size K. This property is expected to negatively affect scalability, especially when the simulation model's running time is non-trivial. A single function evaluation with a mesoscopic traffic simulator can take several minutes, depending on the size of the network and the demand levels (total number of vehicles on the network at a given instant). For a case with a simulation time of 2 minutes and $K = 2000$, the time *per iteration* is as high as 67 hours, or nearly three days!

SPSA (simultaneous perturbation stochastic approximation, Spall (1992, 1994a,b, 1998b,a, 1999)) provides huge savings in *per iteration* cost, by approximating the gradient using just two function evaluations (independent of the value of K):

$$\hat{g}(\boldsymbol{\theta}^i) = \frac{z(\boldsymbol{\theta}^i + c^i\Delta_i) - z(\boldsymbol{\theta}^i - c^i\Delta_i)}{2c^i} \begin{bmatrix} \Delta_{i1}^{-1} \\ \Delta_{i2}^{-1} \\ \vdots \\ \Delta_{iK}^{-1} \end{bmatrix} \tag{3.14}$$

where Δ_i is a K-dimensional perturbation vector consisting of component-wise perturbations Δ_{ik}. Since the numerator in Equation 3.14 is invariant for all $k = 1, 2, \ldots, K$, the computational effort in each iteration is fixed (independent of K). This represents an obvious benefit for scalability, though numerical tests will be required in order to determine if the number of iterations to convergence is also reasonable. Nevertheless, from a theoretical perspective, SPSA represents a promising solution algorithm that must be further evaluated to determine its suitability to the off-line calibration problem.

The following steps describe the SPSA approach in detail:

1. The process is initialized ($i = 0$) so that $\boldsymbol{\theta}^i = \boldsymbol{\theta}^0$, a K-dimensional vector of *a priori* values. The SPSA algorithm's non-negative coefficients a,A,c,α and γ

are chosen according to the characteristics of the problem[3].

2. The number of gradient replications (`grad_rep`) for obtaining the average gradient estimate at $\boldsymbol{\theta}^i$ is selected[4].

3. The iteration counter is incremented: $i = i + 1$. The step sizes a^i and c^i are calculated as $a^i = a/(A + i)^\alpha$ and $c^i = c/i^\gamma$.

4. A K-dimensional vector Δ_i of independent random perturbations is generated. Each element $\Delta_{ik}, k = 1, 2, \ldots, K$, is drawn from a probability distribution that is symmetrically distributed about zero, and satisfies the conditions that both $|\Delta_{ik}|$ and $E|\Delta_{ik}^{-1}|$ are bounded above by constants. The literature indicates the suitability and success of the Bernoulli distribution ($\Delta_{ik} = \pm 1$ with equal probability). Note that the inverse moment condition above precludes the use of the uniform or normal distributions.

5. The objective function is evaluated at two points, on "either side" of $\boldsymbol{\theta}^i$. These points correspond to $\boldsymbol{\theta}^{i^+} = \boldsymbol{\theta}^i + c^i \Delta_i$ and $\boldsymbol{\theta}_{i^-} = \boldsymbol{\theta}^i - c^i \Delta_i$. Lower and upper bound constraints are both imposed on each point before function evaluation.

6. The K-dimensional gradient vector is approximated as

$$\hat{g}(\boldsymbol{\theta}^i) = \frac{z(\boldsymbol{\theta}^{i^+}) - z(\boldsymbol{\theta}^{i-})}{2c^i} \begin{bmatrix} \Delta_{i1}^{-1} \\ \Delta_{i2}^{-1} \\ \vdots \\ \Delta_{iK}^{-1} \end{bmatrix} \tag{3.15}$$

   The common numerator for all K components of the gradient vector distinguishes the SPSA approach from traditional FD methods.

7. Steps 4 to 6 are repeated `grad_rep` times, using *independent* Δ_i draws, and an average gradient vector for $\boldsymbol{\theta}^i$ is computed.

[3]Some guidelines for the selection of SPSA parameters are outlined in a later section.

[4]Gradient smoothing is believed to provide more stable approximations when function evaluations are noisy.

8. An updated solution point $\boldsymbol{\theta}^{i+1}$ is obtained through the application of Equation 3.11. The resulting point is again adjusted for bounds violations.

9. Step 3 is re-visited until convergence. Convergence is declared when $\boldsymbol{\theta}^{i}$ and the corresponding function value $z(\boldsymbol{\theta}^{i})$ stabilize across several iterations.

The convergence characteristics of SPSA depend on the choice of gain sequences $\{a^{i}\}$ and $\{c^{i}\}$, and the distribution of the perturbations Δ_{i}. Spall (1999) states that the two gain sequences must approach 0 at rates that are neither too high nor too low, and that the objective function must be several times differentiable in the neighborhood of $\boldsymbol{\theta}^{0}$. If these properties hold, and if the perturbations are selected according to the requirements stated in Step 4 above, then the author provides the best-case convergence rate as $i^{-1/3}$. The paper also claims the superior efficiency of SPSA compared to FDSA, with a K-fold saving per iteration.

Spall (1998a) provides some intuition for the success of SPSA, by comparing its performance with that of FDSA. When the objective function may be evaluated with little or no noise, FDSA is expected to emulate the steepest descent approach. Mathematical analysis has established this descent direction to be at right-angles to the contour line at each point. SPSA, owing to a random search direction, does not necessarily follow a true descent to the optimum. That is, the gradient approximations may differ from the true gradients, so that the corresponding search directions deviate from those of steepest descent. However, SPSA's gradient approximation is *almost* unbiased:

$$E[\hat{\mathbf{g}}(\boldsymbol{\theta}^{i})] = \mathbf{g}(\boldsymbol{\theta}^{i}) + \mathbf{b}^{i} \tag{3.16}$$

where the bias $\mathbf{b}^{i} = \mu c^{i^2}$, the term μ being a constant of proportionality. As $c^{i} \to 0$ (for large i), the small bias $\mathbf{b}^{i}$ vanishes. The path determined by SPSA is thus expected to deviate only slightly from that of FDSA. The "errors" in the search path will average out across several iterations, so that FDSA and SPSA converge to the same solution in a comparable number of iterations (Figure 3-3).

When the function evaluations are noisy, neither FDSA nor SPSA will trace the steepest descent directions with certainty. However, extending the previous analogy,

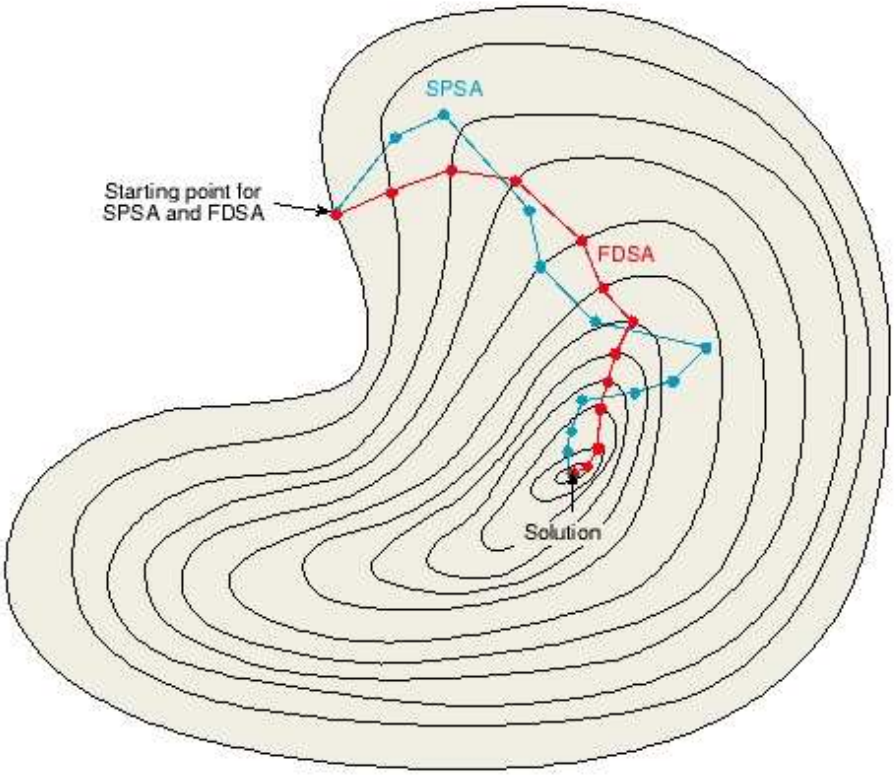

Figure 3-3: SPSA vs. FDSA [Spall (1998a)]

the SPSA directions on average remain close to the optimal directions.

The SPSA method outlined in this section immediately illustrates the potential computational savings for large-scale optimization. Unlike traditional stochastic approximation approaches, the effort expended per SPSA iteration is exactly 2 functional evaluations, and is *independent of the number of parameters,* K. The sequence $\{c^i\}$, if chosen carefully, can overcome the other limitation of FDSA, namely the unreliability of gradient approximations based on very small perturbations. Essentially, c^i must approach zero with i, but must do so at a "reasonable" rate. Rapid convergence of c^i may result in unreliable derivatives, while a slow rate may prevent SPSA from approaching the true gradient near the optimum.

3.6.2 Pattern search methods

Pattern search methods are often listed as direct search methods, since they do not require derivative calculations. Instead, they compare function values to determine the "best" direction of movement from the current solution.

Hooke and Jeeves method

This algorithm (Hooke and Jeeves, 1961) starts from an initial solution $\boldsymbol{\theta}^0$ and begins by determining a "good" search direction for $k = 1$, the first component of $\boldsymbol{\theta}^0$. This is achieved by moving by a certain search step size on either side of θ_1^0, and comparing the function values at both points (while maintaining the remaining parameter dimensions $k = 2, 3, \ldots, K$ constant). If an improved point is not obtained, the step size is reduced, and the search is repeated until a local descent direction is identified for $k = 1$.

An intermediate point is generated by moving along this direction (keeping all other components fixed). By definition, the objective function at the intermediate point is lower than the value computed at $\boldsymbol{\theta}^0$. The search for $k = 2$ starts from this new point, to generate a second intermediate point. The process is repeated until a search direction has been identified for each of the K components in the parameter vector. The iteration ends by collating all K search directions into a single direction, and applying an update that yields $\boldsymbol{\theta}^1$, the "solution" from the first iteration. Subsequent iterations repeat similarly, until some pre-defined termination thresholds are satisfied. Note that an improvement in objective function at each iteration is not guaranteed by this method. However, the iterations are expected to move towards a local optimum as $i \to \infty$. More details about this class of methods may be found in Kolda et al. (2003).

The independence from derivatives is conceptually a very attractive feature of pattern searches, since stochastic DTA models are generally expected to result in unreliable gradient estimates. However, the personal experiences of the author indicate that these methods perform poorly when applied to even small- and medium-sized

simulation optimization problems. The primary reason for failure is the focus on the immediate (component-wise) vicinity of the starting point, which is often far from a global optimum. A new pattern requires at least $K + 1$ function evaluations. When combined with a large-scale and highly non-linear function, the algorithm quickly restricts its search space to a very small radius around the current point (due to a rapidly shrinking search step size), thus making very slow progress even towards the nearest local optimum. Search methods possessing the capability to "jump" rapidly in a promising direction before refining their search, will therefore be better suited to the calibration context.

Nelder-Mead (Simplex) method

The Nelder-Mead method (Nelder and Mead, 1965) maintains a population of $K + 1$ points (called a simplex), and begins by computing the objective function at each point. At every iteration, the worst point is replaced by its reflection about the centroid of the remaining points in the simplex. The objective function is evaluated at the new point, and the process continues until little improvement can be achieved by eliminating the worst point. The simplex can shrink or expand depending on the objective function value of the new point, and ideally shrinks to a unique solution at convergence. The initial simplex may be set up by randomly generating $K + 1$ points that satisfy the lower and upper bounds.

Convergence results for the Nelder-Mead algorithm are scarce, and are largely limited to the one- and two-dimensional cases (see, for example, Lagarias et al. (1998)). This, and other papers, outline the difficulties in developing rigorous convergence analyses when $K > 2$, and provide empirical evidence in support of the assertion that the Nelder-Mead method can terminate at a sub-optimal point in noisy situations.

Box-Complex method

The Box-Complex algorithm (Box, 1965) is an extension to the Nelder-Mead approach: they both begin with a set of randomly-selected feasible points that span the search space. In each iteration, a candidate point (one with the highest objective

function value) is replaced by its reflection about the centroid of the remaining points. If the resulting point is worse than the candidate, the point may be moved closer to the centroid using some contraction scheme. The algorithms aim to move in a direction that eventually collapses the population into a single solution. The primary difference between the two methods lies in the definition of the size of the population: while Simplex requires exactly $K+1$ points in its set, Box-Complex requires a *minimum* of $K+2$. The use of a larger set can potentially increase the speed and accuracy of the search, and also guard against the possibilities of numerical instabilities with the Nelder-Mead approach (see Box (1965) for details).

Below we present an outline of the algorithm, tailored to our specific problem instance (a more elaborate treatment can be found in the original paper by Box):

1. A complex $\mathcal{S}$ of $S > K+1$ points is generated[5]. Each point $\boldsymbol{\theta}^s$ is a complete parameter vector of dimension K. Let θ_k^s denote the k^{th} component of the point $\boldsymbol{\theta}^s$, $s = 1, 2, \ldots, S$. The first point $\boldsymbol{\theta}^1$ (for s=1) is set to $\boldsymbol{\theta}^0$ (determined externally by the modeler), and is assumed to satisfy all the constraints. The remaining $S-1$ points are obtained one at a time, through the following procedure:

$$\theta_k^s = l_k + r_k^s \, (u_k - l_k) \tag{3.17}$$

   where r_k^s is a uniformly distributed random number in (0,1). The complex of S points thus spans the feasible space defined by the bound constraints.

2. The function value $z(\boldsymbol{\theta}^s)$ is evaluated at each point s in the complex. A point s' with the "worst" (largest) objective function value is determined[6], and is replaced by its reflection about the centroid of the remainder of the complex:

$$\tilde{\theta}_k^{s'} = \bar{\theta}_k + \alpha \, \left[\bar{\theta}_k - \theta_k^{s'}\right] \tag{3.18}$$

   where $\bar{\theta}_k = \frac{1}{S-1} \sum_{s \in \mathcal{S} \setminus \boldsymbol{\theta}^{s'}} \theta_k^s$ is the centroid for dimension k. If bound constraints

[5]The need for this constraint on the complex size is motivated in a later discussion.

[6]It should be noted that there can be multiple points with the same worst objective function value. One of these points is chosen at random for replacement.

are violated, the point is moved an infinitesimal distance δ within the violated lower or upper limit. The value of α is chosen to be greater than unity[7] which allows the complex to expand (if necessary), enabling rapid progress if the initial best solution is far from the optimum. Typically, $\alpha = 1.3$.

3. If a new (reflected) point repeats as the worst point on consecutive trials, it is moved one half the distance towards the centroid of the remaining points:

$$\tilde{\theta}_k^{s'} = \frac{\theta_k^{s'} + \bar{\theta}_k}{2} \ , \ k = 1, 2, \ldots, K \tag{3.19}$$

   Bound constraints are checked and adjusted as previously outlined. The selection of $\alpha > 1$ helps to compensate for such adjustments that shrink the complex.

4. The algorithm terminates when the function values of all points in the complex satisfy some pre-determined distance measure. For example, iterations might be stopped when all function values are within a certain percentage of each other across consecutive iterations.

As mentioned earlier, the size S of the complex must be at least $K+2$, where K is the number of parameters being estimated. When S equals $K+1$, the complex could potentially collapse into a subspace defined by the first binding constraint, preventing the exploration of other constraints. The Box-Complex method is therefore preferred over the Nelder-Mead method. Box (1965) suggests a practical setting of $S = 2K$.

The advantages of the Complex algorithm in the context of calibrating simulation-based functions are many. The non-dependence on numerical gradients has already been stressed as a key feature. The random starting complex used by the algorithm has the potential to move quickly towards the optimum before refining its search. In addition, the random initialization of the complex significantly increases the chance of converging to the *global* optimum, even if this is located "far" from the initial point $\boldsymbol{\theta}^0$ (since a subset of the starting complex is highly likely to be spread far away from

[7]Such an approach is known as over-reflection.

the starting solution).

A potential drawback of the Complex method is related to its focus on the worst point in the complex. While the algorithm repeatedly expends effort to improve the points with the highest objective function value, an improvement to the "best" point (with the lowest function value) is not guaranteed at every iteration. When combined with the fact that multiple function evaluations may be required per iteration (in case the worst point repeats), the algorithm tends to display extremely slow convergence rates as the optimization proceeds. Further, model stochasticity may result in an apparently worst point being eliminated from the complex, when it should have been retained.

3.6.3 Random search methods

Random search methods adopt probabilistic mechanisms to randomly select updated parameter vectors with the hope of improving towards an optimum. They are gradient-free, yet are characterized by a large set of tuning parameters that must be selected (often on a case-by-case basis). They are more suited to the context of discrete variable optimization over small search spaces. In the realm of large-scale continuous optimization, random search methods have displayed slow convergence (if they at all move towards the optimum). Simulated annealing and genetic algorithms are two common types of random searches, which we review here.

Simulated annealing

Simulated annealing (Metropolis et al., 1953; Corana et al., 1987) is the optimization equivalent of the physical process of cooling. The method begins with a very high temperature (chosen by the modeler), and attempts to reach the lowest possible temperature (of zero) just as heated metal cools towards room temperature. When metals re-crystallize from a high-energy state, their molecules might attain intermediate states of higher energy while they re-align themselves through an adiabatic (equilibrium) process. The optimization method follows an analogous "learning" process,

assigning a decreasing (non-zero) probability of traveling uphill while maintaining a generally downward trend in the objective function. The early iterations therefore allow for random "jumps" to escape from local optima.

Specific details of the implementation of simulated annealing methods vary widely in the literature. However, the need for the pre-selection of a large number of tuning parameters implies that significant effort may be required in identifying their optimal values for each application. These parameters include (a) the initial temperature, (b) the distribution of the perturbation applied to randomly generate updates, (c) the cooling schedule that determines the sequence of temperatures, and (d) the criteria for lowering the temperature (typically tied to the number of function evaluations of randomly perturbed parameter vectors allowed at each temperature setting). The work by Corana et al. (1987) is widely cited in the literature, and their tuning parameters are largely adopted in applications.

Simulated annealing has been found to be effective in combinatorial optimization problems, and has been widely applied in the area of electronic circuit design (Kirkpatrick et al., 1983). The primary advantage of the method is the ability to reach a global optimum, due to the high initial probability of visiting a much better solution by chance. The experience with continuous variables and large problems, however, is not encouraging. Wah and Wang (1999) adapt the basic algorithm (ascribed to Metropolis et al. (1953)), to the case of constrained optimization with continuous variables. They present many heuristics tailored to a set of benchmark problems. Goffe et al. (1994) demonstrate that the tuning parameters suggested by Corana et al. (1987) result in convergence for a 2-variable test case, after 852,001 function evaluations! The authors further propose modified parameters that reduce this to 3,789 evaluations, which while being a significant reduction, is still high for a small example. They note that the parameters in Corana et al. (1987) are conservative, but acknowledge that they would be better suited for highly non-linear and larger cases (indicating the method's lack of scalability).

Genetic algorithms

Genetic algorithms (GA) (Holland, 1975; Goldberg, 1989) are classified as evolutionary search methods, and are based on the theory of natural selection. A population of starting solutions (chromosomes or individuals) is generated at random, and their fitness (objective function values) evaluated. Solutions that are fitter (i.e. more likely to be closer to an optimum) are retained in the population with a higher probability, while the inferior points are eventually discarded. The population of current feasible solutions progresses through generations (iterations), with several operators acting on the chromosomes to decide the population that survives to the next iteration.

The fittest individuals in the current population are selected for starting a new generation. The chosen individuals are crossed over in pairs with some probability, and the resulting gene string may be further mutated to increase the randomness of the new population (to perturb the search in a hitherto unexplored direction). An elitist strategy might be enforced in order to retain the best solution(s) without change. Fitness is evaluated before repeating the process.

Genetic algorithms have been applied to solve small transportation optimization problems, including the calibration of a subset of microscopic traffic simulation model inputs (see, for example, Kim and Rilett (2004) and Henderson and Fu (2004)), however, they appear to be inferior to other methods reviewed earlier, for large-scale calibration. GA are naturally tailored for integer variables. In this regard, they differ fundamentally from other optimization methods that perform better in the continuous domain. This characteristic is primarily because variables' feasible values are coded as binary strings. Discretizing the feasible ranges of a large number of continuous variables (like the capacities and speed-density parameters encountered in DTA models) would thus result in numerous possibilities, depending on the step size chosen for this purpose. Even OD flows, if treated as integers, would result in an extremely large set of potential values due to their significantly wider bounds.

Secondly, the algorithm requires several parameters to be pre-defined, including crossover and mutation probabilities, a selection method and an elitist strategy.

The choice of these parameters has been found to be critical, as demonstrated by Mardle and Pascoe (1999). The authors report on the high sensitivity of GA to algorithmic implementations, and the large running times when compared to more traditional methods. They conclude with the recommendation that GA (and random search methods in general) be viewed as an option only when other algorithms fail. Henderson and Fu (2004) draw attention to the heuristic nature of GA, with no guarantee of convergence. Unless the algorithm's parameters are carefully selected for each specific application, solutions of poor quality may result due to premature "convergence".

Thirdly, Henderson and Fu (2004) review an application of GA for maximum likelihood estimation (Liu and Mahmassani, 2000), in which the effect of population size and the number of generations (iterations) is briefly explored. The study found that the two quantities varied greatly depending on the search space and the nature of the objective function, and that large populations are necessary if high levels of accuracy are desired. This conclusion is crucial in terms of scalability, especially when function evaluations are expensive.

3.6.4 Summary

A wide variety of simulation optimization algorithms exist in the literature. However, few have been tested on even medium-sized instances of non-linear optimization problems. While several approaches possess theoretically attractive properties desirable for the calibration of large-scale traffic simulation models, their performance (including both accuracy and running time) must be evaluated empirically before a suitable algorithm(s) can be identified. Based on the preceding analysis and some preliminary numerical experiments, three algorithms were short-listed for detailed testing:

- Box-Complex (pattern search)
- SNOBFIT (response surface method)
- SPSA (stochastic approximation - path search)

3.7 Solution of the off-line calibration problem

In this section, we discuss some issues related to the application of the three selected algorithms. First, we suggest the combination of Box-Complex and SNOBFIT to increase the efficiency of the optimization. We then outline some guidelines for the selection of algorithm parameters in the various methods.

3.7.1 Combined Box-SNOBFIT algorithm

We propose a solution approach that combines the Box-Complex and SNOBFIT optimization algorithms in a way that exploits their respective advantages. A review of the merits and potential drawbacks of the two algorithms suggests a natural integration scheme to obtain accelerated convergence and added estimation accuracy.

The Complex algorithm has the ability to cover the feasible space effectively, and rapidly replace the high-objective function points with estimates closer to the initial best point. However, the rate of convergence can drop significantly as it becomes harder to improve the worst point just through reflections about the centroid. SNOBFIT efficiently utilizes the information from every function evaluation to systematically search for local and global minima. However, it might benefit from a good set of starting points that reduce the need for costly quadratic approximations around remote points.

We propose a two-step approach in which the Complex algorithm first shrinks a randomly generated initial set of points to one that is more uniform in function values, without expending the computational resources needed to converge to the optimum. The result is expected to be a complex that is more representative of the various local minima of the non-linear objective function. In the second step, SNOBFIT uses the final complex as the starting set $\mathcal{L}$ to further refine the search through local approximation. A determination of the transition point (when the optimization switches from Box-Complex to SNOBFIT) may be made after reviewing the progress of the first method. A logical option would be to switch when the best objective function in the complex flattens out consistently across successive iterations.

While the per-iteration effort for SPSA represents a K-fold saving over FDSA or population-based methods such as Box-SNOBFIT, analyses of the total time to convergence must take into account the *number of iterations* required by each algorithm to satisfy the stopping criteria. While a determination of the true running-time savings will require extensive empirical work (see Chapter 4), it should be noted that the significant difference in per-iteration cost could, in theory, allow SPSA to perform many more iterations in the same (or less) time, to enable it to quickly identify solutions very close in quality to the final Box-SNOBFIT result.

3.7.2 Some practical algorithmic considerations

Both Box-SNOBFIT and SPSA use coefficients that determine the actual workings and performance of the respective algorithms. The choice of these coefficients significantly impacts their practical rates of convergence. In this section, we borrow from the literature and experimental experience to provide a brief review of the more critical algorithmic coefficients, along with guidelines for their initialization.

Selecting algorithmic coefficients for Box

Perhaps the most critical input to the Box algorithm is the size of the population (or complex) of points it maintains at each iteration. Theoretical considerations require that this be a minimum of $\mathsf{K}+2$ to ensure numerical stability. A less dense set risks a collapse of the complex along one or more dimension(s), with little chance of pulling away into better regions of the feasible domain. Based on experiments on fairly small problems, Box (1965) recommends that the complex size be twice or thrice the problem dimension K, which immediately raises the issue of scalability: the number of function evaluations required in order to set up the optimization iterations equals the size of the complex. It should be noted, however, that each subsequent iteration improves a single point in the set (one with the worst objective function). The corresponding computational effort is generally very small relative to the initial expense, though the per-iteration burden will typically increase sharply after many

iterations. This increase is attributed to the fact that it gets progressively harder to find, through a simple reflection operation about the shrinking centroid, a "better" replacement for the current worst point. In other words, the worst point under consideration may have to be reflected multiple times before its function value falls below that of the second-worst point. Indeed, this behavior is a key contributor to the drastic slow-down in Box's convergence rate after the complex has shrunk significantly in the early iterations.

Selecting algorithmic coefficients for SNOBFIT

Like the Box algorithm, SNOBFIT is also population-based. Therefore, the identification of an appropriate population size is again an important practical aspect. A hard constraint in this regard is the requirement that there be enough points in the population for SNOBFIT to perform its local quadratic fitting around each point. Since this step involves the five nearest points for each point under consideration, a minimum population size of $K + 6$ is necessary (Huyer and Neumaier, 2004).

As in any population-based method, maintaining a larger set of points enhances SNOBFIT's ability to locate a good solution. However, unlike the Box method, increasing the size of the population adds to the computational time in *every* iteration. This is because each SNOBFIT iteration recommends a new set of points (based on its quadratic minimizations) at which the function must be evaluated before progressing to the next iteration. From a scalability perspective, therefore, the minimum size of $K + 6$ is the most attractive setting.

Another key input to SNOBFIT is p, the probability of recommending a type 3 point[8]. A smaller p encourages SNOBFIT to explore hitherto unknown areas of the search space, in the search for a global solution. Larger values of p limit the search to the vicinity of the points at which the objective function is already known. $p = 0.3$ was found to work well for the calibration problem.

[8]Type 3 points are chosen from the list of local minima obtained through quadratic fitting around each point in the population. The best local minimum is designated as type 1, and is removed from contention for a type 3 label.

Selecting algorithmic coefficients for SPSA

SPSA is a random-search method, and does not maintain a population of points. This is perhaps the primary reason for its limited global optimization ability. However, the algorithm requires the initialization of other key constants: a, A, c and α and γ.

Experiments have confirmed that the values of $\alpha = 0.602$ and $\gamma = 0.101$ reported in the literature are adequate in the context of the calibration problem. Spall (1998b) also provides similar values associated with asymptotic optimality. The value of the "stability constant" $A = 50$ was also found to work well. The choice of the remaining two terms, however, can significantly impact the performance of SPSA.

The values of a and c control, respectively, the step sizes for parameter updating and gradient computation. If a is too large, SPSA may overlook a nearby solution and venture too far away. If a is too small, the algorithm may get stuck locally and never effectively search the surrounding space. Similarly, a large c may lead parameter component(s) to hit their bounds (almost) immediately, thus rendering the gradient approximations invalid. This problem would be magnified if the objective function is highly non-linear about the current point. On the other hand, a very small value for c could cause unreliable gradient approximations.

Empirical analyses revealed that suitable values for a and c may be identified by studying the magnitudes of the gradient approximations and subsequently limiting the desired updates to certain percentage of the magnitude of each parameter component.

3.8 Summary

In this chapter, typical DTA model calibration variables were described. The process of data collection was outlined, and the type of data assumed for this research was discussed. The off-line calibration problem was mathematically formulated, and the challenging problem dimensions analyzed. The mechanics of three non-linear optimization algorithms with properties suitable to the problem at hand were presented. An innovative global search approach integrating the Box-Complex and SNOBFIT algorithms was detailed. The SPSA stochastic approximation algorithm was intro-

duced as a scalable alternative for large networks. Some theoretical considerations for the empirical comparison of the computational performances of the two approaches was outlined, and practical guidelines for the selection of algorithmic terms and constants were outlined. The next two chapters focus on detailed case studies on two networks: a test case with synthetic data, and a much larger real network from Los Angeles, CA.

Chapter 4

Synthetic Case Study

Contents

This chapter describes a case study for the systematic evaluation of the estimator developed in Chapter 3. The broad objectives of the case study are outlined, the experimental setup and design described, and detailed numerical results presented and analyzed. Conclusions are drawn about the performance of the calibration approach, its ability to capture various transportation and traffic phenomena, and its scalability to real-sized datasets.

4.1 Objectives

The primary objectives of this case study are to:

- operationalize the calibration estimator proposed in Chapter 3, and demonstrate its ability to replicate traditional traffic sensor measurements.
- systematically evaluate the estimator for a variety of network conditions, on a prototypical network with known underlying parameters and processes.
- Compare the numerical accuracy and computational performance of the Box-SNOBFIT and SPSA algorithms.

4.2 Experimental setup

The setup for the evaluation of a DTA model calibration estimator requires three components: (1) an archived dataset of time-varying aggregate sensor measurements, (2) a DTA model with a set of inputs and model parameters, and (3) a network representation with a functioning surveillance system.

4.2.1 Sensor dataset generation

The generation of an archived sensor dataset must:

- allow for an objective and independent evaluation of the model being calibrated. The data generator must be able to model network processes and assumptions that are different than those used by the DTA model.

- realistically capture network demand patterns, driving behavior and supply phenomena that are consistent with real-world conditions.

- model a wide range of demand and supply conditions resulting in different traffic patterns.

- represent various surveillance system configurations, and record the types of sensor data available through field detectors.

Archived data for this case study were obtained through detailed microscopic simulations using MITSIM (Yang and Koutsopoulos, 1996; Yang et al., 2000).

4.2.2 Overview of DTA model and parameters

DynaMIT (Dynamic network assignment for the Management of Information to Travelers) is the DTA model chosen for demonstrating the methodology developed in this thesis. DynaMIT combines a flexible microscopic demand simulator and a mesoscopic supply simulator to effectively capture complex demand and supply processes and their interactions. Accurate modeling of origin-destination (OD) flows, pre-trip and en-route driver decisions, traffic dynamics, queuing and spillback allow the system to estimate and predict network state in a realistic manner. DynaMIT is designed to prevent overreaction by ensuring that the generated guidance is consistent with the conditions that drivers are expected to experience. This is achieved through explicit modeling of drivers' reaction to information. The flexible simulation system can adapt to diverse ATIS requirements, and is designed to handle a wide range of scenarios including incidents, special events, weather conditions, highway construction activities and fluctuations in demand. A detailed description of the features and functionalities of the DynaMIT system is provided in Ben-Akiva et al. (2001, 2002), and in Appendix A. We focus here on the DynaMIT inputs and parameters that must be calibrated.

Demand model parameters

DynaMIT's demand simulator is comprised of a route choice model and an OD estimation and prediction model. Driver route choices are captured through a Path-Size Logit model (Ramming, 2001):

$$P(i) = \frac{e^{V_i + \ln PS_i}}{\sum_{j \in P} e^{V_j + \ln PS_j}} \tag{4.1}$$

where $P(i)$ is the probability of choosing path i, V_i is the systematic utility of alternative i, PS_i is the size of path i, and P denotes the choice set (paths connecting the relevant OD pair). The utility V_i for each path is a function of attributes such as the travel time along the path. The coefficient of travel time is a calibration variable.

DynaMIT employs a sequential OD estimation module, based on transition and measurement equations (see Equations 2.9 and 2.10). An input to this model is the set of historical time-dependent OD flows (to be calibrated off-line), which will be updated with the latest sensor count measurements as they become available. The OD estimation and prediction algorithms are based on an autoregressive process that captures (through the transition equation) spatial and temporal correlations between the OD flows. We list below the key calibration parameters for the OD estimation and prediction model:

- The historical OD flows, $\mathbf{x}_h^H$.
- The variance-covariance matrix $\mathbf{V}_h$ associated with the indirect measurement errors.
- The variance-covariance matrix $\mathbf{Q}_h$ associated with the direct measurement errors.
- The matrices $\mathbf{f}_h^p$ of autoregressive factors.

Supply model parameters

DynaMIT's mesoscopic supply simulator captures traffic dynamics and models the build-up and dissipation of queues and spillbacks. The links in the network are subdivided into segments to capture changing section geometries. Each segment contains a moving part (with vehicles moving at a certain speed), and a queuing part. The speeds of vehicles in the moving part are governed by macroscopic speed-density relationships:

$$v = v_{max}\left[1 - \left(\frac{k - k_{min}}{k_{jam}}\right)^{\beta}\right]^{\alpha} \tag{4.2}$$

where v is the speed of the vehicle (in mph), v_{max} is the speed on the segment under free-flow traffic conditions, k is the current segment density (in vehicles/mile/lane), k_{min} is the minimum density beyond which free-flow conditions begin to break down, k_{jam} is the jam density, and α and β are segment-specific coefficients. In addition, the speeds computed using Equation 4.2 are subject to a minimum speed v_{min}.

The movement of vehicles from one segment to the next is governed by capacity calculations. The primary quantities of interest are the segment capacities which together with the available physical space on the downstream segments, determine the ability of vehicles to progress downstream. A constraint on either capacity or space would cause vehicles to queue. An important calibration step is therefore the computation of segment capacities that are consistent with prevailing traffic conditions. We list below the key calibration variables on the supply side:

- Segment speed-density parameters (v_{max}, k_{min}, k_{jam}, β, α and v_{min}).
- Segment capacities on freeway and arterial segments.
- Capacity factors that determine capacities on segments affected by incidents.

4.2.3 Network description and calibration variables

The prototypical network (Figure 4-1) was a directed graph consisting of 8 nodes and 8 links. Each link possessed a uniform cross-section along its length, and was divided

into 3 segments in order to better represent the evolution of congestion. Demand was assumed to flow between all three feasible OD pairs (connecting the three origin nodes O1, O2 and O3 to the destination node D). Travelers making trips between O1 and D choose between two alternative paths, while the remaining two OD pairs were each captive to a single path. Link geometries reflect varying numbers of lanes, as indicated in the figure.

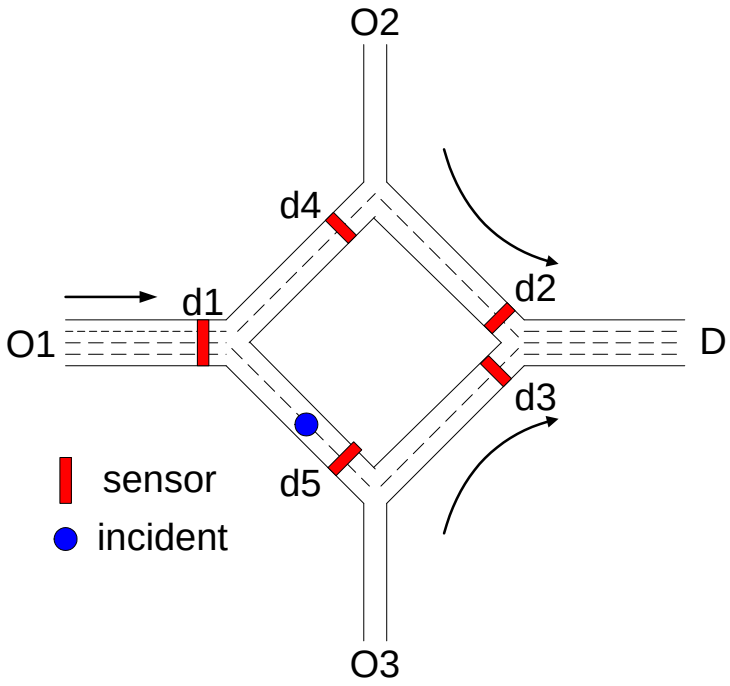

Figure 4-1: Prototypical Network

The simulation period of interest spanned 50 minutes (6:50 – 7:40 AM), which was further divided into 5-minute intervals (so that the number of intervals $H = 10$). An incident was assumed to impact the available capacity on the network for a period of 10 minutes, beginning at 7:05 AM. The location of the incident is indicated in the figure through a filled circle. One out of two available lanes at the incident location was assumed to be blocked during the incident. Drivers traveling between O1 and D who chose to travel through the affected link were likely to experience incident delays, depending on their departure times and prevailing traffic conditions. A total of 5 link-wide traffic sensors (indicated by boxes, and labeled d1 through d5) provided

count and speed observations.

The calibration variables for this case study included both demand and supply parameters. The demand variables consisted of time-dependent flows for each of the three OD pairs, and a travel time coefficient in the route choice model. OD profiles were discretized into 5-minute time slices, yielding $\frac{[50 \text{ min}]}{[5 \text{ min}]} \times [3 \text{ OD pairs}] = 30$ flow variables. A total of 31 demand variables (including the route choice coefficient) were thus estimated.

The supply variables consisted of segment exit capacities, segment speed-density relationship parameters, and a capacity reduction factor during the incident. All supply parameters were assumed to remain constant throughout the 50-minute period. The 24 segments were classified into three groups based on the number of lanes in their cross-sections, and a single speed-density function (with 6 parameters) was estimated for each group. Consistent with the above assumptions, a total of $[24 + (6 \times 3) + 1] = 43$ supply variables were estimated.

4.3 Base case analysis

A base case was defined with the primary objective of demonstrating the feasibility and validity of the methodology developed in Chapter 3. Four MITSIM factors (designated by the letters A, B, C and D, and representing both demand- and supply-side effects) were selected, and their values chosen to reflect typical behaviors and processes observed in prior studies.

The chosen factors and their settings are illustrated in Table 4.1. Factor A represents a demand-side effect that captures the sensitivity of drivers' route choice decisions to the perceived travel times along alternative routes. The factor controls the travel time coefficient in the route choice model. A value of -0.03/minute was selected based on earlier findings (see, for example, Ramming (2001)).

Factor B specifies the spatial distribution of demand between the three OD pairs. The time-dependent profiles of the main OD flow (between nodes O1 and D) and the two side flows were selected so as to generate visible congestion due to the incident,

A Route Choice	B OD: Spatial	C OD: Temporal	D Desired Speed
-0.03 /min	balanced	low-variance	typical

Table 4.1: Base Case Factor Settings

while simultaneously capturing merging and weaving phenomena between the main flow and each of the minor flows (Figure 4-2).

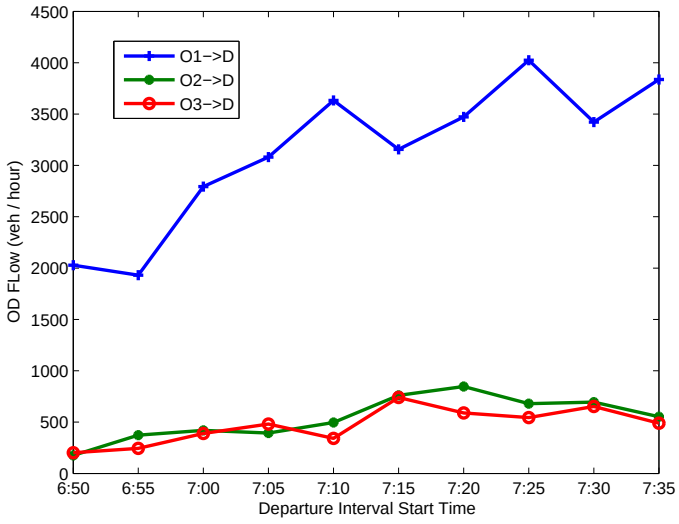

Figure 4-2: Base Case: Historical OD Flow Profiles

Factor B may be perceived as the *historical* demand between each OD pair. The actual demand levels realized on a particular day are thus drawn from a distribution with mean flows equal to the historical values. The spread of this distribution is controlled through factor C, which represents the temporal (within-day) variability in demand. The base setting for factor C allowed variations of up to 25% on either side of the mean. It was assumed that the corresponding noise was independent across time intervals.

Factor D corresponds to the distribution of desired speeds across the population of drivers. This factor represents a supply-side effect that impacts vehicle interactions, weaving behavior, traffic dynamics and network capacities in MITSIM. A typical setting for this factor was chosen from prior calibration exercises based on real data.

The dataset for the base case was simulated using MITSIM. Input files consistent with the chosen factor and incident settings were created. Sensor count and speed data from a single MITSIM run were recorded, representing a realization of the underlying stochastic demand and supply processes.

4.3.1 Estimators

Five estimators were studied under the base case:

- Estimator **Ref**: The supply-side variables were calibrated using existing approaches, while the demand-side parameters (the OD flows and route choice coefficient) were constrained to their known "true" values. Ref thus represents the best fit that can be obtained using current methods, and serves as the reference for evaluating the performance of our proposed estimators. Further, the supply parameter estimates from Ref serve as the starting point for the subsequent estimators.

  Segment capacities for Ref were estimated based on the number of constituent lanes and intersection signal timing plans, using the recommendations of the Highway Capacity Manual (TRB, 2000). The residual incident capacity was approximated through knowledge of the number of affected lanes (and the total number of lanes), based on the Highway Capacity Manual. Speed-density functions for each of the three segment types were identified by independently fitting sensor speed-density data. The data from all sensors belonging to a particular group were pooled for this purpose.

- Estimator **S(c)**: The unknown parameter vector was restricted to the 43 supply variables, while keeping the demand variables fixed at their known true values (an approach similar to Ref). However, our solution approach was applied to identify the optimal supply parameters at the network level. While the dataset contained both count and speed observations, only the count measurements were used to solve the optimization problem.

- Estimator **S(cs)**: The vector of unknown parameters was identical to S(c). However, both count and speed data were used to refine the supply parameters obtained in Ref. S(cs) may be viewed as estimator S(c) with added information.

- Estimator **SD(c)**: The set of variables was augmented by treating the 31 demand parameters as additional unknowns, resulting in the estimation of [43 + 31] = 74 parameters. The number of degrees of freedom in the optimization problem were thus increased. Only count data were included in the objective function.

- Estimator **SD(cs)**: The entire set of variables (encompassing both demand and supply models, as in SD(c)) was estimated, using count and speed data. This estimator represents the best possible situation, that utilizes all the available information.

4.3.2 Measures of performance

The root mean square error (RMSE) and root mean square normalized error (RMSN) statistics were used to measure the performance of the estimators in replicating the observed data:

$$\mathsf{RMSE} = \sqrt{\frac{\sum_{i=1}^{S}(y_i - \hat{y}_i)^2}{S}} \tag{4.3}$$

$$\mathsf{RMSN} = \frac{\sqrt{S\sum_{i=1}^{S}(y_i - \hat{y}_i)^2}}{\sum_{i=1}^{S} y_i} \tag{4.4}$$

where y_i is the i^{th} observed measurement (or "true" parameter value), and $\hat{y}_i$ the corresponding simulated (or estimated) quantity. The measure of fit to observed sensor counts and speeds were designated as RMSE^c and RMSE^s respectively. The corresponding normalized statistics were designated as RMSN^c and RMSN^s. In each case, the value of S in Equation 4.3 was the number of observations used in calculating the statistic.

RMSE^d and RMSN^d were employed to evaluate the fit of the estimated OD flows

against the true flows for the two SD estimators (where OD flows were part of the unknown parameter vector). The route choice coefficient was compared against its "true" value. The stability of the supply parameter estimates was studied to ensure that the proposed approach yielded reasonable speed-density functions and capacities.

Statistics for the Ref estimator were computed for comparison purposes. The fit to counts was represented by $RMSE^c = 17.19$ ($RMSN^c = 0.0943$), while the fit to speed data evaluated to $RMSE^s = 3.85$ ($RMSN^s = 0.0943$).

4.3.3 Numerical results using Box-SNOBFIT

Tables 4.2 and 4.3 summarize the RMSE and RMSN statistics for the base case. Numbers in parentheses denote percent improvement over Ref. It should be noted that the percent improvements for RMSN statistics are identical to those for the RMSE numbers, since RMSN represents a simple normalization of RMSE using the mean of the measurement of interest (refer Equations 4.3 and 4.4).

The statistics for the base case indicate that the proposed methodology improves DynaMIT's ability to replicate the field data. Indeed, the $RMSE^c$ and $RMSE^s$ values for all four S and SD estimators are consistently lower than those for Ref. The enhanced performance is further highlighted by the significantly better fit to speeds and OD flows obtained when sensor speed data are incorporated into the objective function. The motivation for network-wide calibration is thus reinforced, especially for the supply variables that are fit only locally at sensor locations in Ref.

Estimator	Calibration Data					
	Counts (c)			Counts + Speeds (cs)		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
S	15.89 (7.6)	2.86 (25.7)	-	16.86 (1.9)	2.29 (40.5)	-
SD	15.87 (7.7)	3.02 (21.6)	2.50	16.69 (2.9)	2.24 (41.8)	2.11

Table 4.2: Base Case RMSE Statistics: Box-SNOBFIT

Tables 4.2 and 4.3 further confirm expected trends across the five estimators. Adding degrees of freedom, for example, results in better fit to counts. A similar improvement in the fit to speeds is observed across S(cs) and SD(cs). Moreover, the

Estimator	Calibration Data					
	Counts (c)			Counts + Speeds (cs)		
	$RMSN^c$	$RMSN^s$	$RMSN^d$	$RMSN^c$	$RMSN^s$	$RMSN^d$
S	0.0872	0.0700	-	0.0925	0.0561	-
SD	0.0871	0.0740	0.0214	0.0916	0.0549	0.0183

Table 4.3: Base Case RMSN Statistics: Box-SNOBFIT

introduction of constraints (in the form of speed observations, in the S(cs) and SD(cs) estimators) helps fit the speeds better at the expense of the fit to counts. While the loss of fit to counts is not large, the reduction in $RMSE^d$ indicates that better OD flows have been identified through the information contained in speed data. The $RMSE^c$ value is still below Ref levels, underscoring the overall advantages of the new approach.

The estimated demand variables in the two SD estimators compare favorably with the assumed "true" values, as illustrated by the low error in fitting OD flows as well as the visual comparisons in Figures 4-3 and 4-4. In addition, simultaneous OD estimation using the traditional assignment matrix formulation[1] and Ref supply parameters resulted in $RMSN^d = 0.0335$, underlining the ability of the proposed methodology to more accurately capture demand patterns. The travel time coefficient of the route choice model was estimated as -0.0291/minute and -0.0301/minute for SD(c) and SD(cs) respectively, representing errors of less than 3% with respect to the true value of -0.03/minute.

The speed-density parameters estimated for each of the three segment groups are presented in Tables 4.4 through 4.6[2]. It is verified that the parameter estimates are stable across all five estimators, and also for all three segment groups.

The calibrated DynaMIT accurately captured the impact of the incident that disabled one of two lanes at the affected location. The estimated segment capacity during the incident was found to be less than half of the original capacity, consistent with

[1]Note that the popular sequential OD estimation approach fails in this example, since the travel time between any origin node and sensor is greater than the departure interval width of five minutes. Thus few, if any, of the vehicles are counted by sensors during their respective departure intervals.

[2]v_{max} and v_{min} are speed estimates measured in miles/hour. k_{jam} and k_{min} represent densities in vehicles/lane-mile. α and β are parameters.

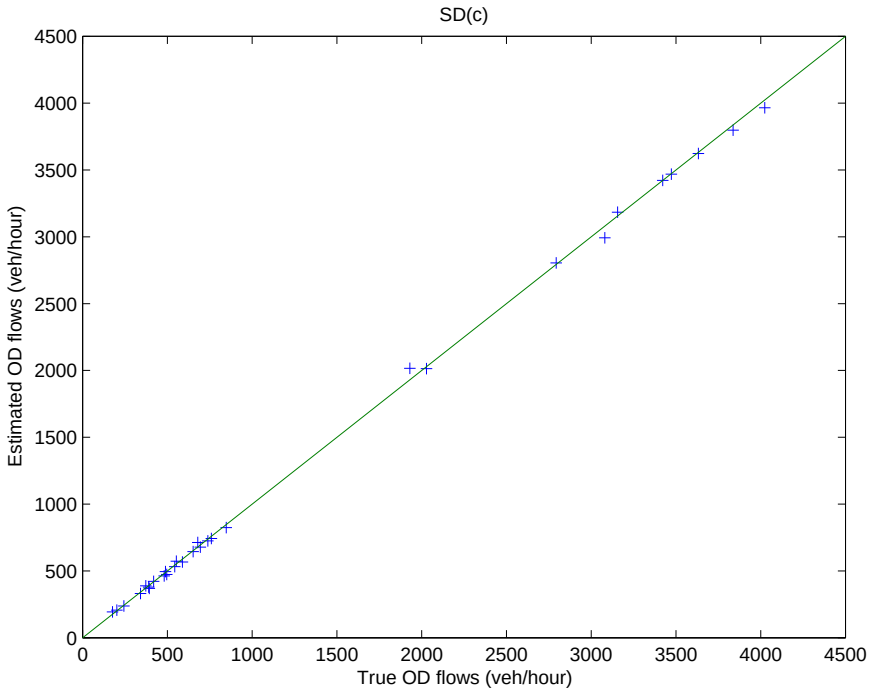

Figure 4-3: Fit to OD Flows (using only counts)

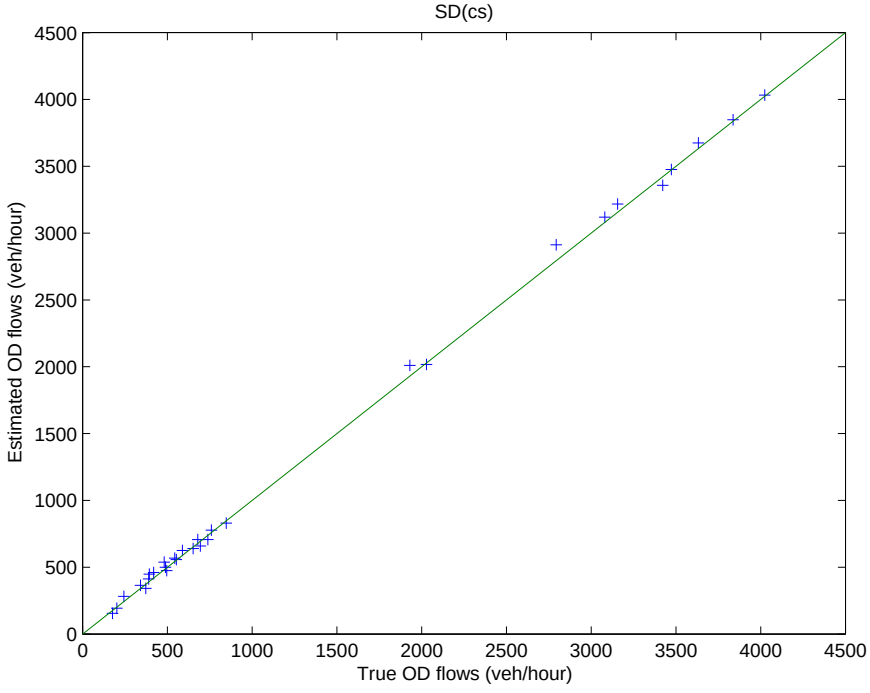

Figure 4-4: Fit to OD Flows (using counts and speeds)

	Ref	S(c)	S(cs)	SD(c)	SD(cs)
v_{max}	51.75	49.90	46.85	49.86	47.74
v_{min}	10.00	10.95	8.46	9.61	9.63
k_{jam}	527.5	501.0	513.3	441.8	507.2
k_{min}	6.8	3.3	9.5	5.1	6.3
α	0.2871	0.2532	0.2656	0.2761	0.2793
β	0.2464	0.2720	0.2804	0.2717	0.2713

Table 4.4: Speed-Density Parameters: Group 1

	Ref	S(c)	S(cs)	SD(c)	SD(cs)
v_{max}	52.06	49.92	51.86	48.92	50.23
v_{min}	10.00	11.36	11.20	10.52	7.50
k_{jam}	500.4	472.0	467.0	476.4	497.1
k_{min}	4.5	5.7	1.8	3.1	2.5
α	0.3779	0.3687	0.3423	0.3413	0.3711
β	0.2803	0.2786	0.2062	0.3215	0.2683

Table 4.5: Speed-Density Parameters: Group 2

	Ref	S(c)	S(cs)	SD(c)	SD(cs)
v_{max}	52.06	51.36	47.74	48.99	51.15
v_{min}	10.00	9.47	12.40	11.89	8.03
k_{jam}	500.4	461.0	473.7	459.0	446.4
k_{min}	4.5	5.2	4.6	3.6	5.2
α	0.3779	0.3448	0.3042	0.3372	0.3516
β	0.2803	0.3158	0.3656	0.3058	0.3049

Table 4.6: Speed-Density Parameters: Group 3

the additional loss of throughput due to the forced merging of vehicles (in MITSIM) immediately upstream of the affected lane.

4.4 Sensitivity analysis

A detailed sensitivity analysis was adopted in order to systematically study the performance of the solution methods developed in Chapter 3 under various demand and supply situations. Four factors were identified for this purpose. The chosen factors, their levels and the selected experimental runs are discussed next.

4.4.1 Factor levels and runs

The four factors A, B, C and D are each represented by three levels, denoted by the integers -1, 0 and +1. Further, we reference a factor and one of it's levels by combining the factor's letter code and the level's integer code. Thus, A(0) represents the level-zero setting for factor A. Other factors and their levels are similarly referenced.

Table 4.7 summarizes the four factors and their settings. Factor A captures the sensitivity of drivers' pre-trip route choice decisions to expected network travel times. The levels of this factor correspond to a range of possible behavioral patterns, represented through the travel time coefficient in the route choice model. A(0) = -0.03/minute represents typical parameter values estimated in earlier route choice studies (see, for example, Ramming (2001)). A(+1)=-0.15/minute captures nearly deterministic behavior, with drivers strongly favoring the path with minimum delays. Such a hypothesis tends to magnify even small differences in travel times between the different paths. A(-1) = -0.01/minute captures the other extreme, when drivers display more "balanced" behavior by being indifferent to significant travel time differences among the available alternative paths.

Factor B controls the the spatial distribution of network demand, through the relative magnitudes of the main and side OD flows. B(-1) combines a low main flow (between O_1 and D) with high side flows. Under this setting, a small number of vehicles pass through the incident (and are impacted by the associated delays), while

Factor	Levels		
	-1	0	+1
Route Choice (A)	-0.01 /min (time-insensitive)	-0.03 /min	-0.15 /min (deterministic)
OD: spatial (B) (historical)	main flow: lower side flows: higher	balanced	main flow: higher side flows: lower
OD: temporal (C)	hist. flow (no variance)	hist. flow + low-variance error	hist. flow + high-variance error
Desired Speed (D)	slower	typical	faster

Table 4.7: Factors and Levels

experiencing more interactions with the side flows originating at nodes O_2 and O_3. B(+1) increases the main flow (and the impact of the incident) while reducing the merging interactions from the two minor OD pairs. B(0) provides an intermediate setting.

Factor B also specifies the historical (or average) dynamic demand profile that generates the transportation demand underlying the network. Actual demands are assumed to be instances of a stochastic process that adds a uniformly distributed error term to the historical flows. The variance of this error term is the focus of factor C. Three variance levels are studied. C(-1) corresponds to a zero-variance situation, when the actual demand profile on a given day is identical to the historical profile specified through factor B. C(0) and C(+1) introduce some temporal noise through low- and high-variance perturbations to the historical flows.

Factor D represents supply-side effects that impact vehicle speeds and network capacities. We focus on individual drivers' desired speeds, which play a role in determining their actual speeds as well as their interactions with surrounding vehicles. The levels for this factor were obtained by controlling the distribution of desired speeds in MITSIMLab. D(-1), D(0) and D(+1) capture increasing mean desired speeds.

Table 4.8 summarizes the nine runs constituting the analysis. Run 1 represents "typical" settings that may be expected in reality, and is identical to the base case described earlier. Each of the subsequent runs are obtained by varying one factor at a time to its two extreme settings of -1 and +1.

The dataset for each run in the sensitivity analysis was simulated using the pro-

Run	A	B	C	D
1	0	0	0	0
2	-1	0	0	0
3	+1	0	0	0
4	0	-1	0	0
5	0	+1	0	0
6	0	0	-1	0
7	0	0	+1	0
8	0	0	0	-1
9	0	0	0	+1

Table 4.8: Sensitivity Analysis Runs

cedure adopted for the base case. A separate set of input files was created for each run, consistent with the individual factor settings outlined in Table 4.8.

4.4.2 Numerical results

The RMSE statistics reveal that the calibrated DynaMIT simulator accurately replicates the observed count and speed data under a variety of demand and supply conditions. Detailed tables documenting the numerical results are presented in Appendix B and Figures 4-5 through 4-9 summarize the RMSE statistics.

Figure 4-5 presents the fit to counts for scenarios S and SD, when only the counts subset of the sensor data is used for calibration. Scenario S consistently fits the counts better than the reference case. As expected, the SD scenario further improves the fit, owing to the increase in degrees of freedom arising from estimating OD flows and the route choice coefficient.

A comparison of the count RMSE statistic identifies runs 1 (base case), 3 (deterministic route choice) and 8 (slower-moving vehicles) as particularly complicated, with fit that is worse than that of the remaining runs. Among these, the base case is expected to be challenging (the corresponding factor settings having been chosen to be representative of actual conditions experienced on real networks). Deterministic route choice (run 3) results in drivers always choosing the shortest route to their destinations. This can increase congestion and add to the complexity, as the side flows will cause a greater impact when merging with the main flows. Lower desired speeds

(run 8) imply that drivers stay on the network longer than in the base case. The potential for interactions with other drivers increases, and delays due to the incident impact downstream sensors more than in the base case (owing to the longer travel times).

The fit to counts when speed observations are added to the objective function, is presented in Figure 4-6. Scenario SD again shows more accurate replications of the observed counts. However, the reference case is found to fit the counts better in certain cases (runs 2 and 5). This may be explained by viewing the speed measurements as added constraints to the counts-only situation, which could narrow the set of feasible solutions. The benefit of adding speeds to the dataset is seen shortly.

Figure 4-7 outlines the fit to observed speeds, when only the count data are used for estimation. Such a comparison contains limited information about the estimator's ability to fit observed data, since the algorithm does not use the speeds to refine the parameter estimates. However, the figure confirms that the counts are not over-fit at the expense of the speeds. Rather, the fit to speeds often improves from the base case (though there is no systematic pattern when it does). These improvements may be attributed to small adjustments to the supply parameters based on the traffic information contained in the counts (supply parameters control the speed-density function, which sets vehicle speeds according to prevailing traffic densities).

Figure 4-8 provides a more meaningful illustration of the impact of speed data. As expected, estimator S provides reductions in the speed RMSE (when compared to the reference case), with SD further improving the speed statistics. The significantly better fit to speeds comes at the expense of the count statistics, though the loss in fit to counts (see Figure 4-6) is minimal.

It is observed that the improvement over S due to estimating demand parameters in addition to supply parameters (SD) is often minimal. This can be ascribed to the use of the "true" OD flows while estimating S (and in Ref). Further, the small number of OD pairs in the synthetic network limits the potential impact of demand calibration (the role of demand is illustrated in the next chapter, using a real network).

A key advantage of this evaluation methodology is the availability of "true" pa-

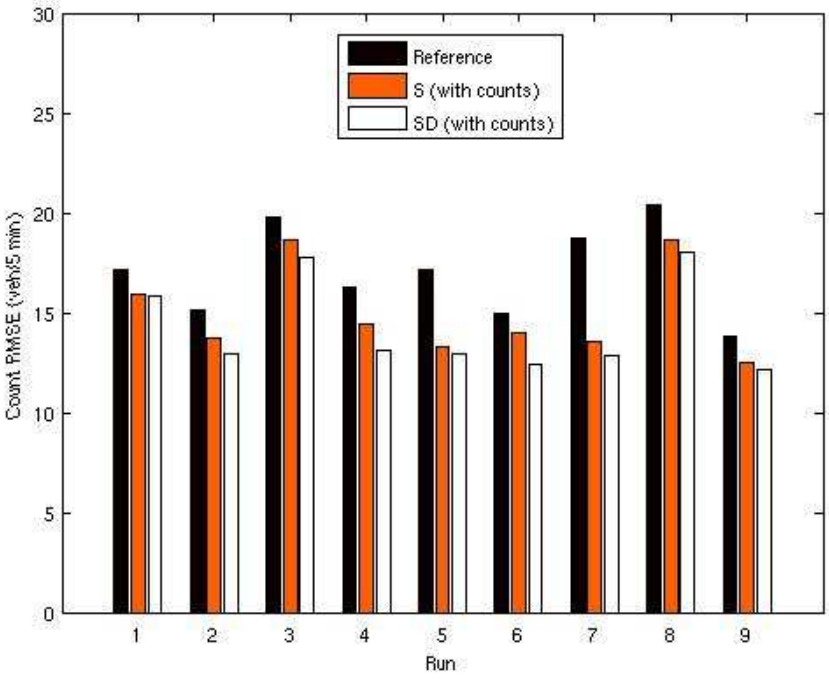

Figure 4-5: Fit to Counts Using Only Counts

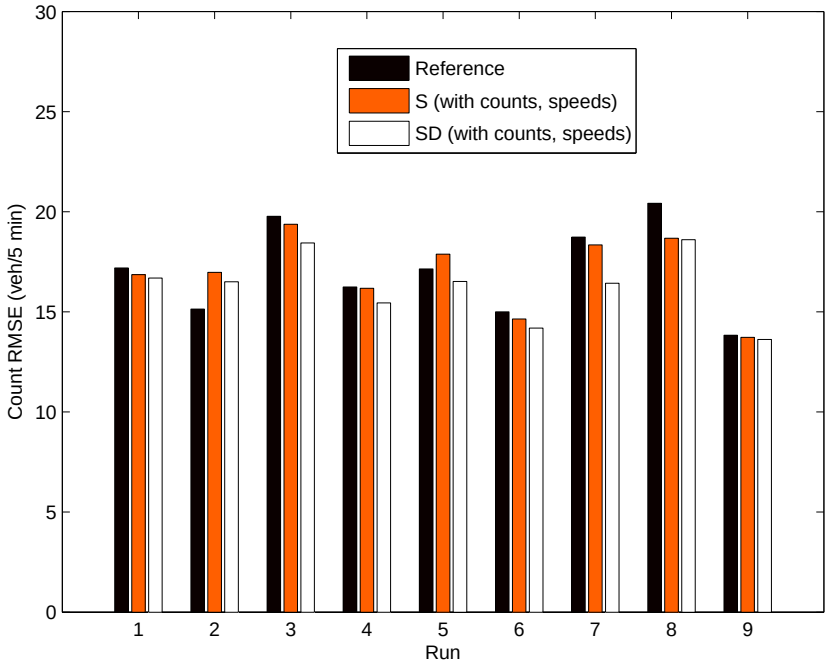

Figure 4-6: Fit to Counts Using Counts and Speeds

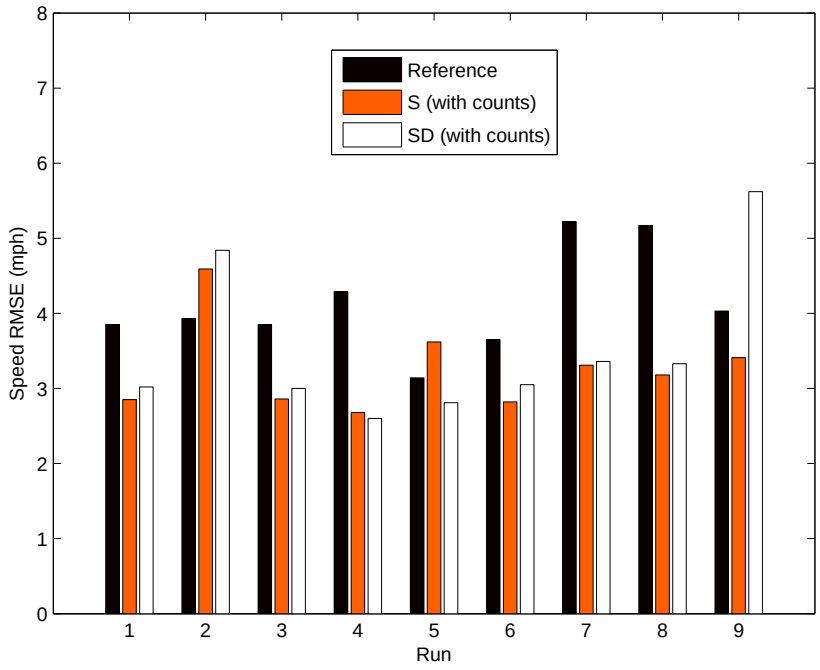

Figure 4-7: Fit to Speeds Using Only Counts

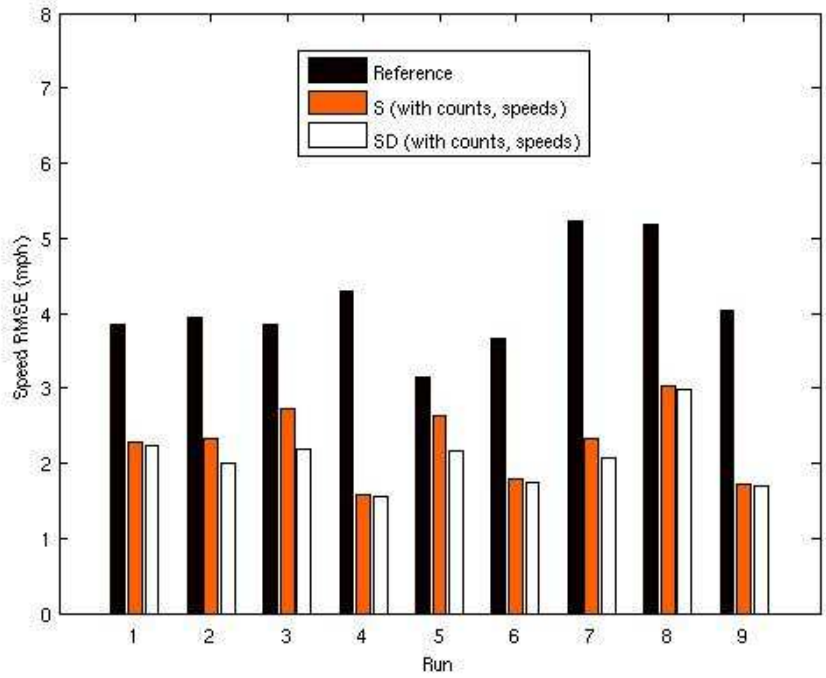

Figure 4-8: Fit to Speeds Using Counts and Speeds

rameter values, which are typically lacking in real datasets. Figure 4-9 focuses on the ability of our estimators to replicate the underlying "true" OD flows (in scenario SD). The RMSE statistic was computed using all three OD pairs, across the ten intervals. The figure indicates good fit to known demand profiles, and indicates that the inclusion of speed information systematically improves the quality of the OD flows. Runs 7 (high interval-over-interval variability) and 8 (slower drivers) display relatively poorer statistics, as may be expected: greater fluctuations across intervals create a harder estimation problem, since the temporal transition of OD flows is not smooth. The algorithm must therefore expend additional effort in each interval, while simultaneously accounting for the vehicle contributions from prior departure intervals. Slower drivers remain on the network longer, and impact sensor counts at more sensor locations and time intervals.

4.4.3 Conclusions and further analysis

It can be concluded, from the discussion in the previous section, that each factor contains one level setting that results in a challenge for calibration. These cases correspond to A(+1), B(0), C(+1) and D(-1), corresponding to deterministic route choice behavior, "balanced" spatial demand distribution, high-variance temporal fluctuations and slower drivers. Similarly, the situations of *least resistance* are A(-1), B(-1), C(-1) and D(+1), corresponding to time-insensitive route choice decisions, lower main flows, no temporal variability and faster drivers. The extreme ("worst" and "best") levels were combined to form two additional runs (Table 4.9), that capture some of the interactions between the factors. These combinations, however, have been selected with an overall eye on several fitness measures (counts, speeds and OD flows). Focusing on a single quantity (such as fit to true OD flows) can yield different "best" and "worst" cases. It is also possible that the factors interact in ways such that their individual effects are either enhanced or cancelled. Other combinations of factor levels may therefore correspond to the true best and worst cases, though their identification is not within the scope of this research.

The numerical results for runs 1, 10 and 11 are compared in Figures 4-10 to 4-14

Run	A	B	C	D
10	+1	0	+1	-1
11	-1	-1	-1	+1

Table 4.9: Additional Experimental Runs

(detailed results are provided in Appendix B). It is observed that the error statistics for runs 10 and 11 are generally higher than before, verifying that they represent two sets of factor levels that yield significantly more challenging estimation situations. In particular, the individual "best" factor settings (in run 11) seem to cancel out, with their interactions resulting in a more difficult calibration situation. The numerical results, however, display trends similar to those in the previous runs: the addition of degrees of freedom causes better fit, while the inclusion of speed information leads to a slight deterioration in the fit to counts. The significance of speed measurements in helping to capture supply-side dynamics, congestion evolution and the underlying OD flows is again emphasized.

4.5 Base case numerical results with SPSA

Performance statistics when adopting the SPSA algorithm for the four base-case estimators are summarized in Tables 4.10 and 4.11. A comparison with the corresponding Box-SNOBFIT results in Tables 4.2 and 4.3 illustrates the accuracy of the new solution algorithm, while revealing a consistent pattern: the parameters estimated using Box-SNOBFIT replicate the observed data marginally better than those obtained using SPSA. Intuitively, this observation may be explained based on knowledge of the mechanisms of the two algorithms. SPSA adopts a search-direction approach that must climb hills to explore global minima. While there are no explicit constraints that preclude SPSA from finding a global optimum, its search path generally follows the gradient (approximation) at each step. A complete exploration of the search space may therefore not occur. SNOBFIT, however, maintains a set of "good" parameter vectors at all times, allowing the method to potentially exploit the "spread" of this population to search more globally for a solution.

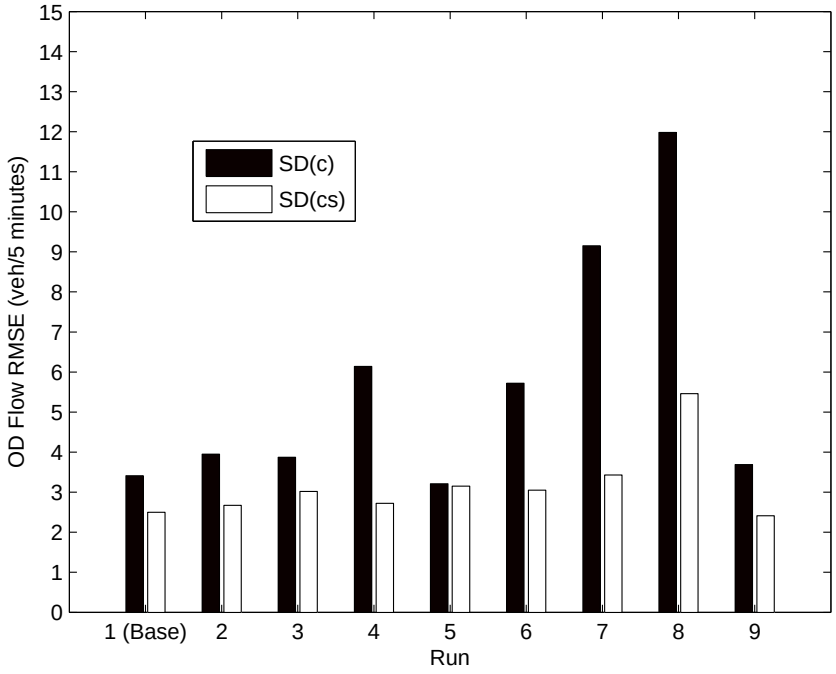

Figure 4-9: Fit to OD Flows: Runs 1 (Base) to 9

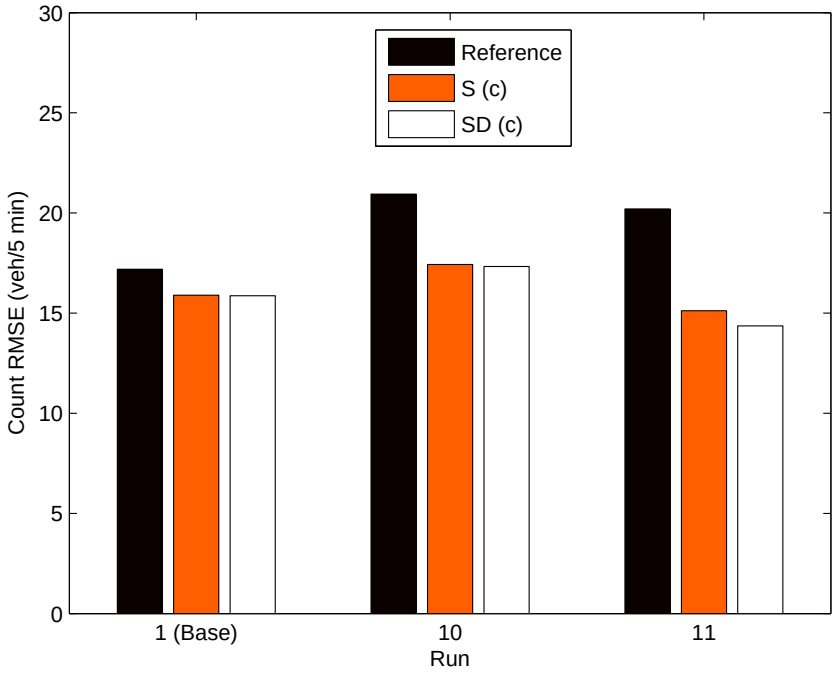

Figure 4-10: Fit to Counts Using Only Counts

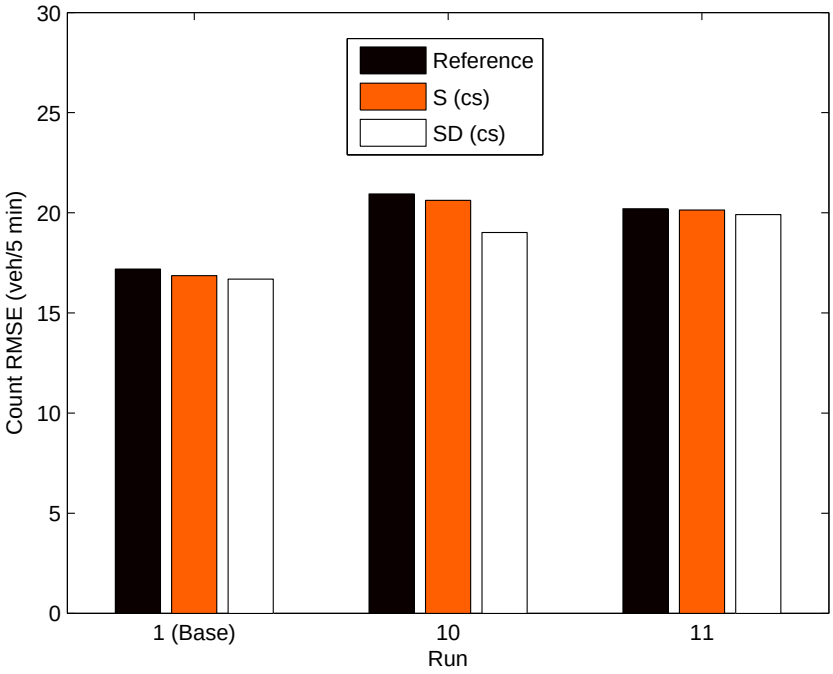

Figure 4-11: Fit to Counts Using Counts and Speeds

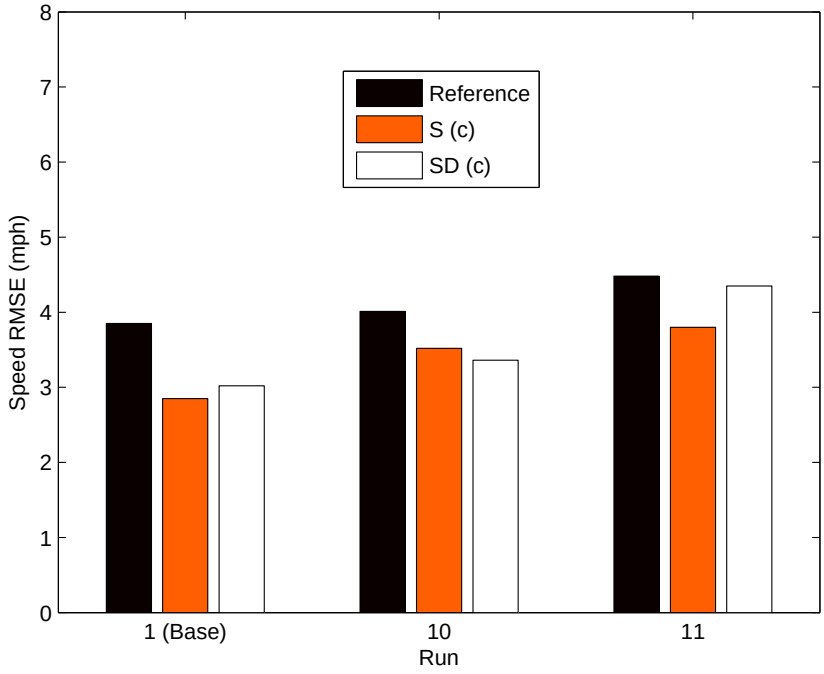

Figure 4-12: Fit to Speeds Using Only Counts

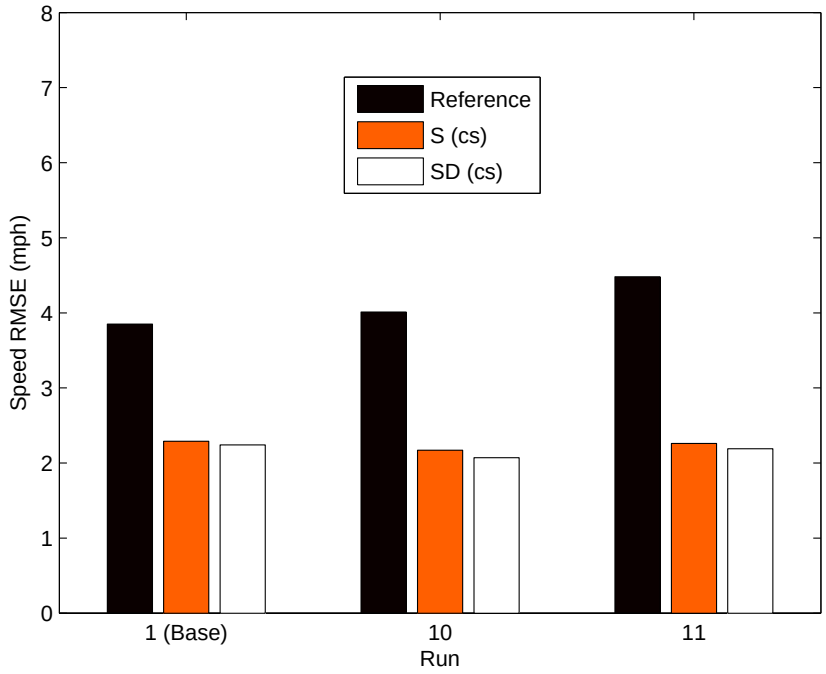

Figure 4-13: Fit to Speeds Using Counts and Speeds

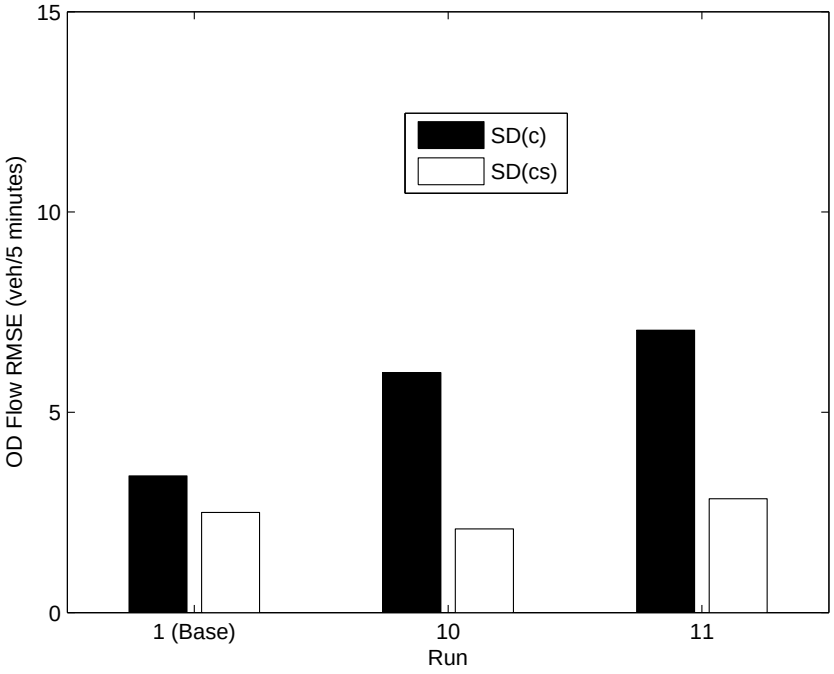

Figure 4-14: Fit to OD Flows

Scenario	Calibration Data			
	Counts (c)		Counts + Speeds (cs)	
	$RMSE^c$	$RMSE^s$	$RMSE^c$	$RMSE^s$
S	15.93 (7.3)	3.13 (18.7)	16.93 (1.5)	2.52 (34.5)
SD	15.90 (7.5)	3.17 (17.7)	16.76 (2.5)	2.48 (35.6)

Table 4.10: Base Case RMSE Statistics: SPSA

Scenario	Calibration Data			
	Counts (c)		Counts + Speeds (cs)	
	$RMSN^c$	$RMSN^s$	$RMSN^c$	$RMSN^s$
S	0.0874	0.0766	0.0929	0.0617
SD	0.0872	0.0776	0.0920	0.0607

Table 4.11: Base Case RMSN Statistics: SPSA

Further graphical comparisons between the two sets of estimated parameters confirm that the final solutions reached by the two approaches are comparable, as indicated by the tightness of the points about the 45-degree line. Figures 4-15 and 4-16 show the proximity of the two sets of OD flows (for estimators SD(c) and SD(cs)). Figure 4-17 is an example of SPSA's ability to identify stable supply model parameters that are close to their Box-SNOBFIT counterparts.

4.5.1 Scalability: Box-SNOBFIT vs. SPSA

The previous section discussed the quality of the solutions obtained using the Box-SNOBFIT and SPSA algorithms. A comparison of the computational burden of the two methods is thus in order. For this discussion, we separate the contributions of different types of computational "cost" to the overall running time needed for convergence. The total running time τ may be represented as:

$$\tau = t_1 + t_2 \tag{4.5}$$

where t_1 is the total time spent in function evaluations, and t_2 is the additional time expended in the optimization steps. Let $m_{snobfit}$ and m_{spsa} denote the number of iterations required for the two algorithms to converge. The time occupied by function

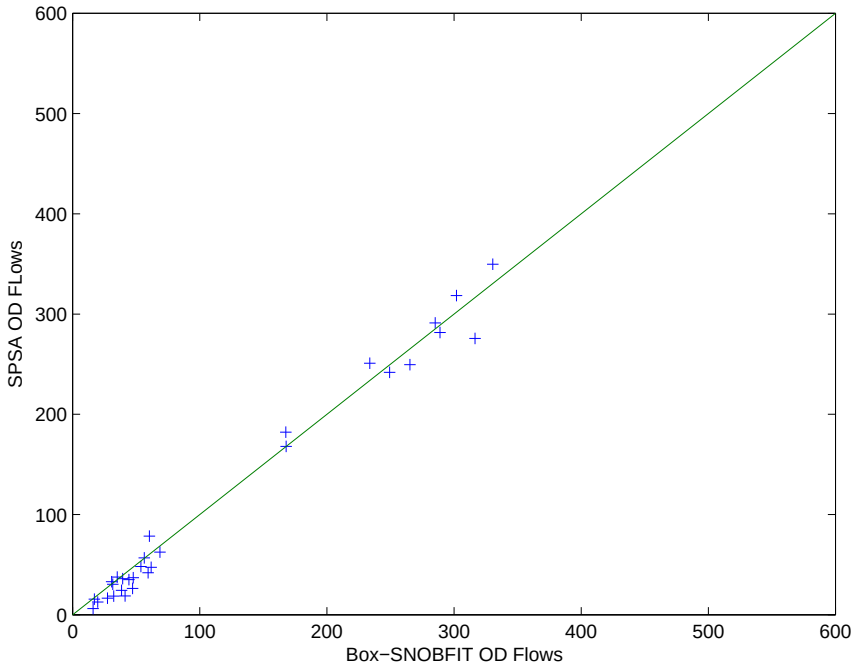

Figure 4-15: SD(c): Box-SNOBFIT vs. SPSA

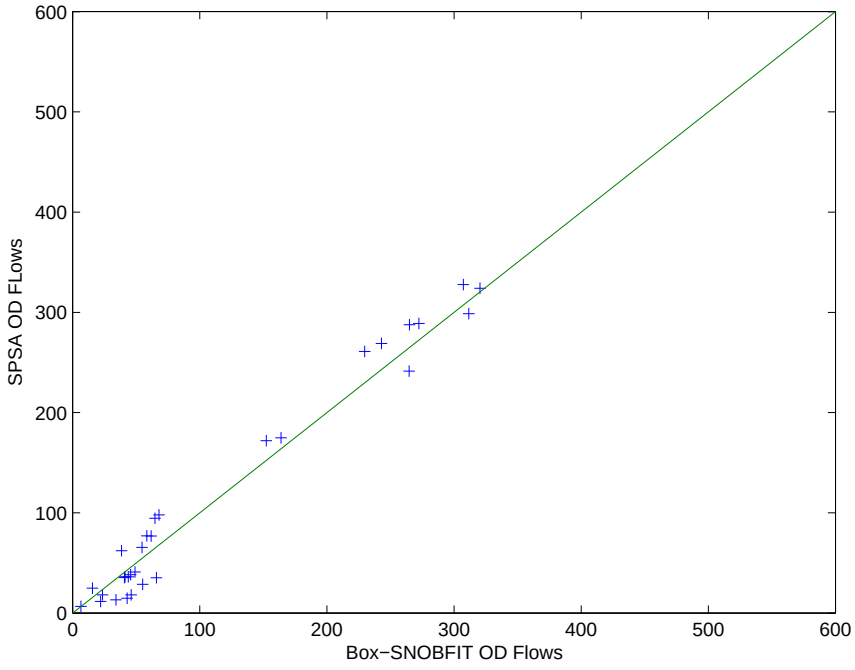

Figure 4-16: SD(cs): Box-SNOBFIT vs. SPSA

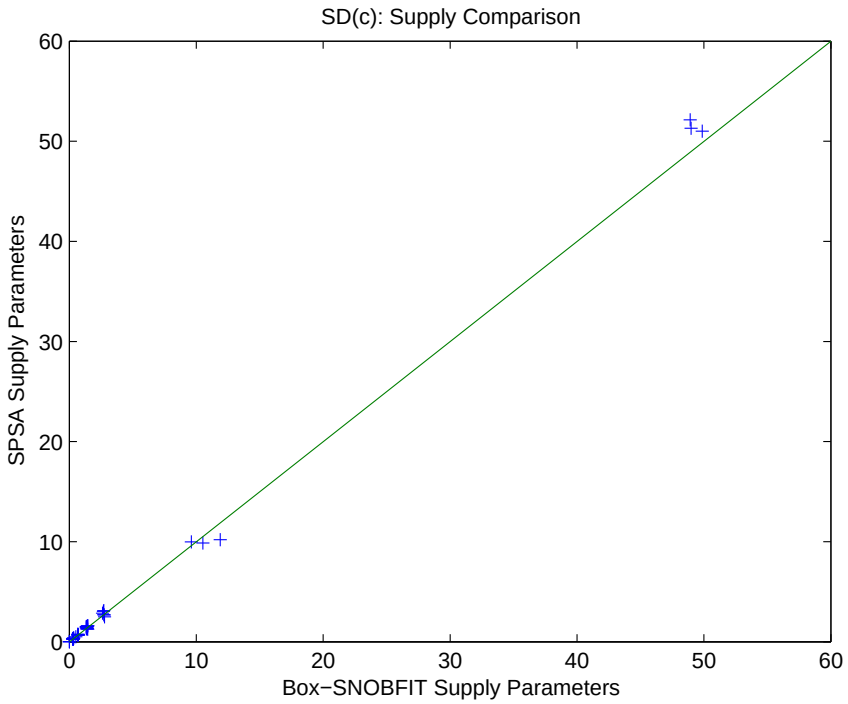

Figure 4-17: SD(c): Box-SNOBFIT vs. SPSA

evaluations may thus be written as:

$$t_1^{\text{snobfit}} = m_{\text{snobfit}}(K+6)(t) \tag{4.6}$$

$$t_1^{\text{spsa}} = m_{\text{spsa}}(\texttt{grad\_reps})(2t) \tag{4.7}$$

Here, K denotes the number of variables we wish to estimate. The constant t represents the time for a single function evaluation. It should be noted that a minimum of $(K+6)$ points must be maintained by SNOBFIT at all times. Since each iteration involves the evaluation of the function at all points in the set, t_1^{snobfit} represents the *minimum* function evaluation time. It should be noted that the time required for a single SPSA gradient calculation is the equivalent of two function evaluations. The term `grad_reps` enters the analysis when more than one gradient replication is averaged during each iteration.

The second running time component is related to how each algorithm utilizes the generated function values. Box-SNOBFIT relies on a series of quadratic local fits

during each iteration. Such a step would depend non-linearly on the problem size K, so that

$$t_2^{snobfit} = m_{snobfit} K^a \tag{4.8}$$

with $a > 1$. On the other hand, SPSA's computational requirement per iteration is limited to the calculation of gradient replications and a one-step parameter update, which takes constant time b:

$$t_2^{spsa} = m_{spsa}(b) \tag{4.9}$$

Generally, it is expected that $m_{snobfit} \ll m_{spsa}$. For small problems, such as the one described in this case study, $t_1^{snobfit}$ and t_1^{spsa} are comparable: the significant time spent by SNOBFIT in evaluating functions in each iteration is offset by the far fewer number of iterations to achieve convergence. Also, $t_2^{snobfit}$ is negligible for small K. $\tau_{snobfit}$ and τ_{spsa} are therefore of the same order of magnitude, as borne out by the empirical evidence: Figure 4-18 tracks the value of the objective function as a function of the number of iterations. For larger problems on real networks, both $t_1^{snobfit}$ and $t_2^{snobfit}$ would grow quickly with K, while t_1^{spsa} and t_2^{spsa} would remain parsimonious. SPSA is thus expected to score heavily over Box-SNOBFIT in terms of scalability. More empirical evidence confirming this hypothesis is provided in Chapter 5, using the Los Angeles dataset.

4.5.2 Conclusions

Analysis of the base case validates the appropriateness and feasibility of the proposed calibration estimators under the assumed network setting. The Box-SNOBFIT algorithm provides more accurate solutions closer to the global optimum, but preliminary analyses raise questions about its scalability to large networks. The SPSA algorithm exhibits very attractive computational performance, and yields estimates that are very close to the Box-SNOBFIT results.

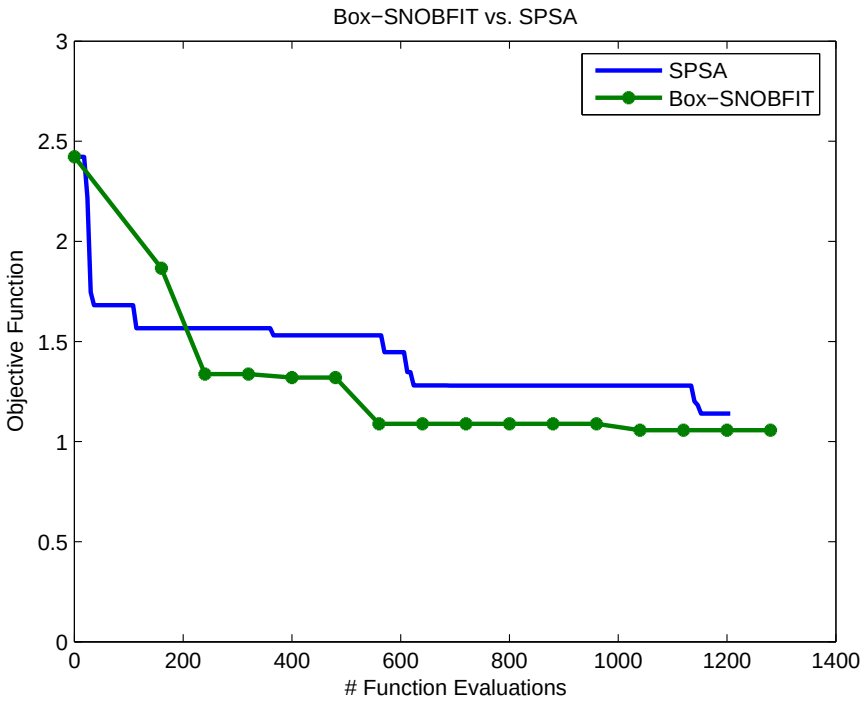

Figure 4-18: Computational Performance of SPSA and Box-SNOBFIT Algorithms

4.6 Synthesis of results and contributions

In this chapter, we evaluated our proposed calibration approach through a systematic sensitivity analysis. A synthetic network and simulated data were used to study the performance of four estimators under varying demand and supply, using the DynaMIT traffic simulator as a test-bed. A main-effects plan based on four factors (at three levels apiece) was implemented. The results were used to identify and test two additional runs representing potentially challenging interaction effects. The calibrated DynaMIT was found to accurately replicate the sensor count and speed observations, and the estimators successfully captured the underlying demand and supply parameters. A global search method and a close approximation with superior computational performance were evaluated. Chapter 5 describes further experiments demonstrating our methodology on a large and actual transportation network.

The synthetic case study presented in this chapter facilitated the evaluation of several key dimensions of the off-line DTA model calibration problem. It served as an

operational implementation of the methodology outlined in Chapter 3, thus demonstrating the feasibility of our proposed model and solution approach. The simultaneous estimation of demand and supply parameters across multiple time intervals was also documented.

An important contribution of this evaluation was the illustration of our estimator's ability to replicate the traffic patterns and behavior mechanisms that generate surveillance measurements. A range of demand- and supply-side factors were varied systematically to test the effectiveness of the estimator under both typical and extreme traveler decision processes, network demand patterns and driver behavior hypotheses. The accuracy of the fit to observed or known quantities, together with the stability of the unobserved parameters, prove the validity and robustness of the estimator.

The results in this chapter demonstrate the applicability of black-box simulation optimization techniques such as the Box-Complex, SNOBFIT and SPSA algorithms, to complex transportation network problems. The experiences from this case study have validated the parameter recapture capability of the approach, as well as its practical computational overhead. A demonstration of scalability to large networks is undertaken in Chapter 5.

Chapter 5

Case Study

Contents

5.1 Objectives

Chapter 4 described, through a detailed sensitivity analysis, an evaluation of the robustness of the calibration methodology developed in this thesis. We concluded that the new estimators have the ability to replicate demand and supply parameters underlying several synthetic sensor data sets. A key property of the synthetic network, however, was the *a priori* knowledge of the true OD flows and route choice parameter. Such data are generally not available for real networks. Further, the size of the problem allowed for extensive experimentation to fine-tune the various algorithms for rapid convergence.

This chapter documents the second application of the calibration methodology, to a network in the South Park region of Los Angeles, CA, with the following objectives:

- Demonstrate the feasibility of the estimator on a real network, with unobserved demand parameters[1].
- Demonstrate the scalability of the solution approach on a complex network with many route choice possibilities.
- Illustrate the advantages of approximation-free simultaneous demand-supply estimation (without the traditional assignment matrix).

We describe the Los Angeles dataset in some detail, to provide context to this case study. This includes an analysis of the archived sensor data collected by a real surveillance system, and a discussion of the extent of variability within the same. Numerical experiments are outlined, and results comparing the new approach to the reference case are presented. The tests include a validation exercise illustrating the benefits of the new off-line methodology in a real-time application.

[1]Supply parameters remain unobserved, as they were in Chapter 4.

5.2 The Los Angeles dataset

In this section the network, surveillance data and a log of special events in the study area are described.

5.2.1 Network description

The Los Angeles network is for the South Park area just south of downtown Los Angeles. This area includes both the Staples Center and the Los Angeles Convention Center, and so is heavily affected by special events. Specifically, the area generates a minimum of 200 events per year ranging from the Democratic National Convention and the Automobile Show, to NBA Lakers and NHL Kings games. Traffic patterns vary significantly with different types of events. Recurring commuter traffic along the Figueroa Corridor, Olympic Boulevard and other one-way streets in the financial district also pose a challenge to traffic management.

Figure 5-1: The Los Angeles Network

The region is crossed by two major freeways: the Harbor Freeway (I-110) and the Santa Monica Freeway (I-10). Traffic along the freeways is very heavy throughout the day, and on weekends. When severe and prolonged traffic congestion develops along

these freeways, diversions to parallel surface streets frequently occur. The traffic getting onto the freeways may also be diverted to other ramps connecting to several major arterials marking the border of the study network (including Hoover Street on the west, Adams Blvd towards the south, Olympic Blvd due north and Grand Avenue on the east).

The South Park area has an Advanced Traffic Control System (ATCS) with 109 traffic signals under the control of this new PC-based traffic system. An extensive video surveillance system and variable message signs are also available to confirm incidents and provide information to motorists.

The computer representation of the network consisted of 243 nodes connected by 606 directed links. The links were further divided into a total of 740 segments to capture variations in section geometry and traffic dynamics along the length of each link.

5.2.2 Surveillance data

The data for this case study was obtained from a set of freeway and arterial loop detectors that reported time-dependent vehicle counts and detector occupancies for the month of September 2004. Archived traffic records and location information for a total of 203 detectors were obtained through two sources. Freeway and ramp data were extracted from the on-line PeMS (UC Berkeley and Caltrans, 2005) database. Arterial sensor data were provided by the Los Angeles Department of Transportation (LADoT). Both sources contained traffic data by lane. Detector occupancies were converted into density estimates using standard assumptions regarding average vehicle and detector lengths. Speeds were obtained from counts and densities, using the fundamental relationship:

$$\mathsf{q} = \mathsf{k}\,\nu$$

where q is the flow rate (vehicles/hour); k is the traffic density (vehicles/lane-mile); ν is the space mean speed (miles/hour).

In addition to loop detector data, the surveillance information included a record

of incidents that were reported on the network. While the records provided details such as the incident's end time, location and general description, they did not contain any indicators of the start time.

Since the incident start time and duration are key exogenous inputs to the DTA model, the count data were analyzed to identify abnormalities that could potentially be ascribed to specific records in the incident log. No such deviations were observed. While the level of sensor coverage does not preclude major incidents on links without sensors, the probability of such an event in a relatively small section of the city may well be low. Minor incidents that occur when flows are below the link capacities (adjusted to account for the reduction in throughput due to the incident) are not expected to affect the calibration process, and may be left out of the dataset.

5.2.3 Special events and weather logs

Logs of weather conditions and scheduled special events in and around the study area were reviewed to identify days expected to have significantly modified travel demand and/or driver behavior patterns. According to the Weather Underground website (The Weather Underground, Inc., 2004), there was no precipitation in the Los Angeles area during the entire month of September 2004 (except for minor showers on the 14th). Temperatures also remained uniform and high throughout the month.

A list of weekday holidays and special events at the Convention Center was reviewed to determine if planned and scheduled events might be a factor in determining demand patterns. Labor Day counts were found to be markedly different from those measured on other weekdays in the month. This is to be expected, as the day is marked by a holiday with a high fraction of shopping trips.

5.2.4 The historical database

A total of one month (September 2004) of freeway and arterial data was analyzed, to ascertain its sufficiency for the calibration task. Figures 5-2 and 5-3 illustrate the temporal distribution of sensor counts at two representative counting locations

on freeway sections, for different days of the week. Figure 5-4 depicts a similar analysis for a sample arterial sensor. The days were randomly selected from groups of Mondays, Tuesdays, etc spanning the entire month. It was observed that 5:15-8:00 AM is the most challenging time period from the calibration view-point, as it includes a sharp (almost linear) build-up of commuter trips over a short duration, and covers the AM peak period.

Sensor count profiles were compared by time of day and day of the week to help in the classification of data into day types. Weekdays and weekends displayed markedly different traffic patterns, with Saturdays and Sundays also significantly different from each other. Weekdays exhibited similar build-up and dissipation of congestion, with no clear day-of-the-week effects. The available data was thus classified into three groups: weekdays, Saturdays and Sundays. The application of the methodology developed in this thesis is demonstrated for weekdays.

5.3 Application

5.3.1 Reference case

A detailed presentation of the reference case (summarized in Section 2.5) applied to the Los Angeles dataset is available in Gupta (2005). The work represents the best current methods applied to off-line DTA system calibration, with the following salient features:

- Segment capacities (under normal conditions and with incidents) are approximately determined according to sensor flow data, the number of freeway lanes, arterial signal timing plans and the recommendations of the Highway Capacity Manual.

- Segments are classified according to appropriate physical attributes (facility type, number of lanes, etc.). Speed-density functions for each segment type are estimated through local regressions between sensor speed and density (occupancy) measurements.

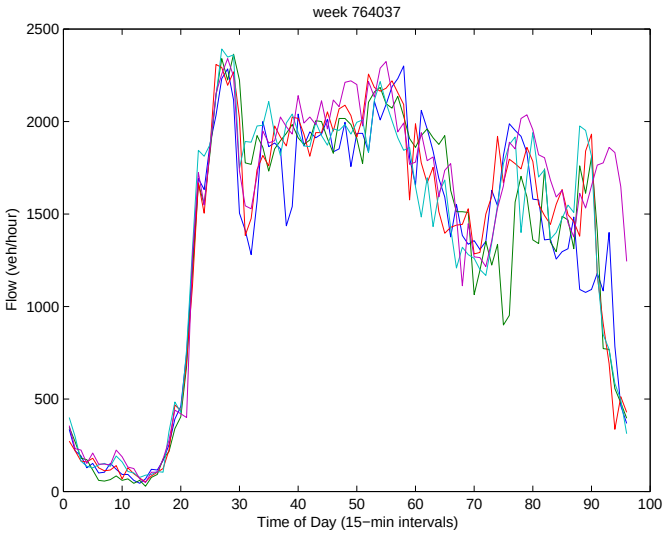

Figure 5-2: Freeway Flows by Day of Week (Sensor ID 764037)

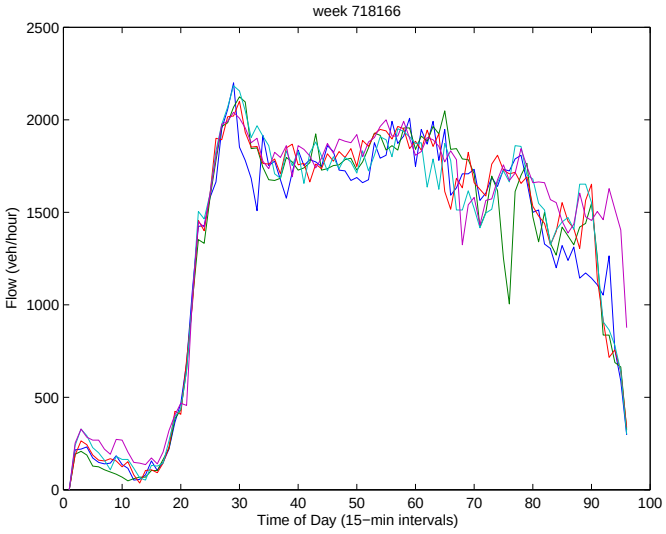

Figure 5-3: Freeway Flows by Day of Week (Sensor ID 718166)

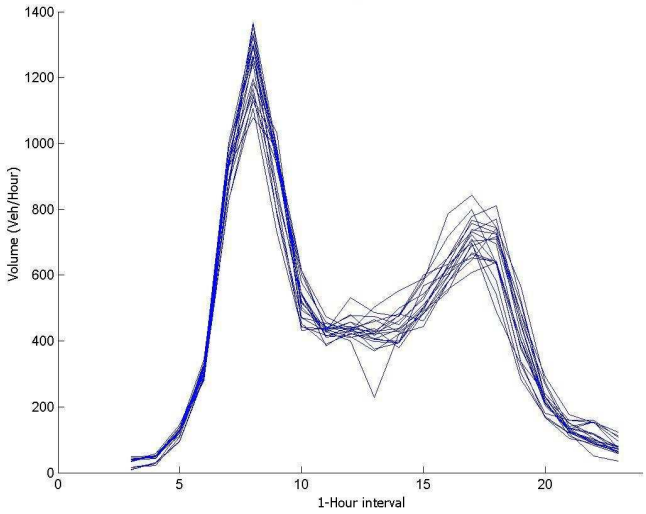

Figure 5-4: Arterial Flows by Day of Week

- The supply parameters (capacities and speed-density functions) are held constant once they are estimated through the steps described above.

- Time-varying OD flows are estimated with a sequential approach using only sensor count data. The restriction on the type of data stems from the use of a linear assignment matrix mapping between OD flows and link counts. Corresponding mappings for speed or density measurements are generally intractable.

- The route choice parameter is estimated through a manual line search.

Gupta (2005) describes the application of the above reference methodology to create a historical database of demand and supply parameters for the entire day (3:00 AM -midnight).

5.3.2 Network setup and parameters

The time period of 3:00 AM–9:00 AM was selected, so as to include the peak 5:15–7:00 window and to provide extension into the demand plateau region. Given the low levels of traffic in the early hours of the day, the period from 3:00-5:15 AM was also

used to warm up and load the network. The focus of the evaluation was thus limited to 5:15-9:00 AM, which was divided into 15-minute time intervals.

A total of $n_{OD} = 1129$ active OD pairs was identified by Gupta (2005) for the Los Angeles network, by tracking the non-zero flows over several time intervals of sequential OD estimation. The flow between every OD pair was estimated for each time interval in the study period. The set of demand parameters was augmented by a travel time coefficient used by DynaMIT's route choice model.

Supply parameters (segment capacities and speed-density function parameters) were also part of the calibration. Segments were grouped based on physical attributes such as their position on the network, and the number of lanes in their sections. Supply parameters were estimated for each group.

5.3.3 Estimators

Five estimators were employed in addition to the reference case. As in Chapter 4, S(c) and SD(c) correspond to the count-based estimation of supply parameters only, and all (supply+demand) parameters, respectively[2]. D_1(c) corresponds to the estimation of demand parameters alone (OD flows and route choice parameters), using Ref supply parameters. In addition, estimator D_2(c) was developed to identify only demand parameters using the supply parameters from S(c) as given. D_2(c) thus corresponds to sequential demand-supply calibration, as in the reference case. Further, the two D estimators will help verify the impact of the demand component on the outcome of the calibration process. Since the OD flows typically outnumber the other parameters, D would be expected to provide a more significant improvement in fit than S, over the reference case[3]. D also includes the effect of simultaneous OD estimation across intervals, and the direct use of simulator output without linear assignment matrix approximations. The final estimator, SD(cs), corresponds to simultaneous demand-supply calibration using both count and speed data.

[2]Demand parameters from Ref were used while estimating S(c).

[3]This effect was not prominent in the previous case study, due to the relatively small number of demand parameters.

While the combination of S(c) and D_2(c) denotes a sequential approach, it should be noted that the individual demand and supply calibration methods are those developed in this thesis, and differ markedly from those used in the reference case. Critically, the new approach eliminates the dependence on an assignment matrix, thus also providing the flexibility to incorporate any generic traffic measurement into the calibration framework. Supply calibration is also performed at the network level, rather than locally at individual sensor locations.

A comparison of D_2(c) and SD(c) could provide information about the benefits (if any) of simultaneous demand-supply calibration over the more traditional sequential approach.

5.3.4 Measures of performance

The Root Mean Square Normalized (RMSN) error statistic defined in Chapter 4 was used to document and analyze the performance of the various calibration estimators. The fit to counts and speeds were computed across all sensors, as well as for freeway and arterial sensors, to study the accuracy by type of roadway facility. Since true OD flows, route choice and supply parameters are not available for real networks, evaluations were limited to the fit to measured data.

Tests of the accuracy of the calibration were augmented with additional validation analyses. Estimation and prediction tests through a rolling-horizon implementation of the calibrated DynaMIT system were employed together with a new day of sensor count measurements, in order to validate the real-time performance obtained as a result of the above off-line calibration.

5.3.5 Solution algorithm

In Chapter 4, we analyzed the computational effort associated with the Box-SNOBFIT and SPSA methods as a function of the number of unknowns to be estimated through calibration. We concluded that the two algorithms estimate comparable parameters, though SPSA does so at a fraction of the computational cost (measured by running

time to convergence) for large-scale problems. The validity of this conclusion was confirmed by initial tests on the Los Angeles dataset.

The test case involved 7 consecutive time intervals, resulting in $K = 4629$ variables. A solution to the problem was attempted with both algorithms, on a dedicated Pentium 4 processor with 2 GB of physical memory and 750 GB of total hard disk capacity. The machine ran the Fedora Core 3 Linux operating system. The time taken for a single function evaluation was approximately 1.5 minutes. While this number may seem small, a simple calculation of per-iteration effort illustrates the significant savings provided by SPSA.

A single iteration of SNOBFIT requires $K + 6 = 4635$ function evaluations, while the corresponding number for SPSA is 6 (corresponding to averaging across 3 gradient replications). Each SNOBFIT iteration thus takes nearly 116 hours (about 5 days), while SPSA requires just 9 minutes! Empirical evidence strongly supported the argument that SPSA can therefore make significant progress towards the optimum solution (through many more iterations) well before SNOBFIT is even ready to perform its first set of quadratic minimizations and recommend potential solution points. SPSA thus represents a scalable solution approach (though not strictly a global estimator), while Box-SNOBFIT may be used on smaller instances to locate a more precise global optimum. The results in the remainder of this chapter were obtained through the application of SPSA.

Most of the theoretical and empirical results involving stochastic approximation (SA) techniques correspond to cases where the components of the parameter vector $\boldsymbol{\theta}$ have similar magnitudes. Under this condition, it generally suffices to adopt uniform a_j and c_j for all components of $\boldsymbol{\theta}$. The DTA calibration problem obviously deviates from this requirement. While the OD flows are definitely positive and often fairly large, the supply parameters are more diverse in their range. The travel time coefficient in the route choice model is small in magnitude and possesses a negative sign. A further complication is that the different parameter magnitudes depend on the units used to measure them.

Spall (1998b) mentions the potential need for scaling in order to "regularize" the

problem and ensure fast convergence. One method of achieving this objective is to scale all parameters to similar magnitude (Gill et al., 1984), so that the usual SPSA steps can be directly applied. Other approaches for parameter-specific scaling are suggested in Spall (1998b).

5.4 Results

The empirical results are presented in two sections. The first section details the performance of the various estimators in an off-line setting, employing the same sensor dataset used by the reference estimator. Subsequently, a new weekday is selected, and the calibrated DynaMIT system is operated in real-time fashion to validate the improvement in the system's ability to estimate and predict traffic conditions in a real-time setting.

5.4.1 Calibration results

Table 5.1 contains the fit-to-counts statistics for the various estimators. As expected, all four estimators provide significant levels of improvement over the reference case.

The importance of capturing network-wide effects in the calibration procedure is underscored by a comparison of estimator S(c) against the reference case. The significant improvement in fit to both counts and speeds is principally due to the shift away from the local sensor-by-sensor supply calibration adopted in Ref.

Estimator	Fit to Counts ($RMSN^c$)		Fit to Speeds ($RMSN^s$)	
	Freeway	Arterial	Freeway	Arterial
Ref	0.218	0.239	0.181	0.203
S (c)	0.149	0.178	0.119	0.131
D_1 (c)†	0.114	0.143	0.118	0.125
D_2 (c)‡	0.103	0.126	0.107	0.112
SD (c)	0.090	0.113	0.088	0.093
SD (cs)	0.098	0.114	0.048	0.058

†Using Reference supply parameters
‡Using S (c) supply parameters

Table 5.1: Fit to Counts: RMSN (15-minute counts)

The practical usefulness of estimator S extends beyond the above results. Realistic sensor coverage levels mean that a large fraction of a network's links are not instrumented. Currently, such links (or their constituent segments) are grouped together with the closest segment type, and a common speed-density relationship is fitted. Estimator S allows the estimation of a potentially greater number of relationships, with the closest segment type only providing *a priori* parameter estimates that may be revised for better network-wide fit. Further, in applications where only one of the three basic traffic data (counts, speeds and densities) is available, the reference case would be faced with insufficient information to derive speed-density relationships. Estimator S may then be used to update supply parameters transferred from another location.

$D_1(c)$ fits the counts better than S(c), thus highlighting the importance of demand parameters for calibration. Since demand is the basic and primary driver of traffic conditions on the network, this result is expected. From an optimization perspective, the number of unknown OD flows is typically far larger than the supply parameter set. Consequently, D provides many more degrees of freedom that allow the algorithm more flexibility in finding a better solution. Interestingly, calibrating demand parameters using only count data also results in an improvement in speeds, potentially due to better density estimates arising from demand patterns that are closer to the true values.

Demand calibration through $D_1(c)$ also provides insight into the limitations of traditional OD estimation approaches. First, the linearizing assignment matrix transformation of current methods approximates the complex relationship between OD flows and sensor counts. Second, the often-adopted sequential approach to estimating OD flows across successive intervals may fail when trip times are much larger than the width of the estimation interval. A majority of the vehicles departing on half-hour-long trips, for example, could affect counts in future intervals, and may not be observed during the fifteen-minute departure interval. The sequential approach ignores the impact of this lag between a vehicle's departure time and when it is actually "measured" on the network. While the assumption may be reasonable on

small networks with short trips, it is unrealistic on large and congested networks with multi-interval trips. The interval width also plays a crucial role, with shorter intervals accentuating the limitation. Estimator D removes both drawbacks, thus providing more accurate and efficient OD flow estimates. The improvement in fit for $D_1(c)$ over the reference case highlights this important contribution [4].

$D_2(c)$ completes one iteration of sequential demand-supply calibration, in which the supply parameters obtained from S(c) are used while estimating only OD flows and route choice model parameters through the D estimator. The results show better fit than either S(c) or $D_1(c)$, indicating the benefits of joint model calibration. Further iterations between the demand and supply estimators may be performed until convergence criteria are satisfied. However, it must be remembered that each iteration consists of two complex, large-scale optimization problems. Further, the rate of convergence of this iterative method is difficult to establish, and the expected number of iterations is consequently unavailable.

SD(c) represents the simultaneous equivalent of $D_2(c)$, with both demand and supply parameters estimated together. This estimator thus does not involve demand-supply iterations, and terminates after the solution of a single optimization problem. This approach is preferable, as it provides both added efficiency (through simultaneous calibration) and rapid convergence. SD(c) improves upon $D_2(c)$, though the reduction in RMSN is relatively small. It should however be preferred in practice, owing to the advantages outlined earlier.

The final estimator, SD(cs), extends the SD(c) case to match speed observations in addition to counts. As in the previous case study, the introduction of speed measurement equations results in a marginal loss of fit to counts. This must be anticipated, as the objective function being minimized now includes additional terms. From a practical perspective, the speed information plays the role of selecting the right set of demand parameters among many that may capture the counts accurately. In Park et al. (2005), for example, traditional count-based OD estimation reliably

[4] $D_1(c)$ was estimated with reference supply parameters. The improvement in fit over the reference case therefore captures the differences in demand estimation between the two approaches.

captured temporal count profiles, yet failed to replicate path travel times measured using probe vehicles. A calibration methodology that can significantly improve the model's ability to explain traffic dynamics, with minimum impact to the fit to counts, would thus be invaluable in travel time reliability studies and route guidance applications. The fit to speed data improves significantly from SD(c) to SD(cs), underscoring a key contribution of the proposed calibration methodology arising from the ability to include general traffic data.

Figure 5-5 compares the fitted counts from Ref and SD(c), against the actual counts for all sensors and time intervals. The fit to sensor count data was also analyzed by time of day, to ensure that the calibration methodology effectively tracked the profiles of observed counts across the network. Figure 5-6 shows cumulative counts (summed across all sensors). Visual inspection illustrates the accuracy of SD(c), and the improvement over Ref.

Some sample plots of cumulative counts at individual sensor locations are presented in Figures 5-7 to 5-12, which further indicate the accuracy of the calibration approach developed in this thesis. Indeed, the SD(c) case results in a more accurate fit to counts when compared with the reference case. The six sensors shown here were selected at random to provide a spatial spread across the entire network.

The fit to count data across the entire network is illustrated in Figure 5-13. $RMSN^c$ values for a sample of sensors distributed spatially on both freeways and arterials are provided, as further proof of calibration accuracy.

5.4.2 Validation results

The results in the previous section illustrated the ability of the methodology to better fit observed data in an off-line scenario. Additional validation tests were performed to evaluate the ability of the calibrated system to provide real-time traffic estimations and predictions of higher quality compared to the reference case. For this purpose, the DynaMIT-R DTA system was run in a rolling horizon, with sensor count data from a new day (not used for calibration) used to simulate the real-time data collection

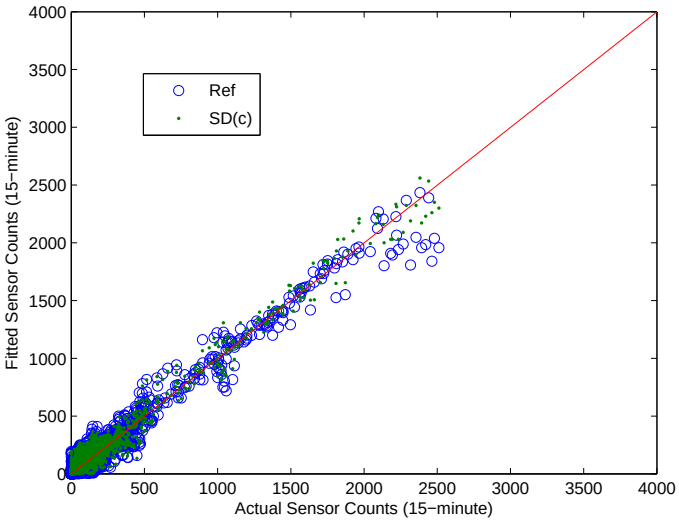

Figure 5-5: Sensor Counts (all sensor locations)

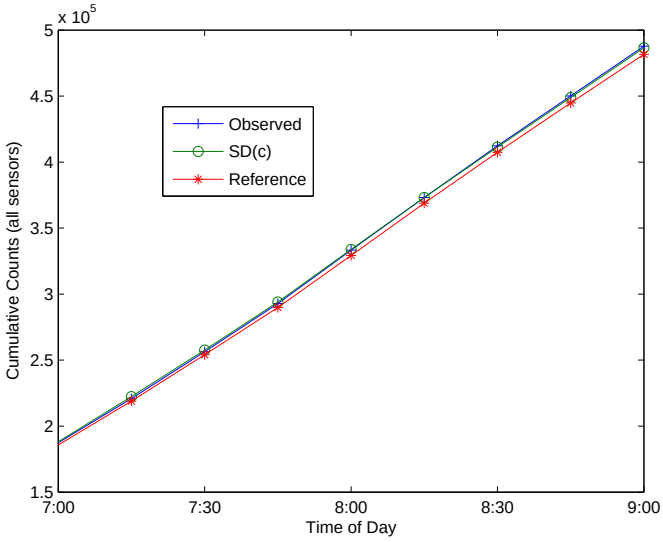

Figure 5-6: Cumulative Counts (all sensor locations)

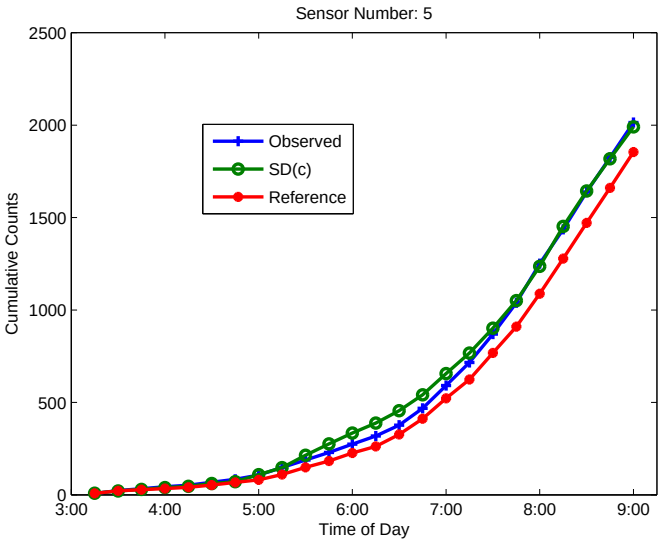

Figure 5-7: Cumulative Counts (Sensor 5)

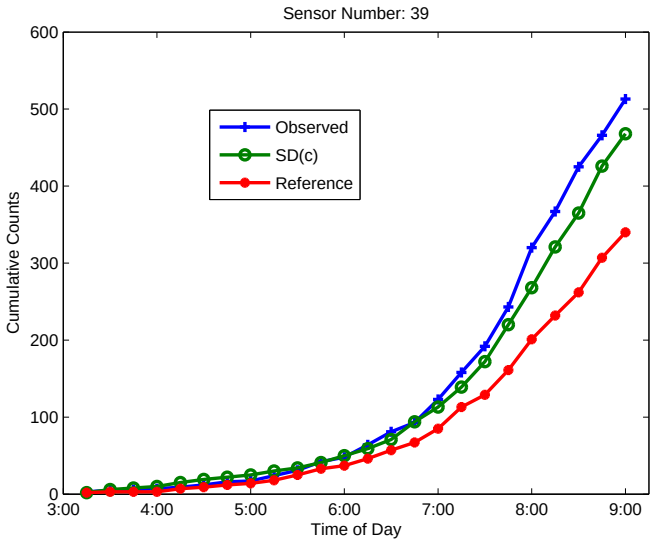

Figure 5-8: Cumulative Counts (Sensor 39)

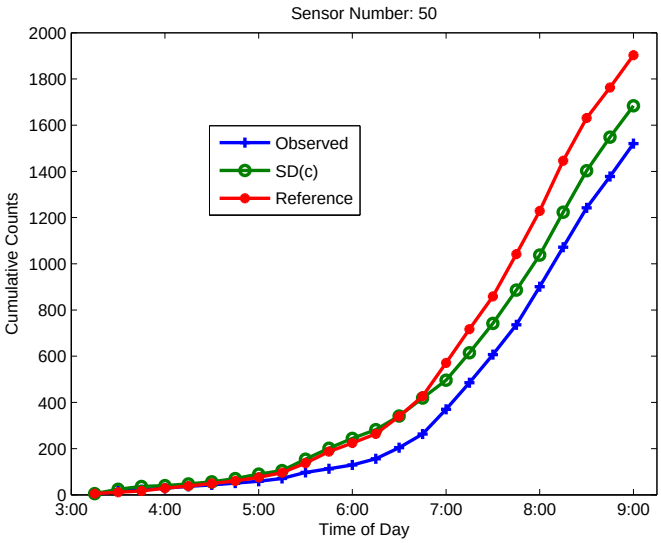

Figure 5-9: Cumulative Counts (Sensor 50)

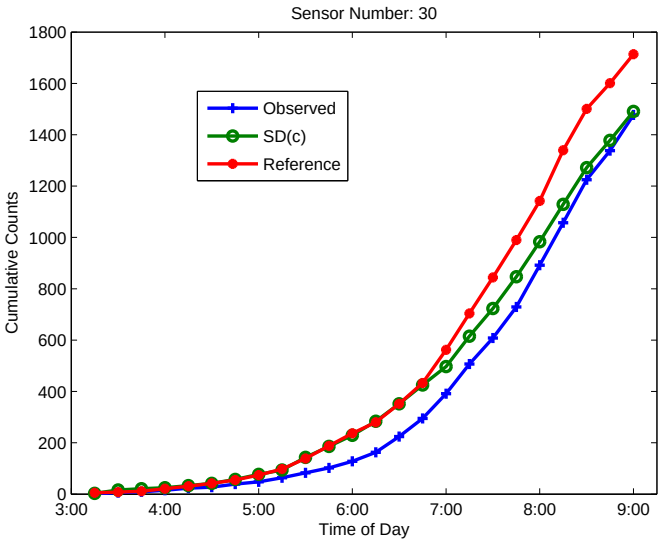

Figure 5-10: Cumulative Counts (Sensor 30)

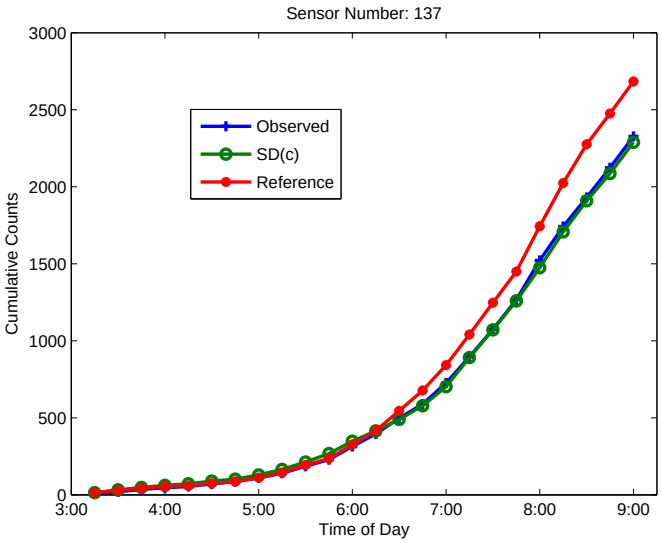

Figure 5-11: Cumulative Counts (Sensor 137)

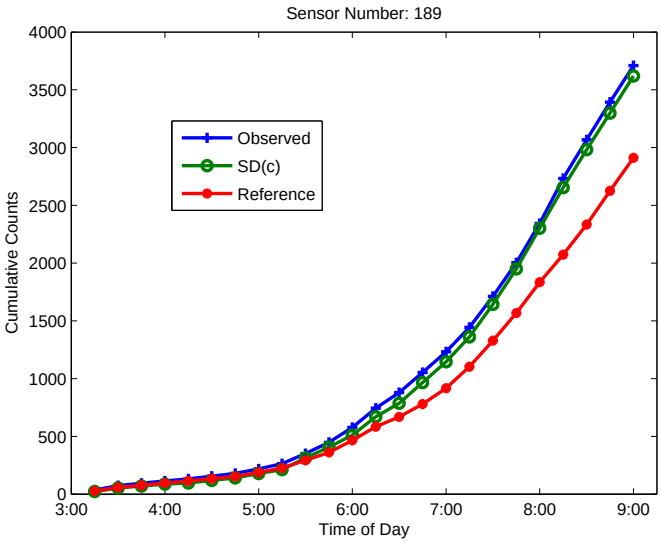

Figure 5-12: Cumulative Counts (Sensor 189)

duties of the network's surveillance system[5].

In the tests, the calibrated demand and supply parameters from SD(c) and the reference case were in turn input to DynaMIT-R as a historical database representative of real traffic patterns and conditions. The real-time system was then executed from 5:15 to 9:00 AM in a rolling horizon comprised of 15-minute estimation phases and one-hour prediction phases. Figure 5-14 portrays the 14 horizons contained in the analysis period. An OD prediction process based on an autoregressive (AR) process of degree 3 was employed, meaning that the latest estimations and predictions of the past 45 minutes were used to predict the flows for each interval (one-hour predictions were generated in four 15-minute steps). The factors in the AR process were estimated based on the OD flow estimates obtained through off-line calibration in the reference case.

It should be noted that DynaMIT-R's OD estimation module uses the formulation based on the assignment matrix, which represents an approximation of the formulation used in this thesis. Further, the rolling horizon implementation employs a sequential OD estimation framework, which could cause further deterioration in the quality of the real-time OD flow estimates and the corresponding fit to sensor counts. While the simultaneous approach adopted for calibration is superior to the sequential method, the latter possesses computational benefits that are suitable for real-time operation.

Evaluating the real-time performance of the calibrated DynaMIT system involves the comparison of the 1-step, 2-step, 3-step and 4-step predictions for each time interval against those obtained by assigning the corresponding historical and estimated OD flows. The most direct analysis of this kind involves the OD flows themselves. The historical (calibrated) OD flows represent the expected demand patterns for the specific *type* of day. The data for validation, though obtained from the same day type, will typically deviate from the historical flows. The estimated OD flows thus represent the best knowledge of demand patterns on the given day, since they are

[5]A detailed description of DynaMIT-R and the rolling horizon framework is presented in Appendix A.

obtained by updating the historical flows using the latest sensor count information reported by the surveillance system.

Predicted OD flows from each horizon represent the best estimates of the OD flows in the absence of the critical count data. They incorporate all prior knowledge contained in the historical flows as well as the estimated flows until the previous estimation interval. Extending this argument, the historical flows would provide the lowest fit to counts, since they contain no additional information about the deviation of the current day from expected conditions for that type of day.

Table 5.2 provides a summary of the fit to counts across all horizons. Average $RMSN^c$ statistics from using two different sets of historical inputs (corresponding to SD(c) and Ref) are compared, for six different cases: estimation, four steps of prediction, and the historical. The last case corresponds to evaluating the fit to the counts on a new day, without any adjustments to the historical database.

Clearly, DynaMIT's performance in a real-time setting is vastly enhanced due to the higher quality of the SD(c) historical database, thus providing a clear validation of the benefits of the proposed methodology. Further, the statistics follow the expected pattern across the six cases: state estimation results in the lowest possible error, since it is based on known count data. Prediction quality deteriorates with the horizon, ultimately leveling at the historical value. This behavior is consistent when SD(c) is used as input, but is not evident with the reference inputs.

An interesting observation is the limited variability between the error statistics between the six cases, justifying the assumption that the different week days belong to the same type (and are characterized by similar underlying demand and supply processes).

Figures 5-15 to 5-19 illustrate the fit to counts by time of day, and underline the improvement over the reference case. As expected, the gap between the predicted and historical statistics shrinks as one looks farther into the future. For example, the 2-, 3- and 4-step predictions are progressively closer to the corresponding historical numbers than the 1-step predictions. Since the AR process predicts *deviations* of OD flows from the historical, this finding confirms that the ability to predict diminishes

	Fit to Counts (RMSN^c)	
	SD(c)	Reference
Estimation	0.102	0.303
1-Step Prediction	0.110	0.304
2-Step Prediction	0.113	0.295
3-Step Prediction	0.114	0.301
4-Step Prediction	0.124	0.300
Historical	0.127	0.336

Table 5.2: Fit to Counts: Average RMSN for 5:15 AM - 9:00 AM

with the length of the horizon, with the impact approaching zero far enough away from the start. The accuracy of the calibration, estimation and prediction capabilities are thus reinforced.

The empirical results from the validation tests are thus consistent with the calibration results and expected trends, and validate the feasibility and accuracy of the methods developed through this research.

5.5 Synthesis of results and major findings

This chapter served as a real-world validation of the off-line DTA calibration methodology proposed and tested in previous chapters. Empirical findings justifying the practical applicability and scalability of the developed methodology were presented. Primarily, the method's ability to provide significant improvements over existing calibration approaches was illustrated on a large and real network with actual sensor data. The scalability of the SPSA stochastic approximation algorithm was convincingly demonstrated as a solution approach for large-scale calibration. The results were validated through prediction tests using data from an independent day. The numerical results are consistent with standard *a priori* hypotheses, and indicate the validity and usefulness of the methods developed in this thesis.

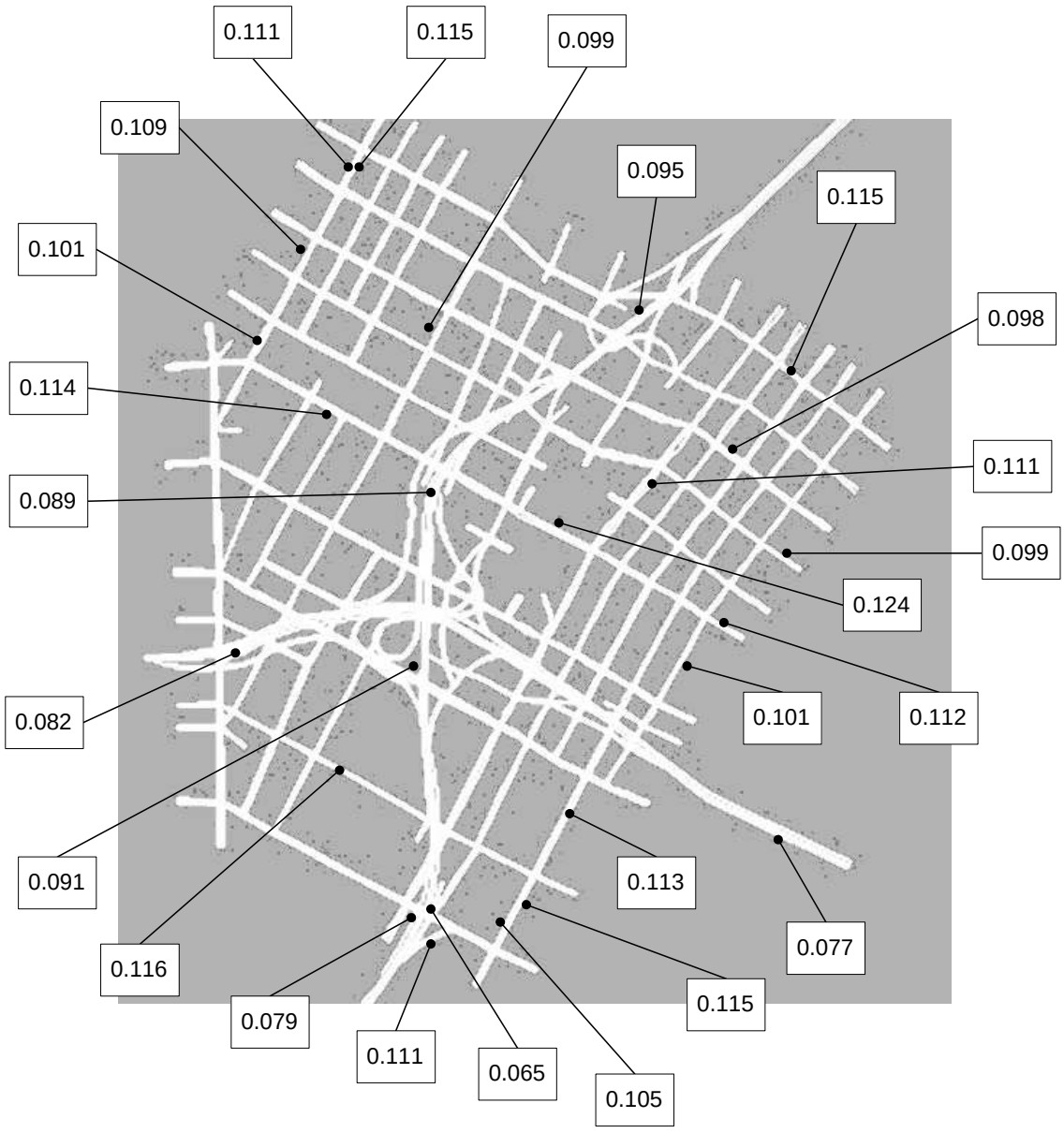

Figure 5-13: Sample Count RMSN Statistics

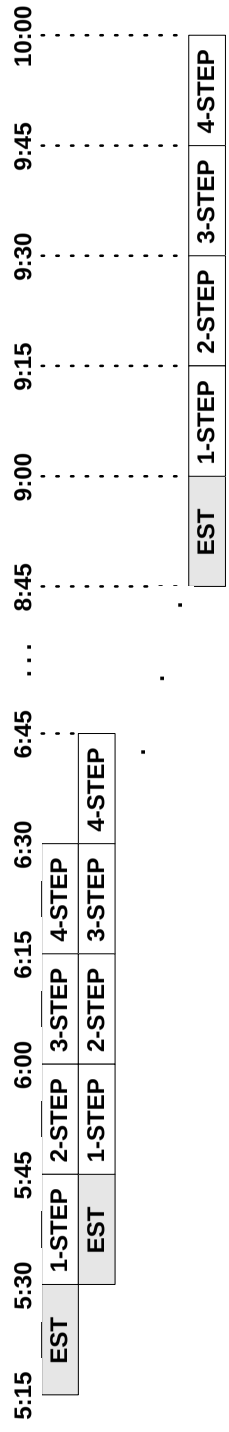

Figure 5-14: Rolling Horizons for Validation Study

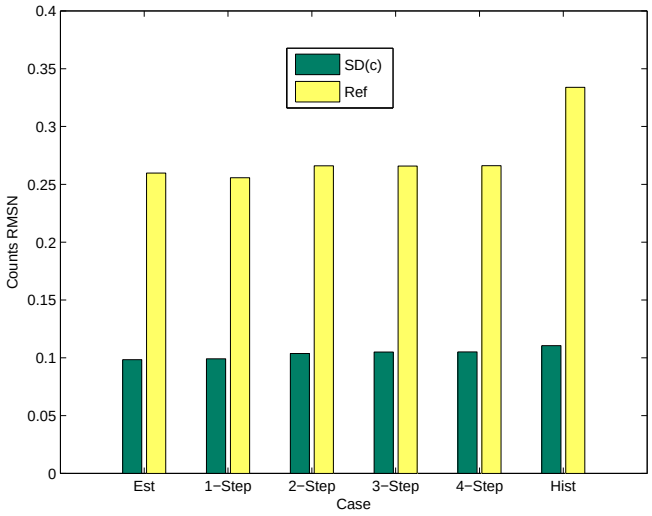

Figure 5-15: Fit to Counts: 6:15-6:30

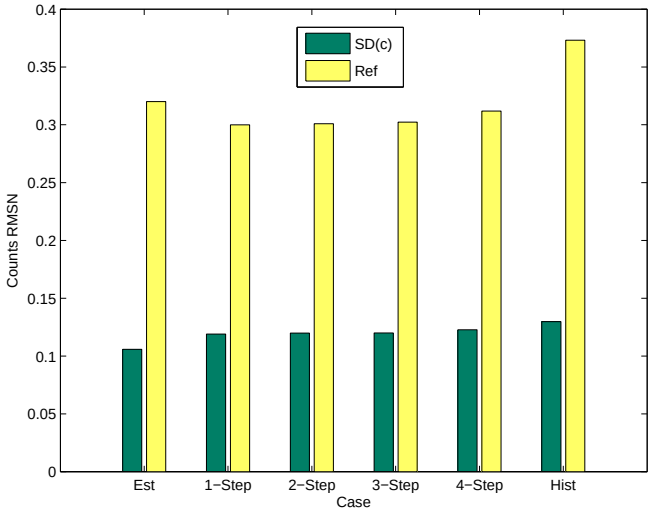

Figure 5-16: Fit to Counts: 6:30-6:45

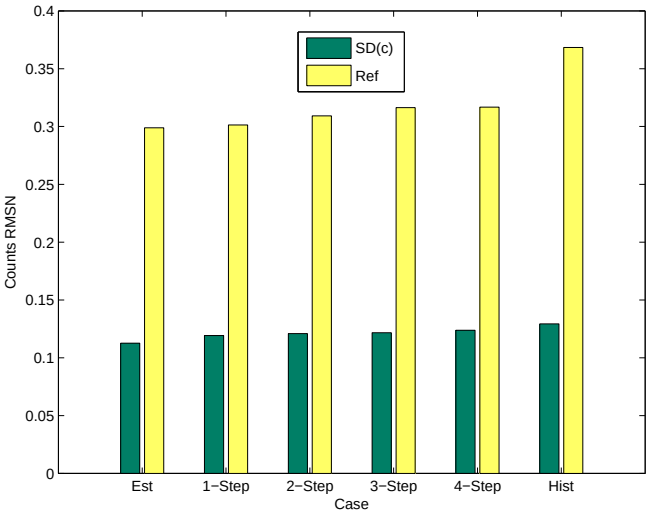

Figure 5-17: Fit to Counts: 6:45-7:00

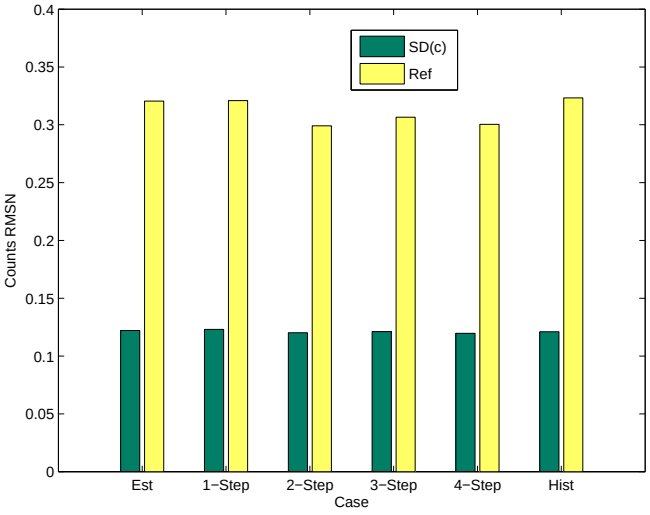

Figure 5-18: Fit to Counts: 7:30-7:45

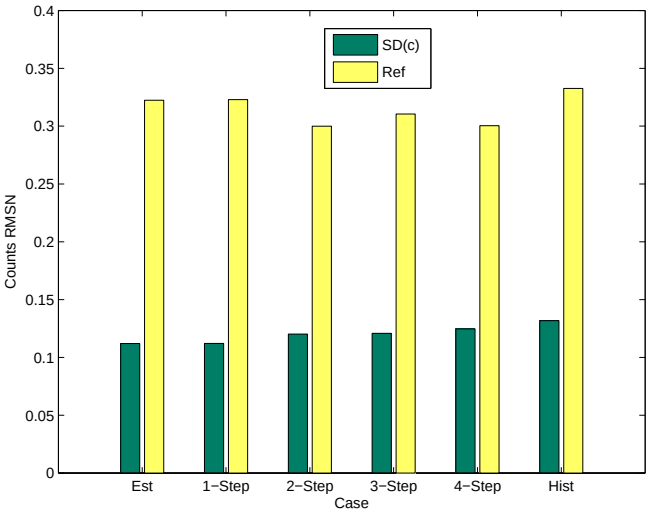

Figure 5-19: Fit to Counts: 8:30-8:45

Chapter 6

Conclusion

Contents

We conclude this thesis with a summary of the context and scope of this research, its principal contributions, and some natural avenues for future work.

6.1 Summary

Traffic data collection efforts in the past have relied on manual procedures such as surveys and vehicle counts, that are both costly and time-consuming. They have therefore been applied infrequently, and often do not capture the full range of demand and supply patterns on the network. The extensive deployment of traffic surveillance technologies has now resulted in the collection and archiving of time-varying traffic data at the network level and across multiple days, providing rich datasets for the calibration of complex DTA models. This thesis develops an off-line methodology that simultaneously calibrates both demand and supply components of DTA models using such traffic data. Solution algorithms suitable for the resulting non-linear, stochastic optimization problem are identified and evaluated through detailed case studies. The benefits of this approach over current practices that estimate supply and demand parameters sequentially are presented, and its ability to replicate underlying network parameters in a robust manner is demonstrated. The role of speed data in improving the underlying demand parameters is also highlighted. Further, the scalability of the methodology is illustrated through a real network with actual sensor data, which represents a key finding from a practical perspective.

The benefits of the proposed calibration methodology are many. First, the simultaneous calibration of all demand and supply parameters provides the most efficient estimates of the DTA model's inputs. The Los Angeles case study clearly demonstrates the superiority of this approach over the sequential state-of-the-art. Second, the direct use of the model's output (without approximating the complex relationship between the calibration variables and the observed data) improves the accuracy of the estimates and introduces the flexibility to use general traffic data. Further, simultaneous demand estimation across multiple time intervals helps capture the effect of long trips (a sequential estimator has a more local outlook that ignores the

contributions of OD departures to measurements in future time intervals). Finally, the calibration of supply parameters at the network level significantly improves over the current local-fitting approach, by capturing the spatial and temporal correlations between the various measurements.

It should be noted that the approach developed in this thesis can be applied to any general DTA model. Though the case studies in Chapters 4 and 5 are based on the real-time, mesoscopic DynaMIT traffic simulation tool, the methodology in Chapter 3 makes no restrictive assumptions on the function $\mathsf{f}(\bullet)$, which can be (a) analytical or simulation-based, (b) deterministic or stochastic, (c) designed for network planning, operations or management. Simulation-based models can further be microscopic, mesoscopic or macroscopic, with the approach currently being tested on the MITSIMLab microscopic simulator. The applications of this research are therefore diverse.

We now summarize the main contributions of this research (Section 6.2), and outline some directions for future work (Section 6.3).

6.2 Research contributions

The following are the primary contributions of this research:

- Development of an off-line methodology for the simultaneous demand-supply calibration of general DTA models, that:
  - uses model outputs to directly capture complex relationships between the data and model parameters (does not use linear approximations).
  - accommodates general traffic data (beyond loop detector counts).
  - is robust under various demand and supply situations.
  - scales to large networks.
  - was validated on real data.

The above contributions were systematically developed through the following steps:

- General formulation of the off-line calibration problem through an optimization framework.
- Analysis of problem characteristics, and the identification of suitable solution algorithms suited to its non-linear, non-analytical and stochastic nature.
- Proof of concept of the proposed methodology on a synthetic network with simulated traffic data, and a systematic sensitivity analysis demonstrating the ability to recover known underlying parameters for a wide range of demand and supply settings.
- Tests on a large traffic network with actual sensor data, that demonstrated both the practical nature of the methodology and its scalability to large, real traffic problems. Confirmation of benefits, including the advantage of model estimation without the linear assignment matrix approximation. Validation of calibration benefits for an on-line application.

6.3 Future research directions

While the theoretical calibration framework developed in this thesis is general, simplifying assumptions were made in the case studies in order to demonstrate the primary contributions of the methodology. Relaxing these assumptions could yield potentially useful research extensions, some of which are outlined next.

6.3.1 Equilibrium and day-to-day effects

Network travel times $\mathbf{TT}^{rc}$ used by the route choice model were assumed to be exogenous inputs in the two case studies. Consistency between the estimated model parameters and travel times could however be enforced by embedding an equilibrium model (such as the day-to-day learning mechanism in Equation 2.11) within $\mathsf{f}(\bullet)$.

The demonstration of our approach on a dataset with potential travel time learning effects across multiple days will represent another interesting application of this work. However, a dataset that supports a learning hypothesis will have to be obtained first. Such a dataset must span a significant length of time, such as several months, in order to allow for the identification of drivers' long-term learning processes when faced with evolving traffic patterns.

6.3.2 Observability and optimal sensor coverage

Empirical tests (Gupta, 2005) have indicated that the dynamic OD estimation problem is observable[1]: when the number of OD pairs is larger than the number of independent sensor counts, sequential OD estimation with a suitable transition equation could yield OD flows that do not depend on the (arbitrary) choice of *a priori* flows for the first interval. The ratio of the size of the OD vector to the size of the counts vector is an indicator of the number of "warm-up" intervals before stable OD estimates are obtained. The methodology developed in this thesis presents the opportunity to test if added information through speed measurements can hasten the onset of observability, thus shortening the "warm-up" period required.

Observability tests help determine if the existing sensor coverage is adequate to uniquely estimate all OD flows. An interesting trade-off in this context is the relative benefit of count and speed information for varying sensor coverage levels. Increased count data coverage potentially provides more information about the underlying OD flows, which might limit the additional benefit from speed measurements. A related issue of some practical significance is the identification of optimal spatial distributions of sensors. This question is pertinent in the design stage when sensor deployment locations are flexible, and can thus be optimized. While this problem has clear objectives, its solution is non-trivial. An obvious requirement would be that the final solution minimize the number (or cost) of sensors deployed, while maximizing estimation accuracy. A more complex variant would impose a budget constraint that limits the total number of sensors that can be installed. Apart from the large number of pos-

[1]The concept of observability was discussed in Section 3.3.

sible sensor locations, the potentially high correlations between measurements from neighboring sensors adds to the complexity (since additional sensors close to existing ones provide diminishing information about the unknown parameters). A valuable avenue for future research is a sensitivity analysis that explores the impact of sensor coverage on calibration accuracy, and the development of guidelines for the better design of surveillance systems with multiple sources of traffic data.

6.3.3 Impact of incidents

Incidents can have a significant impact on the estimation and prediction accuracy of DTA models, particularly in on-line applications. Though the true severity and duration of an incident may not be known until well into the future, a good estimate of the reduction in capacity is essential for maintaining the accuracy of the system's predictions. Knowledge of capacity reduction factors for incidents observed in archived datasets can therefore allow operators of Traffic Management Centers to choose an appropriate factor for disruptions detected in real-time on the network. The estimation of such factors was demonstrated through the synthetic case study. However, complete incident data for the Los Angeles application were unavailable. A log of severe incidents from real networks, including location, start and end times, and a description of severity, can be used to classify incidents and estimate representative capacity reduction factors for planning and on-line applications.

6.3.4 Historical database updating

The methods in this thesis have advanced the state-of-the-art of DTA system calibration for a single day of data. In the context of real-time and on-line systems, these methods may be applied at the end of each day, using the most recent archived set of traffic measurements to update the historical database. Recent research efforts have proposed heuristic approaches for maintaining a "current" historical database that reflects all the information contained in the days leading up to the previous day. A logical next step of immense value to traffic system operators would be the empirical

testing of different updating schemes, to determine the most appropriate procedure(s) for on-line traffic systems. Ashok (1996), for example, lists several possibilities for updating OD flows, such as the use of a moving average from the past few days. Research into the consistent updating of other parameters such as error covariances will also be useful.

6.3.5 Networks, models and modeling error

The results in the two case studies are derived using a mesoscopic traffic simulation model and one real network. Tests on more networks possessing different structures (combinations of freeways and arterials with various degrees of overlapping routes) should be performed. Further, the performance of the proposed methodology on different types of DTA models should be analyzed. Apart from identifying calibration guidelines by model type, such tests could reveal the impact of modeling errors on calibration. Alternatively, different model specifications (such as speed-density functional forms) could be implemented in the same DTA model to ascertain if the calibration methodology is capable of recovering the true underlying demand parameters in each case.

6.3.6 More detailed travel behavior models

The case studies in this thesis focus on drivers' pre-trip route choice behavior, while capturing their departure time preferences implicitly through the time-dependent OD matrix. Other possible decisions relate to choice of mode and response to en-route traveler information. The use of a DTA model with transit capabilities, for example, could be used to demonstrate the impact of calibrating parameters in a mode choice model. Advanced datasets that include detailed records of traveler information provision (data dissemination methods, message sets, and displayed messages by time of day), if available, can be used to estimate models of en-route response to information. Survey data for the calibration of such models using traditional techniques are rare, and an approach based on aggregate data will be of value.

6.3.7 Emerging traffic data

The methodology and solution approach developed in this research are flexible, and allow for the use of any available traffic data in the off-line calibration process. Indeed, this flexibility was clearly demonstrated in both case studies through the use of sensor speeds in addition to traditional counts data. The sensor data however consisted of point measurements. Future applications could combine such observations with point-to-point data recorded through automatic vehicle identification (AVI) or GPS technologies. Such path (or sub-path) information may be expected to improve the efficiency of the estimated parameters, especially on the demand side.

6.4 Conclusion

A comprehensive optimization framework for the off-line calibration of complex dynamic traffic assignment systems was developed in this research. The framework was operationalized by adapting state-of-the-art simulation optimization algorithms to suit the unique characteristics of the problem at hand. Detailed demonstrations on both synthetic and large, real networks have validated the efficiency and practical nature of the developed methodology, and confirm that the new approach of simultaneous demand-supply calibration significantly out-performs the sequential state-of-the-art.

Appendix A

Overview of the DynaMIT System

Contents

This appendix introduces DynaMIT (Ben-Akiva et al. (1997)), a state-of-the-art DTA system with both real-time and planning applications. The features and functionalities of the DynaMIT system are presented, along with an overview of its model components. The system's unknown quantities (both model inputs and parameters) are enumerated, both in order to illustrate the dimensionality of the problem, and provide an introduction to the case studies presented in this thesis.

The models in a DTA system can be broadly categorized into two classes. The *demand* simulator captures aggregate flows of vehicles between points on the network, and models individual drivers' route choice decisions at various stages of their trips. In addition, the demand models play a key role in the prediction of future network flows. The *supply* simulator models vehicle movements on the links of the network. The outputs of the supply simulator include link, path and sub-path travel times, link flows, speeds and densities, and queue lengths upstream of bottlenecks.

Section A.1 provides an overview of DynaMIT-R, a DTA-based real-time mesoscopic traffic simulation model. Section A.2 reviews the framework for DynaMIT-P, a version of DynaMIT for short-term planning applications. The various demand and supply inputs to DynaMIT were reviewed earlier, in Sections 4.2.2 and 4.2.2.

A.1 Overview of DynaMIT-R

DynaMIT (Dynamic Network Assignment for the Management of Information to Travelers) is a state-of-the-art traffic simulation system based on the principle of dynamic traffic assignment. Its real-time version (DynaMIT-R) is designed for traffic estimation and prediction, and the generation of traveler information and consistent anticipatory route guidance. DynaMIT-R supports the operation of Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) at Traffic Management Centers (TMC). A planning version of DynaMIT, codenamed DynaMIT-P, employs DTA for short-term planning scenarios such as work zones, optimal VMS locations and OD estimation. Sponsored by the Federal Highway Administration (FHWA), DynaMIT was designed and developed at the Intelligent

Transportation Systems Program at the Massachusetts Institute of Technology.

A.1.1 Features and Functionality

The key to DynaMIT's functionality is its detailed network representation, coupled with models of traveler behavior. Through an effective integration of historical databases with real-time inputs from field installations (surveillance data and control logic of traffic signals, ramp meters and toll booths), DynaMIT is designed to efficiently achieve:

- Real time estimation of network conditions.
- Rolling horizon predictions of network conditions in response to alternative traffic control measures and information dissemination strategies.
- Generation of traffic information and route guidance to steer drivers towards optimal decisions.

To sustain users' acceptance and achieve reliable predictions and credible guidance, DynaMIT incorporates *unbiasedness* and *consistency* into its core operations. Unbiasedness guarantees that the information provided to travelers is based on the best available knowledge of current and anticipated network conditions. Consistency ensures that DynaMIT's predictions of expected network conditions match what drivers would experience on the network.

DynaMIT has the ability to trade-off level of detail (or resolution) and computational practicability, without compromising the integrity of its output.

A.1.2 Overall Framework

DynaMIT is composed of several detailed models and algorithms to achieve two main functionalities:

- Estimation of current network state using both historical and real-time information.

- Generation of prediction-based information for a given time horizon.

The estimation and prediction phases operate over a rolling horizon. This concept is illustrated with a simple example (Figure A-1).

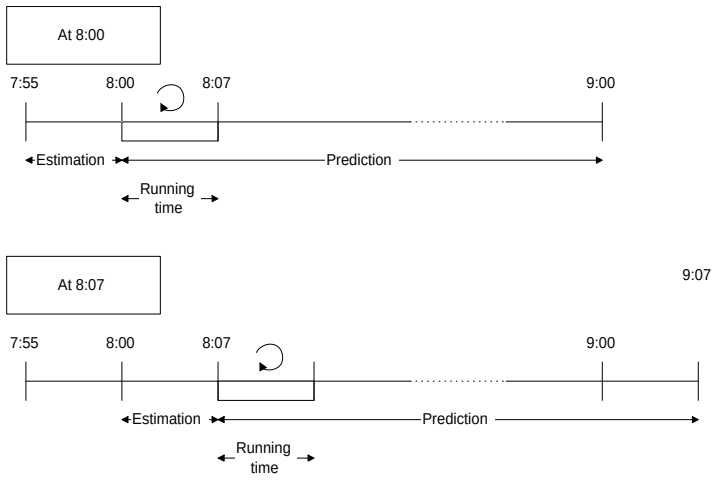

Figure A-1: The Rolling Horizon

It is now 8:00am. DynaMIT starts an execution cycle, and performs a state estimation using data collected during the last 5 minutes. When the state of the network at 8:00 is available, DynaMIT starts predicting for a given horizon, say one hour, and computes a guidance strategy which is consistent with that prediction. At 8:07, DynaMIT has finished the computation, and is ready to implement the guidance strategy on the real network. This strategy will be in effect until a new strategy is generated. Immediately following that, DynaMIT starts a new execution cycle. Now, the state estimation is performed for the last 7 minutes. Indeed, while DynaMIT was busy computing and implementing the new guidance strategy, the surveillance system continued to collect real-time information, and DynaMIT will update its knowledge of the current network conditions using that information. The new network estimate is used as a basis for a new prediction and guidance strategy. The process continues rolling in a similar fashion during the whole day.

The overall structure with interactions among the various elements of DynaMIT is illustrated in Figure A-2. DynaMIT utilizes both off-line and real-time information.

The most important off-line information, in addition to the detailed description of the network, is a database containing historical network conditions. This database might combine directly observed data and the results of off-line models. The historical database contains time-dependent data, including origin-destination matrices, link travel times and other model parameters. Clearly, the richer the historical database, the better the results. Such a rich historical database requires substantial data collection and careful calibration.

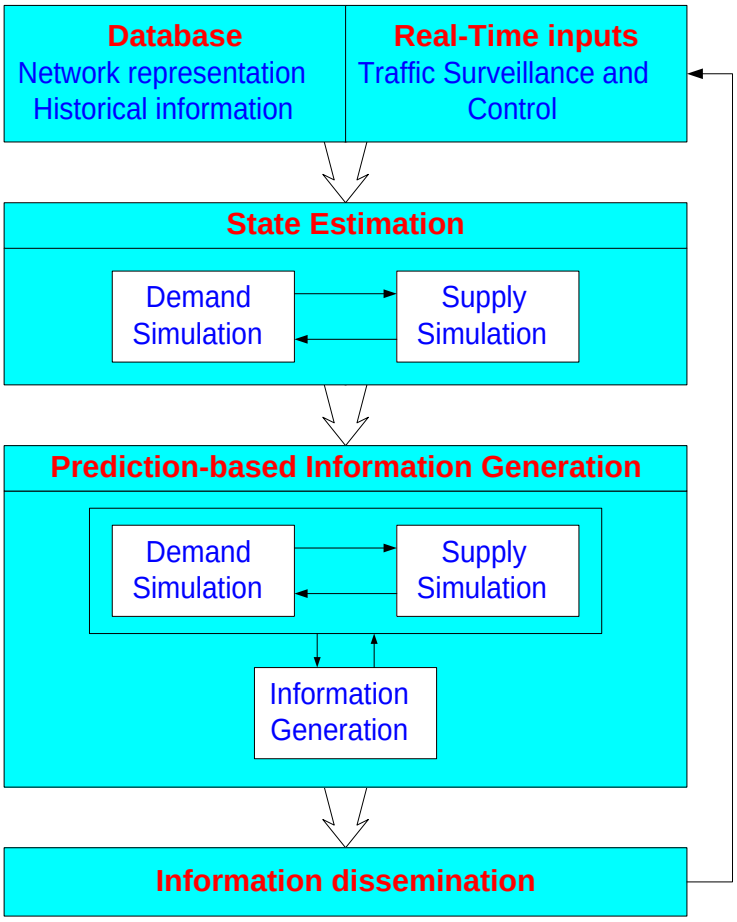

Figure A-2: The DynaMIT Framework

Real-time information is provided by the surveillance system and the control system. DynaMIT is designed to operate with a wide range of surveillance and control systems. The minimum real-time information required by DynaMIT is time-dependent link flows, incident characteristics (location, starting time, duration and

severity), and traffic control strategies.

State Estimation

The state estimation module provides estimates of the current state of the network in terms of OD flows, link flows, queues, speeds and densities. This step represents an important function of DTA systems, since information obtained from the traffic sensors can vary depending on the type of surveillance system employed. In an ideal system where there is two-way communication between the traffic control center and every vehicle in the network, perfect information about the vehicle location and possibly its origin and destination can be obtained. While such perfect systems are possible in the future, most existing surveillance systems are limited to vehicle detectors located at critical points in the network. The information provided by these traffic sensors therefore must be used to infer traffic flows, densities and queue lengths in the entire network.

The main models used by the State Estimation module are:

- A demand simulator that combines real-time OD estimation with user behavior models for route and departure time choice.
- A network state estimator (also known as the supply simulator) that simulates driver decisions and collects information about the resulting traffic conditions.

The demand and supply simulators interact with each other in order to provide demand and network state estimates that are congruent and utilize the most recent information available from the surveillance system (Figure A-3).

Demand Simulation

Demand estimation in DynaMIT is sensitive to the guidance generated and information provided to the users, and is accomplished through an explicit simulation of pre-trip departure time, mode and route choice decisions that ultimately produce the OD flows used by the OD estimation model. The pre-trip demand simulator updates

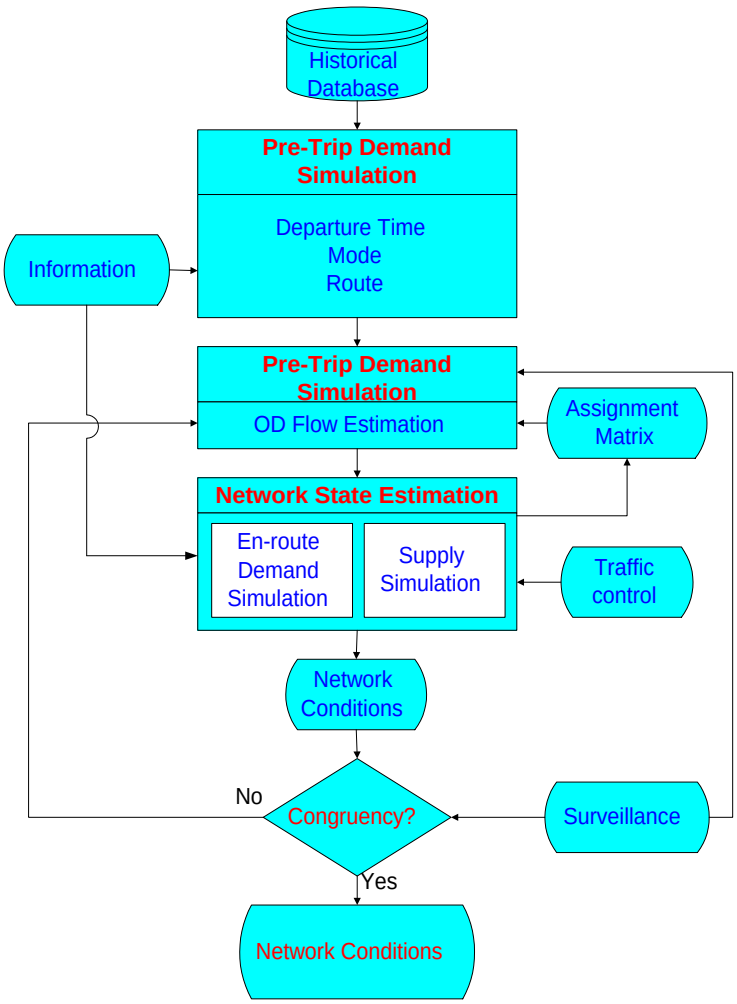

Figure A-3: State Estimation in DynaMIT

the historical OD matrices by modeling the reaction of each individual to guidance information. The consequent changes are then aggregated to obtain updated historical OD matrices. However, these updated historical OD flows require further adjustments to reflect the actual travel demand in the network. Reasons for the divergence of actual OD flows from historical estimates include capacity changes on the network (such as the closure of roads or lanes), special events that temporarily attract a large number of trips to a destination, and other day-to-day fluctuations. Consequently, one of the requirements for dynamic traffic modeling is the capability to estimate (and predict) OD flows in real time. The OD model uses updated historical OD flows, real-time measurements of actual link flows on the network, and estimates of assignment

fractions (the mapping from OD flows to link flows based on route choice fractions and travel times) to estimate the OD flows for the current estimation interval.

Note on OD Smoothing

The fixed point nature of the OD estimation procedure, coupled with the real-time requirements of a prediction-based DTA system, necessitates the use of an efficient solution scheme that will converge quickly. The OD estimation module within DynaMIT utilizes an algorithm similar to the Method of Successive Averages with Decreasing Re-initializations (MSADR)[1] to compute the target OD flows for successive iterations. Stated mathematically,

$$\mathbf{x}^{k^*} = \hat{\mathbf{x}}^{k-1} + \alpha_k(\hat{\mathbf{x}}^k - \hat{\mathbf{x}}^{k-1}) \quad (A.1)$$

$$= \alpha_k\hat{\mathbf{x}}^k + (1-\alpha_k)\hat{\mathbf{x}}^{k-1} \quad (A.2)$$

where $\mathbf{x}^{k^*}$ is the new target OD flow vector, $\hat{\mathbf{x}}^k$ and $\hat{\mathbf{x}}^{k-1}$ are the estimated flows from iterations k and $k-1$ respectively. The weighting parameter α_k is computed so as to accelerate the convergence of the iterative algorithm:

$$\alpha_k = \left[\frac{1}{ae^{-ka}}\right]\left[\frac{1}{\sum_{j=1}^{k}\frac{1}{ae^{-ja}}}\right] \quad (A.3)$$

where k is the iteration counter. The parameter a in the above expression assumes a default value of 1.0.

Supply Simulation

The network state estimator utilizes a traffic simulation model that simulates the actual traffic conditions in the network during the current estimation interval. The inputs to this model include the travel demand (as estimated by the demand simulator), updated capacities and traffic dynamics parameters, the control strategies implemented and the traffic information and guidance actually disseminated. The

[1] See Cascetta and Postorino (2001).

driver behavior model captures the responses to ATIS in the form of en route choices.

Demand-Supply Interactions

One of the inputs to the OD estimation model is a set of assignment matrices. These matrices map the OD flows from current and past intervals to link flows in the current interval. The assignment fractions therefore depend on the time interval, and also on the route choice decisions made by individual drivers. The flows measured on the network are a result of the interaction between the demand and supply components. It may be necessary to iterate between the network state estimation and the OD estimation models until convergence is achieved. The output of this process is an estimate of the actual traffic conditions on the network, and information about origin-destination flows, link flows, queues, speeds and densities.

A.1.3 Prediction and Guidance Generation

The prediction-based guidance module (Figure A-4) consists of several interacting steps:

- Pre-trip demand simulation
- OD flow prediction
- Network state prediction
- Guidance generation

The OD prediction model uses as input the aggregate historical demand adjusted by the pre-trip demand simulator to account for departure time, mode and route choices in response to guidance, and provides the required estimates of future OD flows. The network state prediction function undertakes the important task of traffic prediction for a given control and guidance strategy and predicted set of OD flows, using the current network conditions estimated by the state estimation module as

a starting point. The performance of the network over the prediction horizon is evaluated using a traffic simulation model and en-route behavioral models.

The traffic information and guidance generation function uses the predicted traffic conditions to generate information and guidance according to the various ATIS in place. Traffic control is loosely coupled with DynaMIT in the current version of the system. Control strategies are assumed to be generated outside the DTA system, using the predictions as an input.

The generated traffic information and guidance must be consistent and unbiased. Under such conditions, there would be no better path that a driver could have taken based on the provided information. An iterative process is employed in order to obtain guidance that satisfies these requirements. Each iteration consists of a trial strategy, the state prediction (comprising both demand prediction and network state prediction) under the trial strategy, and the evaluation of the predicted state for consistency. Since, in general, the updated historical OD flows depend on future guidance and information, the update of the historical OD flows (using the departure time and mode choice models) and the OD prediction models are included in the iteration. This general case represents the situation where pre-trip guidance is available to the drivers. In the special case where only en-route guidance is available, the pre-trip demand simulator is bypassed in the iterations. The initial strategy could then be generated from the prediction and guidance generation of the previous period.

A.2 Overview of DynaMIT-P

Apart from its real-time applications, DTA has the potential to significantly improve the transportation planning process for networks with congested facilities. DynaMIT-P is a DTA-based planning tool developed at MIT that is designed to assist planners in making decisions regarding proposed investments and operational changes in local and regional transportation networks. DynaMIT-P efficiently adapts the modules contained in the real-time DynaMIT system for off-line planning applications.

A.2.1 Features and Functionality

DynaMIT-P is designed to assist the evaluations of proposed changes to local and regional transportation networks. Such changes could affect infrastructure, operations, or information. Through the continual interaction between the demand simulator and the supply simulator, DynaMIT-P can effectively achieve:

- Predictions of day-to-day evolutions of travel demand and network conditions.
- Predictions of within-day patterns of traffic flows and travel times.
- Comparisons of different alternatives.

DynaMIT-P incorporates both equilibrium and day-to-day learning into its core operations involving the modeling of demand and supply interactions. Its features include:

- Modular structure with a mesoscopic traffic simulator and a demand microsimulator.
- Expands on the real time DynaMIT System.
- Flexible modeling of demand-supply interactions including both equilibrium algorithms and day-to day learning behavior.
- Uses disaggregate behavioral models.
- Supply simulator uses speed-density functions and queuing and captures the locations and impacts of queues and spillbacks.
- Behavioral models capture the inherent stochasticity of transportation demand.
- Captures the effects of segment-level operational changes, such as ramp meters and traffic signals.
- Predicts the effects of the introduction or enhancement of Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) on travel behavior and network performance.

- Enumeration of drivers facilitates fine distinctions among vehicle types and driver behaviors.
- Distinguishes between informed and uninformed drivers.
- Distinguishes between long-term, short-term, and within-day behaviors.
- Different and interacting approaches for modeling habitual and switching behavior.
- Visualizes and graphically compares alternative strategies using flow/speed/density (over time) charts.

A.2.2 Overall Framework

Travel-related choices vary with regard to the time horizon over which they are made. Individuals make long-term, short-term and within-day travel decisions (Figure A-5). Long-term mobility decisions could include choices on residential location and auto ownership. Short-term (or day-to-day) travel decisions include choice of trip frequency, destination, departure time, mode and route. Adjustments in short-term decisions are made in response to changes in long-term decisions (such as auto ownership) and changes in the network. Individuals form habitual travel patterns that they follow regularly. Within-day decisions capture deviations from these habitual travel patterns. These deviations could be in response to real-time information, unusual weather conditions, incidents, or other special events.

DynaMIT-P focuses on modeling the short-term and within-day travel decisions, assuming that the long-term decisions are given. The inputs to DynaMIT-P include the potential users of the system, their demographic characteristics, residential location, etc. The output is the performance of the transportation system in terms of consumption of resources and benefits. Several important characteristics distinguish DynaMIT-P from traditional planning approaches:

- Microsimulation ensures accurate depiction of individual traveler behavior.

- Detailed modeling of spillbacks, queue formation and dissipation captures the essence of network dynamics.
- Sensitivity to ATMS/ATIS facilitates the evaluation of ITS strategies.
- Time-dependent interactions between the demand and supply components presents a realistic picture of equilibrium.

DynaMIT-P employs three main components to achieve the functionality described above:

- The supply simulator
- The demand simulator
- The day-to-day learning model

The supply simulator is a mesoscopic traffic simulation model. For a given set of travelers and control strategies, it predicts the performance of the network by measuring time-dependent flows, travel times, queue lengths, etc. The simulator is designed to operate at different levels of granularity, depending on the requirements of each application. The main elements of the demand simulator are the OD matrix estimation and the behavioral models. The OD estimation model takes link counts and historical OD flows as inputs, and produces an updated time-dependent OD matrix to match the observed counts. The behavioral models are used to predict the travel behavior of individual travelers as a function of network level of service characteristics, perceptions and past experiences, information access and socioeconomic characteristics. Driver behavior is modeled using the path-size logit model (PS-Logit, Ramming (2001)), which is an extension of C-Logit (Cascetta and Russo (1997)). This model accounts for the degree of overlap among alternative routes while simulating individual route choice. The day-to-day learning model updates travelers perceptions of travel times based on past experiences and expectations, according to the following model:

$$\bar{T}_k^t = \lambda T_k^{t-1} + (1-\lambda)\bar{T}_k^{t-1} \tag{A.4}$$

where $\bar{T}_k^t$ is the expected time-dependent travel time along path k on day t, and T_k^t is the time-dependent travel time experienced along path k on day t. λ captures the learning rate, which may vary across market segments. The value of λ lies between 0 and 1, and is affected by the use of ATIS.

The demand and supply simulators interact with the learning models in a systematic way to capture both the day-to-day and within-day (short-term) demand-supply interactions (Cantarella and Cascetta (1995)). The structure of the short-term dynamics module is shown in Figure A-6. The model is based on an iterative process between the demand and supply simulators. The main input to short-term dynamics is an OD matrix of potential travelers. The demand simulator uses the corresponding behavioral models to update their frequency, destination, departure time, mode, and route, choices. The travelers are then loaded onto the supply simulator and new network performance is obtained. Based on the learning model, travelers update their decisions in response to the observed level of service and network performance. When supply and demand converge, the process ends. The output of the short-term dynamics component is the travelers' habitual travel behavior.

The purpose of the within-day dynamics model is to evaluate the performance of the transportation network in the presence of stochastic factors such as unusual weather, incidents, and special events (concerts, sports, etc.), which could substantially affect traffic conditions. The habitual travel behavior, obtained from the short-term dynamics, is input to the within-day model. Figure A-7 summarizes the interactions among the different elements of the within-day dynamics component.

The outputs from both the short-term and within-day behavior components are used to generate the desired resource consumption and benefits (such as total savings in travel delays, costs, revenues, air pollution, safety, fuel consumption, etc.) DynaMIT-P's open system of demand models, detailed representation of network dynamics, and flexible structure make it a useful tool for a host of planning applications:

- Impact studies of Work Zone Activity, and minimum-impact work zone scheduling

- Special Events
- High-occupancy Vehicle (HOV) and High-occupancy Toll (HOT) facilities
- Congestion Pricing strategies
- Effectiveness of ATMS and ATIS

While both DynaMIT and DynaMIT-P capture the interaction between the OD estimation and route choice model components within the DTA system, DynaMIT-P is better suited for calibration purposes. This stems primarily from DynaMIT-P's functionality to compute equilibrium network travel times. Network equilibrium for a given travel demand level results from a balance between the demand and supply elements. The fixed point nature of the calibration problem requires equilibrium at each calibration stage in order to ensure that the route choice fractions, assignment matrices and estimated OD flows are consistent in each iteration.

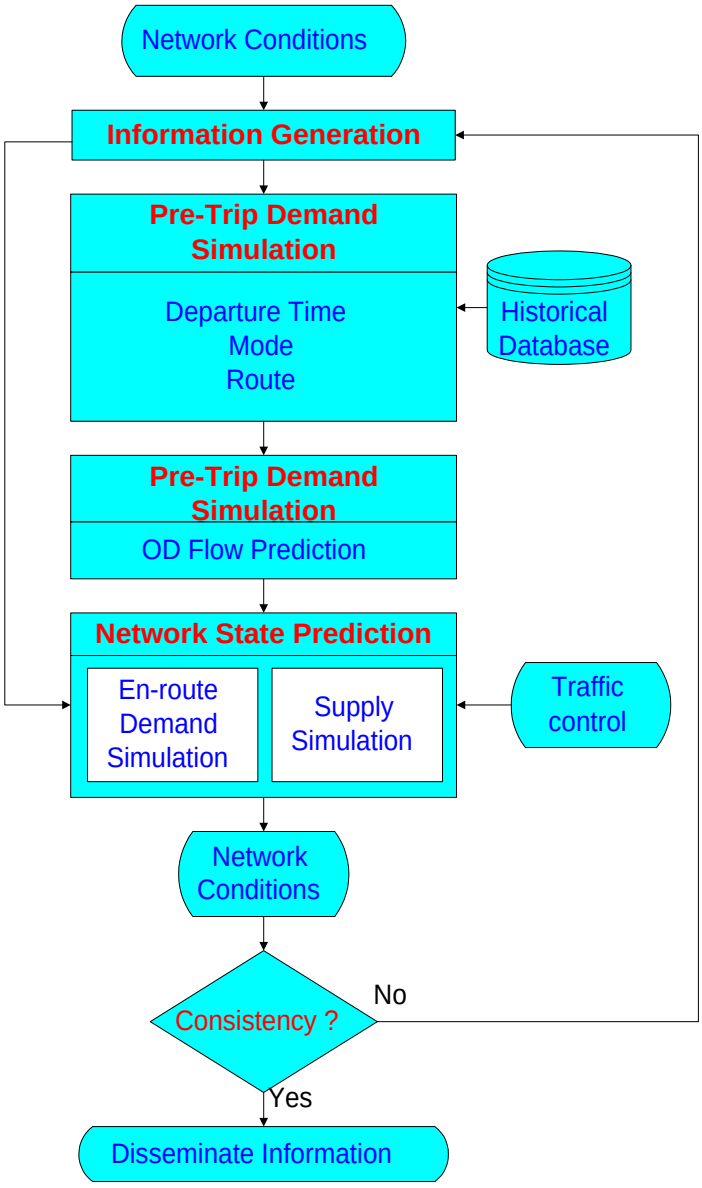

Figure A-4: Prediction and Guidance Generation in DynaMIT

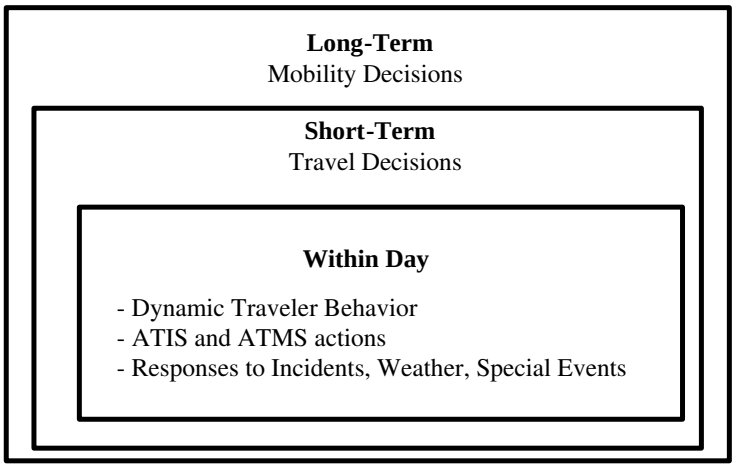

Figure A-5: Framework for Travel Behavior

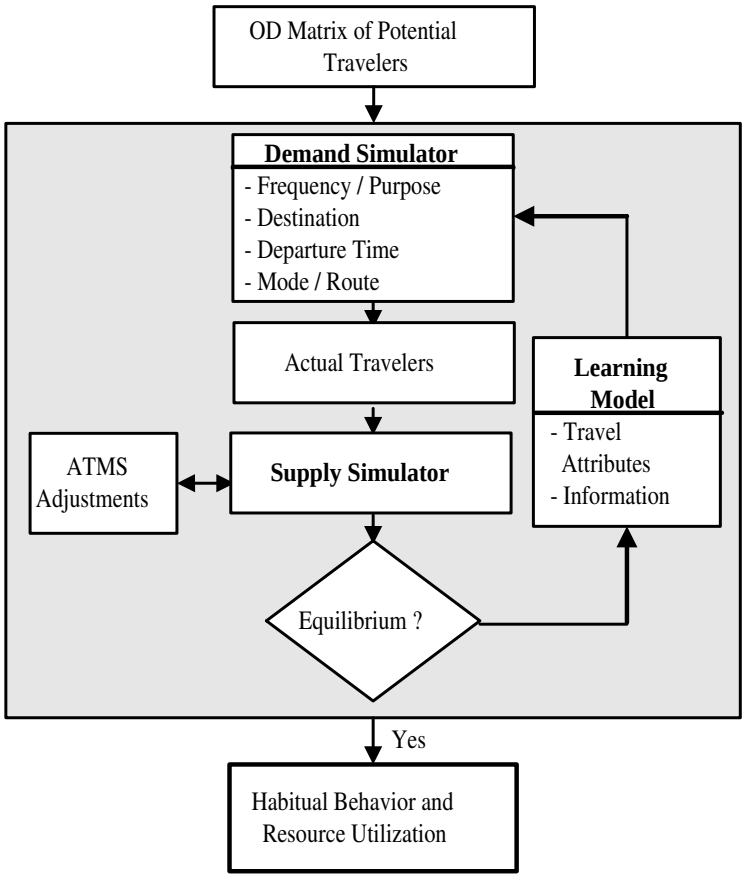

Figure A-6: Short-Term Dynamics

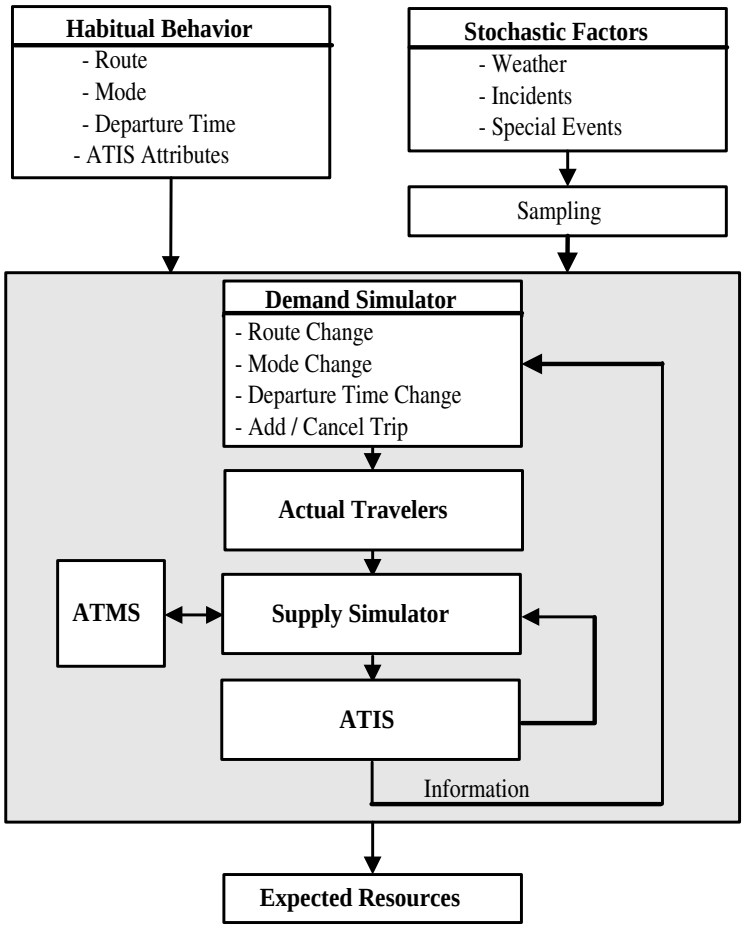

Figure A-7: Within-Day Dynamics

Appendix B

Prototypical Evaluation: Detailed Numerical Results

B.1 Fit to counts, speeds and OD flows

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	15.14	3.93	-
S	13.75	4.59	-	16.97	2.33	-
SD	12.94	4.84	3.95	16.50	2.01	2.67

Table B.1: Run 2

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	19.77	3.85	-
S	18.66	2.89	-	19.37	2.72	-
SD	17.74	3.00	3.87	18.44	2.18	3.02

Table B.2: Run 3

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	16.24	4.29	-
S	14.40	2.68	-	16.18	1.57	-
SD	13.10	2.60	6.14	15.45	1.56	2.72

Table B.3: Run 4

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	17.14	3.14	-
S	13.32	3.62	-	17.88	2.63	-
SD	12.91	2.81	3.21	16.52	2.17	3.15

Table B.4: Run 5

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	15.00	5.22	-
S	14.00	2.82	-	14.64	1.79	-
SD	12.40	3.05	5.72	14.19	1.75	5.19

Table B.5: Run 6

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	18.73	3.65	-
S	13.57	3.31	-	18.34	2.33	-
SD	12.89	3.36	9.15	16.43	2.07	3.43

Table B.6: Run 7

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	20.42	5.17	-
S	18.60	3.18	-	18.68	3.03	-
SD	18.04	3.33	11.98	18.60	2.99	5.46

Table B.7: Run 8

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	13.83	4.03	-
S	12.54	3.41	-	13.73	1.71	-
SD	12.15	5.62	3.69	13.62	1.69	2.41

Table B.8: Run 9

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	20.94	4.01	-
S	17.43	3.52	-	20.62	2.17	-
SD	17.33	3.36	5.99	19.02	2.07	2.09

Table B.9: Run 10

Scenario	Calibration Data					
	Counts			Counts + Speeds		
	$RMSE^c$	$RMSE^s$	$RMSE^d$	$RMSE^c$	$RMSE^s$	$RMSE^d$
Base	-	-	-	20.20	4.48	-
S	15.12	3.80	-	20.14	2.26	-
SD	14.36	4.35	7.05	19.91	2.19	2.84

Table B.10: Run 11

Bibliography

Abdulhai, B., Sheu, J. B., and Recker, W. (1999). Simulation of ITS on the Irvine FOT Area using the 'PARAMICS 1.5' Scalable Microscopic Traffic Simulator. Phase I: Model Calibration and Validation. Technical report, PATH research report, UCB-ITS-PRR-99-12.

Ahmed, K. (1999). *Modeling Drivers' Acceleration and Lane Changing Behavior*. PhD thesis, Massachusetst Institute of Technology.

Antoniou, C. (2002). Development of ITS Analysis Tools for Lower Westchester County. Technical report, Intelligent Transportation Systems Program, Massachusetts Institute of Technology.

Antoniou, C. (2004). *On-Line Calibration for Dynamic Traffic Assignment*. PhD thesis, Massachusetts Institute of Technology.

Antoniou, C., Ben-Akiva, M., Bierlaire, M., and Mishalani, R. (1997). Demand Simulation for Dynamic Traffic Assignment. In *8th IFAC Symposium on Transportation Systems, Chania, Greece*.

Ashok, K. (1996). *Estimation and Prediction of Time-Dependent Origin-Destination Flows*. PhD thesis, Massachusetts Institute of Technology.

Ashok, K. and Ben-Akiva, M. E. (2000). Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin-Destination Flows. *Transportation Science*, 34(1):21–36.

Balakrishna, R. (2002). Calibration of the Demand Simulator within a Dynamic Traffic Assignment System. Master's thesis, Massachusetts Institute of Technology.

Balakrishna, R., Ben-Akiva, M., Wen, Y., and Ashok, K. (2006). Observability in Estimating Time-Dependent Origin-Destination Flows from Traffic Counts. Accepted for presentation at DTA 2006, Leeds, UK.

Balakrishna, R., Ben-Akiva, M. E., Koutsopoulos, H. N., and Toledo, T. (2004). Traffic Simulation Model Calibration Framework Using Aggregate Data. Proceedings of the Triennial Symposium on Transportation Analysis (TRISTAN V).

Balakrishna, R., Koutsopoulos, H. N., and Ben-Akiva, M. (2005a). Calibration and Validation of Dynamic Traffic Assignment Systems. In Mahmassani, H. S., editor, *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction*, pages 407–426. Proceedings of the 16$^{\text{th}}$ International Symposium on Transportation and Traffic Theory, Elsevier.

Balakrishna, R., Koutsopoulos, H. N., Ben-Akiva, M., Ruiz, B. M. F., and Mehta, M. (2005b). A Simulation-Based Evaluation of Advanced Traveler Information Systems. *Transportation Research Record*.

Barcelo, J. and Casas, J. (2002). Dynamic Network Simulation with AIMSUN. In *Proceedings of the International Symposium on Transport Simulation, Yokohama*. Kluwer.

Barcelo, J., Casas, J., Garcia, D., and Perarnau, J. (2005). Methodological Notes on Combining Macro, Meso and Micro Models for Transportation Analysis. In *Presented at the Sedona Modeling Workshop*.

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete Choice Models with Applications to Departure Time and Route Choice. In Hall, R., editor, *Handbook of Transportation Science, 2nd edition*. Kluwer.

Ben-Akiva, M., Bierlaire, M., Bottom, J., Koutsopoulos, H. N., and Mishalani, R. G. (1997). Development of a Route Guidance Generation System for Real-Time Ap-

plication. Proceedings of the Eighth IFAC Symposium on Transportation Systems, Chania, Greece.

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H. N., and Mishalani, R. (2001). Network State Estimation and Prediction for Real-Time Transportation Management Applications. *Networks and Spatial Economics*, 1(3/4):291–318.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H. N., and Mishalani, R. (2002). Real-Time Simulation of Traffic Demand-Supply Interactions within DynaMIT. In Gendreau, M. and Marcotte, P., editors, *Transportation and Network Analysis: Miscellenea in honor of Michael Florian*, pages 19–36. Kluwer.

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge, Massachusetts.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Bierlaire, M. and Crittin, F. (2004). An Efficient Algorithm for Real-Time Estimation and Prediction of Dynamic OD Tables. *Operations Research*.

Birge, J. R. and Ho, J. K. (1993). Optimal Flows in Stochastic Dynamic Networks with Congestion. *Operations Research*, 41(1):203–216.

Box, M. J. (1965). A New Method of Constrained Optimization and a Comparison With Other Methods. *Computer Journal*, 8(1):42–52.

Boyce, D., Lee, D.-H., and Ran, B. (2001). Analytical Models of the Dynamic Traffic Assignment Problem. *Networks and Spatial Economics*, 1(1):377–390.

Boyce, D. E., Ran, B., and LeBlanc, L. J. (1995). Solving an Instanteneous Dynamic User-Optimal Route Choice Model. *Transportation Science*, 29:128–142.

Bradley, S. P., Hax, A. C., and Magnanti, T. L. (1977). *Applied Mathematical Programming*. Addison-Wesley Pub. Co., Reading, Massachusetts.

Brockfeld, E., Kuhne, R. D., and Wagner, P. (2005). Calibration and Validation of Microscopic Traffic Flow Models. Presented at the 84th annual meeting of the Transportation Researc Board.

Caliper (2006). About Transmodeler. http://caliper.com/transmodeler/default.htm, accessed on 27 April 2006.

Cantarella, G. and Cascetta, E. (1995). Dynamic Processes and Equilibrium in Transportation Networks: Towards a Unifying Theory. *Submitted to Transportation Science*, 29(4).

Carey, M. (1987). Optimal Time Varying Flows on Congested Networks. *Operations Research*, 35(1):58–69.

Carey, M. and Subrahmanian, E. (2000). An Approach to Modeling Time-Varying Flows on Congested Networks. *Transportation Research*, 34B(3):157–183.

Carson, Y. and Maria, A. (1997). Simulation Optimization: Methods and Applications. Proceedings of the 1997 Winter Simulation Conference, pages 118–126.

Cascetta, E. (1984). Estimation of Trip Matrices from Traffic Counts and Survey Data: A Generalized Least Squares Estimator. *Transportation Research*, 18B(4/5):289–299.

Cascetta, E., Inaudi, D., and Marquis, G. (1993). Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts. *Transportation Science*, 27(4):363–373.

Cascetta, E. and Nguyen, S. (1988). A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research*, 22B(6):437–455.

Cascetta, E., Nuzzolo, A., Russo, F., and Vitetta, A. (1996). A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and some Calibration Results for Interurban Networks. Proceedings of the 13th International

Symposium on Transportation and Traffic Theory, Leon, France, pages 697–712. Pergamon.

Cascetta, E. and Postorino, M. (2001). Fixed Point Approaches to the Estimation of O/D Matrices using Traffic Counts on Congested Networks. *Transportation Science*.

Cascetta, E. and Russo, F. (1997). Calibrating Aggregate Travel Demand Models with Traffic Counts: Estimators and Statistical Performance. *Transportation*, 24:271–293.

Chen, H. K. and Hsueh, C. F. (1998). A Model and an Algorithm for the Dynamic User-Optimal Route Choice Problem. *Transportation Research*, 32B(3):219–234.

Chen, Y.-S., Giessen, T., Hendriks, H., and Mahmassani, H. S. (2004). Calibrating and Validating a Large-Scale Dynamic Traffic Assignment Model under European Context. Presented at the 83rd Annual Meeting of the Transportation Research Board.

Chu, L., Liu, H. X., Oh, J.-S., and Recker, W. (2004). A Calibration Procedure for Microscopic Traffic Simulation. Proceedings of the 83rd annual meeting of the Transportation Research Board.

Corana, A., Marchesi, M., Martini, C., and Ridella, S. (1987). Minimizing Multimodal Functions of Continuous Variables with the 'Simulated Annealing' Algorithm. *ACM Transactions on Mathematical Software*, 13:262–280.

Daganzo, C. (1994). The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research*, 28B(4):269–287.

Daigle, G., Thomas, M., and Vasudevan, M. (1998). Field Applications of CORSIM: I-40 Freeway Design Evaluation, Oklahoma City, OK. In *Winter Simulation Conference*.

Darda, D. (2002). Joint Calibration of a Microscopic Traffic Simulator and Estimation of Origin-Destination Flows. Master's thesis, Massachusetts Institute of Technology.

de Palma, A. and Marchal, F. (2002). Real Cases Applications of the Fully Dynamic METROPOLIS Tool-Box: An Advocacy for Large-Scale Mesoscopic Transportation Systems. *Networks and Spatial Economics*, 2(4):347–369.

FHWA (2005). CORSIM website, http://ops.fhwa.dot.gov/trafficanalysistools/corsim.htm. Accessed on 8 May 2005.

Friesz, T. L., Bernstein, D., Mehta, N. J., Tobin, R. L., and Ganjalizadeh, S. (1989). Dynamic Network Traffic Assignment Considered as a Continuous Time Optimal Control Problem. *Operations Research*, 37(6):893–901.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L., and Wie, B. W. (1993). A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Operations Research*, 41(1):179–191.

Fu, M. C. (2001). Simulation Optimization. Proceedings of the 2001 Winter Simulation Conference, pages 53–61.

Fu, M. C. (2002). Optimization for Simulation: Theory vs. Practice. *INFORMS Journal on Computing*, 14(3):192–215.

Gabriel Gomes and Adolf May and Roberto Horowitz (2004). A microsimulation model of a congested freeway using VISSIM. Presented at the 83rd annaul meeting of the Transportation Research Board.

Gill, P. E., Murray, W., and Wright, M. H. (1984). *Practical Optimization*. Academic Press.

Goffe, W. L., Ferrier, G. D., and Rogers, J. (1994). Global Optimization of Statistical Functions with Simulated Annealing. *Journal of Econometrics*, 60(1-2).

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, USA.

Gomes, G., May, A., and Horowitz, R. (2004). A Microsimulation Model of a Congested Freeway using VISSIM. Presented at the 83rd annual meeting of the Transportation Research Board.

Gupta, A. (2005). Observability of Origin-Destination Matrices for Dynamic Traffic Assignment. Master's thesis, Massachusetts Institute of Technology.

Halati, A. and Boyce, D. E. (1991). Framework of Simulation Evaluation of In-Vehicle Navigation Systems. Presented at the 71st annual meeting ofthe Transportation Research Board.

Halati, A., Boyce, D. E., and Torres, J. F. (1991). Comparative Assessment of the Potential Traffic Benefits of the In-Vehicle Navigation and Information Systems. Presented at the 71st annual meeting of the Transportation Research Board.

Hall, M. D., Vliet, D. V., and Willumsen, L. G. (1980). SATURN - A Simulation Assignment Model for the Evaluation of Traffic Management System Schemes. *Traffic Engineering Control*, 21:167–176.

Hawas, Y. E. (2002). Calibrating Simulation Models for Advanced Traveler Information Systems/Advanced Traffic Management Systems Applications. *Journal of Transportation Engineering*, 128(1):80–88.

Hazelton, M. L. (2000). Estimation of Origin-Destination Matrices from Link Flows on Uncongested Networks. *Transportation Research*, 34B:549–566.

Hazelton, M. L. (2001). Inference for Origin-Destination Matrices: Estimation, Prediction and Reconstruction. *Transportation Research*, 35B:667–676.

He, R., Miaou, S., Ran, B., and Lan, C. (1999). Developing an On-Line Calibration Process for an Analytical Dynamic Traffic Assignment Model. *Presented at the 78th Annual Meeting of the Transportation Research Board*.

He, R. and Ran, B. (2000). Calibration and Validation of a Dynamic Traffic Assignment Model. *Presented at the 79th Annual Meeting of the Transportation Research Board*.

Hegyi, A., Ngo Duy, D., De Schutter, B., Hellendoorn, J., Hoogendoorn, S., and Stramigioli, S. (2003). Suppressing Shock Waves on the A1 in The Netherlands - Model Calibration and Model-Based Predictive Control. In *Proceedings of the IEEE 6th International Conference on Intelligent Transportation Systems (ITSC'03)*, pages 672–677, Shanghai, China.

Henderson, J. and Fu, L. (2004). Applications of Genetic Algorithms in Transportation Engineering. Presented at the 83rd annual meeting of the Transportation Research Board.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, USA.

Hooke, R. and Jeeves, T. A. (1961). Direct Search Solution of Numerical and Statistical Problems. *Journal of the ACM*, 8:212–229.

Huyer, W. and Neumaier, A. (2004). SNOBFIT - Stable Noisy Optimization by Branch and Fit. *Submitted to ACM Transactions on Mathematical Software*.

INRO (2006). About EMME/2. http://www.inro.ca/en/products/emme2/index.php, accessed on 4 May 2006.

Janson, B. N. (1991). Dynamic Traffic Assignment for Urban Road Networks. *Transportation Research*, 25B(2/3):143–161.

Jha, M., Gopalan, G., Garms, A., Mahanti, B. P., Toledo, T., and Ben-Akiva, M. (2004). Development and Calibration of a Large-Scale Microscopic Traffic Simulation Model. *Transportation Research Record*, 1876:121–131.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic Estimation of the Maximum of a Regression Function. *Annals of Mathematical Statistics*.

Kim, K.-O. (2002). *Optimization Methodology for the Calibration of Transportation Network Micro-Simulation Models*. PhD thesis, Texas A&M University.

Kim, K.-O. and Rilett, L. R. (2003). Simplex-Based Calibration of Traffic Microsimulation Models with Intelligent Transportation Systems Data. *Transportation Research Record*, 1855:80–89.

Kim, K.-O. and Rilett, L. R. (2004). A Genetic Algorithm Based Approach to Traffic Micro-Simulation Calibration Using ITS Data. Presented at the 83rd annual meeting of the Transportation Research Board.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, 220:671–680.

Kleijnen, J. P. C. (1987). *Statistical Tools for Simulation Practicioners*. Marcel Dekker.

Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods. *SIAM Review*, 45(3):385–482.

Kunde, K. K. (2002). Calibration of Mesoscopic Traffic Simulation Models for Dynamic Traffic Assigment. Master's thesis, Massachusetts Institute of Technology.

Kurian, M. (2000). Calibration of a Microscopic Traffic Simulator. Master's thesis, Massachusetts Institute of Technology.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 9(1):112–147.

Leclercq, L. (2005). Calibration of Flow-Density Relationships in Urban Streets. Presented at the 84th annual meeting of the Transportation Research Board.

Lee, D. H., Yang, X., and Chandrasekhar, P. (2001). Parameter Calibration for PARAMICS using Genetic Algorithm. Presented at the 80th annual meeting of the Transportation Research Board.

Lee, S., Kim, J., and Chen, A. (2006). Dynamic Travel Demand Estimation Using Real-Time Traffic Data. Presented at the 85th annual meeting of the Transportation Research Board.

Leonard, D., Gower, P., and Taylor, N. B. (1989). CONTRAM: Structure of the Model. Technical report, Transportation an Road Research Laboratory, Crowthorne, UK. TRRL Research Report RR 178.

Leonard, D., Tough, J. B., and Baguley, P. C. (1978). CONTRAM: A Traffic Assignment Model for Predicting Flows and Queues During Peak Periods. Technical report, Transport and Road Research Laboratory, Crowthorne, UK. TRRL Report LR 841.

Liu, H. X., Ding, L., Ban, J. X., Chen, A., and Chootinan, P. (2004). A Streamlined Network Calibration Procedure for California SR41 Corridor Traffic Simulation Study. Presented at the 83rd annual meeting of the Transportation Research Board.

Liu, Y. H. and Mahmassani, H. S. (2000). Global Maximum Likelihood Estimation Procedure for Multinomial Probit (MNP) Model Parameters. *Transportation Research*, 34B:419–449.

Lo, H.-P. and Chan, C.-P. (2003). Simultaneous Estimation of an Origin-Destination Matrix and Link Choice Proportions using Traffic Counts. *Transportation Research*, 37A:771–788.

Mahanti, B. P. (2004). Aggregate Calibration of Microscopic Traffic Simulation Models. Master's thesis, Massachusetts Institute of Technology.

Mahmassani, H., Qin, X., and Zhou, X. (2004). DYNASMART-X Evaluation for Real-Time TMC Application: Irvine Test Bed. Technical report, Maryland Transportation Initiative, University of Maryland.

Mahmassani, H. S. (2002). Dynamic Network Traffic Assignment and Simulation Methodology for Advanced System Management Applications. Presented at the 81st annual meeting of the Transportation Research Board.

Mahmassani, H. S., Sbayti, H., Victoria, I., Zhou, X., and Chiu, Y.-C. (2003). DYNASMART-P Calibration and Evaluation. Technical report, Maryland Transportation Initiative, University of Maryland.

Mahut, M., Florian, M., Florian, D., Velan, S., and Tremblay, N. (2005). Equilibrium Dynamic Traffic Assignment for Large, Congested Networks. Technical report, INRO. White paper.

Mahut, M., Florian, M., Tremblay, N., Campbell, M., Patman, D., and McDaniel, Z. K. (2004). Calibration and Application of a Simulation-Based Dynamic Traffic Assignment Model. *Transportation Research Record*, 1876:101–111.

Mardle, S. and Pascoe, S. (1999). An Overview of Genetic Algorithms for the Solution of Optimisation Problems. *Computers in Higher Education Review*, 13(1).

May, A. D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, New Jersey 07632.

Mcdougall, W. and Millar, G. (2001). Evaluation of road safety benefits with PARAMICS. Hobart, Tasmania, Australia.

Merchant, D. K. and Nemhauser, G. L. (1978). A Model and an Algorithm for the Dynamic Traffic Assignment Problems. *Transportation Science*, 12(3):183–199.

Messmer, A. and Papageorgiou, M. (2001). Freeway Network Simulation and Dynamic Traffic Assignment with METANET Tools. *Transportation Research Record*, (1776):178–188.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092.

MIT (2002). Development of a Deployable Dynamic Traffic Assignment System - Evaluation Report (Part A): Evaluation of Estimation and Prediction Capabilities. Technical report, Intelligent Transportation Systems Program, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Mouskos, K., Niver, E., Pignataro, L., Lee, S., Antoniou, N., and Papadopoulos, L. (1998). TRANSMIT System Evaluation. Technical report, New Jersey Institute of Technology.

Muñoz, L., Sun, X., Sun, D., Gomes, G., and Horowitz, R. (2004). Methodological Calibration of the Cell Transmission Model. In *Proceedings of the Annual Control Conference*, pages 798–803, Boston, MA.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *Computer Journal*, 7(4):308–313.

Ngoduy, D. and Hoogendoorn, S. P. (2003). An Automated Calibration Procedure for Macroscopic Traffic Flow Models. Proceedings of the 10^{th} IFAC CTS, Tokyo.

Ngoduy, D., Hoogendoorn, S. P., and Van Zuylen, H. J. (2006). Cross-Comparison of Numerical Schemes for Macroscopic Traffic Flow Models. Presented at the 85^{th} annual meeting of the Transportation Research Board.

Park, B. B., Pampati, D. M., Cao, J., and Smith, B. L. (2005). Field Evaluation of DynaMIT in Hampton Roads, Virginia. Presented at the 84th annual meeting of the Transportation Research Board.

Peeta, S. (1994). *System Optimal Dynamic Traffic Assignment in Congested Networks with Advanced Information Systems*. PhD thesis, University of Texas at Austin.

Peeta, S. and Zhou, C. (2002). A Hybrid Deployable Dynamic Traffic Assignment Framework for Robust Online Route Guidance. *Networks and Spatial Economics*.

Peeta, S. and Ziliaskopoulos, A. K. (2001). Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, pages 233–265.

Pflug, G. C. (1996). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Kluwer Academic Publishers.

PTV (2006). http://www.english.ptv.de/cgi-bin/traffic/traf_vision.pl, accessed on 4 May 2006.

Ramming, M. S. (2001). *Network Knowledge and Route Choice*. PhD thesis, Massachusetts Institute of Technology.

Ran, B. and Boyce, D. E. (1996). A Link-Based Variational Inequality Formulation of Ideal Dynamic User-Optimal Route Choice Problem. *Transportation Research*, 4C(1):1–12.

Ran, B., Boyce, D. E., and LeBlanc, L. J. (1993). A New Class of Instantaneous Dynamic User-Optimal Traffic Assignment Models. *Operations Research*, 41(1).

Ran, B. and Shimazaki, T. (1989). A General Model and Algorithm for the Dynamic Traffic Assignment Problems. Proceedings of the Fifth World Conference on Transport Research, Yokohama, Japan.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics*.

Smith, M., Duncan, G., and Druitt, S. (1995). PARAMICS: Microscopic Traffic Simulation for Congestion Management. In *Colloquium on Dynamic Control of Strategic Inter-Urban Road Networks*, London, England.

Spall, J. C. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341.

Spall, J. C. (1994a). A Second Order Stochastic Approximation Algorithm Using Only Function Measurements. In *Proceedings of the 33rd Conference on Decision and Control, Lake Buena Vista, FL*, pages 2472–2477.

Spall, J. C. (1994b). Developments in Stochastic Optimization Algorithms with Gradient Approximations based on Function Measurements. In Tew, J. D., Manivannan, S., Sadowski, D. A., and Seila, A. F., editors, *Proceedings of the 1994 Winter Simulation Conference*.

Spall, J. C. (1998a). An Overview of the Simultaneous Perturbation Method for Efficient Optimization. *Johns Hopkins APL Technical Digest*, 19(4):482–492.

Spall, J. C. (1998b). Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3):817–823.

Spall, J. C. (1999). Stochastic Optimization, Stochastic Approximation and Simulated Annealing. In Webster, J. G., editor, *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 529–542. Wiley-Interscience.

Sundaram, S. (2002). Development of a Dynamic Traffic Assignment System for Short-Term Planning Applications. Master's thesis, Massachusetts Institute of Technology.

Swisher, J. R., Hyden, P. D., Jacobson, S. H., and Schruben, L. W. (2000). A Survey of Simulation Optimization Techniques and Procedures. Proceedings of the 2000 Winter Simulation Conference, pages 119–128.

Taylor, N. B. (2003). The CONTRAM Dynamic Traffic Assignment Model. *Networks and Spatial Economics*, 3:297–322.

The Weather Underground, Inc. (2004). Weather Underground. http://www.wunderground.com. Accessed 22nd April 2006.

Toledo, T., Ben-Akiva, M. E., Darda, D., Jha, M., and Koutsopoulos, H. N. (2004). Calibration of Microscopic Traffic Simulation Models with Aggregate Data. *Transportation Research Record*, 1876:10–19.

Toledo, T., Koutsopoulos, H. N., Davol, A., Ben-Akiva, M. E., Burghout, W., Andreasson, I., Johansson, T., and Lundin, C. (2003). Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study. *Transportation Research Record*, (1831):65–75.

TRB (2000). *Highway Capacity Manual*. Transportation Research Board.

Tsavachidis, M. (2000). Aggregate Analysis of Driver Response to Collective Route Guidance and Implications for System Control. In *Proceedings of the 10th International Conference on Road Transport Information and Control*. Commonwealth Institute, London, UK, Institution of Electrical Engineers, UK.

UC Berkeley and Caltrans (2005). Freeway Performance Measurement System (PeMS) 5.4. http://pems.eecs.berkeley.edu/Public, accessed 30th June 2005.

UMD (2005). DYNASMART Homepage. http://www.dynasmart.umd.edu/index.html, accessed 17th June 2005.

Van Aerde, M. and Rakha, H. (1995). TRAVTEK Evaluation Modeling Study. Technical report, Federal Highway Administration, US DOT.

van der Zijpp, N. (1996). *Dynamic Origin-Destination Matrix Estimation on Motorway Networks*. PhD thesis, Delft University of Technology. Department of Civil Engineering.

van der Zijpp, N. (2002). Software Package for Estimation of Dynamic OD Matrices. Delft University of Technology.

Vliet, D. V. (1982). SATURN - A Modern Assignment Model. *Traffic Engineering Control*, 23:578–581.

Wah, B. W. and Wang, T. (1999). Constrained Simulated Annealing with Applications in Nonlinear Continuous Constrained Global Optimization. In *11th International Conference on Tools with Artificial Intelligence*, page 381.

Wie, B. W. (1991). Dynamic Analysis of User Optimized Network Flows with Elastic Travel Demand. Presented at the 70th annual meeting of the Transportation Research Board.

Wilde, D. J. (1965). *Optimum Seeking Methods*. Prentice-Hall, Inc., Englewood Cliffs, N. J.

Yang, Q. and Koutsopoulos, H. N. (1996). A Microscopic Traffic Simulator for Evaluation of Dynamic Traffic Management Systems. *Transportation Research*, 4(C)(3):113–129.

Yang, Q., Koutsopoulos, H. N., and Ben-Akiva, M. E. (2000). A Simulation Model for Evaluating Dynamic Traffic Management Systems. *Transportation Research Record*, 1710:122–130.

Yu, L., Li, X., and Zhuo, W. (2005). GA-Based Calibration of VISSIM for Intercontinental Airport of Houston (IAH) Network Using GPS Data. Presented at the 84th annual meeting of the Transportation Research Board.

Yue, P. and Yu, L. (2000). Travel Demand Forecasting Models: A Comparison of EMME/2 and QRS II Using a Real-World Network. Technical report, Center for Transportation Training and Research, Texas Southern University.

Ziliaskopoulos, A. K. (2000). A Linear Programming Model for the Single Destination System Optimum Dynamic Traffic Assignment Problem. *Transportation Science*, 34(1):37–49.