

**Optimizing the Efficiency and Equity of Traffic Flow:**

**DEVELOPMENT OF A CONTROL STRATEGY FOR FREEWAY  
CORRIDORS USING DYNAMIC BAYESIAN DECISION NETWORKS**

A thesis submitted to  
the Faculty of Graduate and Postdoctoral Affairs  
in Partial Fulfillment of the requirements for the degree

Doctor of Philosophy

by

Jennifer Armstrong

Department of Civil and Environmental Engineering  
Carleton University

Ottawa-Carleton Institute of Civil and Environmental Engineering

September 2011

©2011 Jennifer Armstrong



Library and Archives  
Canada

Published Heritage  
Branch

395 Wellington Street  
Ottawa ON K1A 0N4  
Canada

Bibliothèque et  
Archives Canada

Direction du  
Patrimoine de l'édition

395, rue Wellington  
Ottawa ON K1A 0N4  
Canada

*Your file* *Votre référence*  
ISBN: 978-0-494-83238-7  
*Our file* *Notre référence*  
ISBN: 978-0-494-83238-7

**NOTICE:**

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

**AVIS:**

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

---

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

  
**Canada**

# Abstract

Ramp metering has the potential to improve the safety and efficiency of existing transportation infrastructure. By controlling the flow of vehicles onto the freeway, the extent of mainline congestion can be reduced, resulting in a system-wide reduction in travel time even with the introduction of ramp delays. However, despite the potential benefits of ramp meters, their use remains controversial, in part due to equity concerns: while some drivers must wait at entrance ramps, others are allowed to access the freeway with little or no delay.

To effectively balance competing objectives, new algorithms are needed – algorithms which leverage new technologies, bringing together the latest research from fields as diverse as artificial intelligence and industrial control. It is not enough to measure the success of an algorithm by how well it improves traffic flow. It must also meet the needs of system users, who may value ramp delays differently than delays due to freeway congestion, and who may be willing to trade-off some of the potential gains in operational performance for a system that operates more fairly.

In this research, a new ramp control algorithm was developed which balances efficiency and equity objectives. To capture the uncertainty inherent in traffic systems, the algorithm has been implemented as a dynamic Bayesian decision network, and incorporates a probabilistic model of freeway flow. This thesis describes the development of the algorithm, its key features, and the results of simulation tests which demonstrate its effectiveness compared to more traditional efficiency-maximizing algorithms such as ALINEA.

Comparison with the ALINEA algorithm suggests that the new algorithm is able to achieve similar congestion benefits when operated on an efficiency basis alone. When equity considerations are also included, the system operates less efficiently, but with ramp delays that are more fairly distributed amongst drivers. Such a system has the potential to enhance public support for ramp metering – support which is essential for the successful deployment of ramp metering in a Canadian context.

# Acknowledgements

This research work would not have been possible without the support of a number of individuals and organizations. In particular, I would like to offer my sincere thanks to:

- Prof. Ata Khan, my thesis supervisor, for his encouragement, advice, and feedback
- The City of Ottawa, for providing much of the data needed to develop the Ottawa model, specifically:
  - Kornel Mucsi (signal timing data)
  - Doug Bowron and Carolyn Feghali (traffic count data)
  - Mona Abouhenidy and Ahmad Subhani (origin-destination data)
- Chan Trinh, Keyur Hindocha, and Jeff Liu, for their enthusiastic assistance in coding the Ottawa model
- The support staff at PTV America, who helped resolve several VISSIM issues
- Morrison Hershfield Ltd., for allowing me the flexibility to pursue my research adventures while holding a full-time job
- My family and friends, who pushed me to finish when I would have given up, who cooked and cleaned for me so I could work all week-end, who sorted out my computer difficulties and provided a sounding board even when they had no idea what I was talking about, who put up with me through it all.

Thank-you.

# Table of Contents

<b>Abstract.....</b>	<b>ii</b>
<b>Acknowledgements .....</b>	<b>iii</b>
<b>Table of Contents .....</b>	<b>iv</b>
<b>List of Tables .....</b>	<b>vii</b>
<b>List of Figures.....</b>	<b>viii</b>
<b>List of Appendices.....</b>	<b>x</b>
<b>Abbreviations .....</b>	<b>xi</b>
<b>1 Introduction.....</b>	<b>1</b>
<b>2 Background .....</b>	<b>3</b>
2.1 Ramp Metering Defined .....	3
2.2 Benefits of Ramp Metering.....	5
2.3 Negative Impacts .....	6
2.4 Role of Traffic Diversion.....	7
2.5 Public Perceptions.....	11
2.6 Ramp Metering Algorithms .....	13
2.7 Limitations of Existing Methodologies .....	16
<b>3 Exploration of Research Needs.....</b>	<b>17</b>
3.1 Introduction.....	17
3.2 Opportunities for Improvement .....	17
3.2.1 Equity .....	17
3.2.2 Diversion .....	32
3.2.3 Integrated Operations .....	34
3.2.4 Summary.....	37
<b>4 Bayesian Networks As a Framework For Ramp Metering.....</b>	<b>39</b>
4.1 Overview.....	39
4.2 Introduction to Dynamic Bayesian Decision Networks.....	39
4.3 Bayesian Networks in Transportation.....	41

<b>5 Research Objectives.....</b>	<b>44</b>
5.1 Research Statement.....	44
5.2 Guiding Principles .....	47
<b>6 Research Methodology .....</b>	<b>48</b>
6.1 General Framework .....	48
6.2 Literature Review.....	49
<b>7 The Ramp Metering Algorithm.....</b>	<b>51</b>
7.1 General Algorithm Structure.....	51
7.2 Equity versus Efficiency .....	52
7.3 Bayesian Network Formulation .....	53
7.4 The Freeway Traffic Model .....	58
7.4.1 Prediction of Flow Breakdown.....	76
7.4.2 Demand Estimation .....	102
7.5 Inference Using Particle Filters .....	110
7.6 Formulation of the Utility Function.....	115
7.7 Solving the Control Problem .....	125
7.7.1 Overview .....	125
7.7.2 Predictive Control.....	126
7.7.3 Calculation of Control Parameters .....	128
<b>8 Validation of the Freeway Traffic Model .....</b>	<b>133</b>
8.1 Test Philosophy.....	133
8.2 The Freeway Test Network.....	133
8.2.1 Modelling of On-Ramp Merging.....	136
8.2.2 The Ramp Signal Controller in VISSIM.....	137
8.3 Methodology & Key Assumptions .....	139
8.4 Calibration of Model Parameters .....	141
8.5 Performance of the Freeway Traffic Model with Evidence.....	142
8.6 Performance of the Freeway Traffic Model without Evidence .....	149
<b>9 Algorithm Performance Under Test Conditions.....</b>	<b>157</b>
9.1 Overview of the Test Environment.....	157
9.2 Key Assumptions .....	161

9.3	Implementation Issues .....	162
9.4	Algorithm Performance .....	165
9.4.1	Base Case with No Ramp Metering .....	165
9.4.2	Utility based on Efficiency Only .....	168
9.4.3	Utility based on Efficiency and Equity .....	173
9.5	Comparison with the ALINEA Ramp Control Algorithm.....	176
9.5.1	Overview of the ALINEA Algorithm.....	177
9.5.2	Rationale for Selecting ALINEA as a Basis for Comparison .....	178
9.5.3	Implementation in the VISSIM Test Network .....	180
9.5.4	Determination of Control Parameters.....	181
9.5.5	Network Performance Under ALINEA Control.....	184
9.5.6	Comparison of Results .....	188
<b>10</b>	<b>Algorithm Performance in a Real-World Network .....</b>	<b>192</b>
10.1	General Approach .....	192
10.2	Development & Calibration of the Ottawa Simulation Model .....	192
10.2.1	Study Area .....	193
10.2.2	Simulation Period .....	195
10.2.3	Data Requirements .....	195
10.2.4	Model Development – Network Specification .....	196
10.2.5	Model Development – Travel Demand Estimation.....	196
10.2.6	Model Calibration / Validation.....	200
10.3	Implementation of the Ramp Metering Algorithm .....	206
10.4	Summary of Key Results .....	207
<b>11</b>	<b>Conclusions.....</b>	<b>211</b>
<b>12</b>	<b>Recommendations.....</b>	<b>214</b>
12.1	Overview.....	214
12.2	Potential Future Enhancements.....	214
<b>13</b>	<b>References.....</b>	<b>221</b>

## List of Tables

Table 2-1 Typical Ramp Metering Results.....	6
Table 7-1 Summary of Model Parameters .....	64
Table 7-2 Model Features, Assumptions & Key Issues.....	66
Table 7-3 Comparison of SAS and WinBUGS Parameter Estimates.....	95
Table 7-4 Flow Breakdown Model Parameters .....	96
Table 7-5 Recommended Freeway Performance Measures.....	117
Table 7-6 Utility Functions Used in Current Algorithm.....	122
Table 7-7 Multi-Attribute Utility Function Weights .....	125
Table 9-1 Algorithm Assumptions.....	161
Table 9-2 Network Performance Statistics – Base Case with No Ramp Metering .....	168
Table 9-3 Network Performance Statistics – New Algorithm (Efficiency Only).....	171
Table 9-4 Network Performance Statistics – New Algorithm (Efficiency + Equity)....	175
Table 9-5 ALINEA Results Under Different Target Occupancies ( $K_R = 70$ vph) .....	184
Table 9-6 Network Performance Statistics – ALINEA Algorithm.....	186
Table 9-7 Comparison of Average Network Travel Speed Results.....	191
Table 10-1 Peak Hour Volume Calibration Results .....	203
Table 10-2 Highway 417 Congestion Patterns during the Afternoon Peak Period .....	205
Table 10-3 Comparison of Algorithm Results for Average Network Travel Speed .....	209

# List of Figures

Figure 3-1 Lorenz Curve & Gini-Coefficient .....	22
Figure 6-1 Research Methodology.....	49
Figure 7-1 General Structure of the Ramp Metering Algorithm .....	52
Figure 7-2 Simplified Bayesian Network of the Freeway Control Problem .....	53
Figure 7-3 Bayesian Network Illustrating the Nodes Interacting with Segment ‘i’ .....	55
Figure 7-4 Flow Chart of the Probabilistic Freeway Traffic Model.....	61
Figure 7-5 Calculation of the Uncongested Travel Speed.....	63
Figure 7-6 Typical Results from the VISSIM Test Network.....	80
Figure 7-7 Counting Algorithm for Probability Updating.....	85
Figure 7-8 Flow Breakdown As Modelled in Netica.....	86
Figure 7-9 Probability of Breakdown Assuming Independent Parameters .....	87
Figure 7-10 Probability of Breakdown Based on Logit Formulation.....	93
Figure 7-11 Graphical Depiction of Breakdown Model for various Merging Lengths... ..	97
Figure 7-12 Allocation of Ramp Observations to Time Steps.....	104
Figure 7-13 Estimation of Future Demand .....	105
Figure 7-14 Graphical Illustration of a Particle Filter .....	112
Figure 7-15 Particle Filter Algorithm .....	113
Figure 7-16 Particle Filter Extension for Dynamic Bayesian Networks .....	114
Figure 7-17 MCMC Step .....	115
Figure 7-18 Impact of Utility Assessment Techniques for Freeway Speed .....	121
Figure 7-19 The Control Problem.....	126
Figure 8-1 The VISSIM Test Network .....	135
Figure 8-2a Model Performance With Evidence – Selected Segments .....	143
Figure 8-2b Model Performance With Evidence – Selected Segments.....	144
Figure 8-3 Model Performance Without Evidence – Selected Segments.....	155
Figure 8-4 Six-Minute Model Predictions Without Evidence – Selected Segments.....	156
Figure 9-1 VISSIM & MATLAB Integration.....	158
Figure 9-2 Algorithm Calculations – Real-World Network .....	159
Figure 9-3 Algorithm Calculations – Simulated Network.....	160
Figure 9-4 Congestion Maps – Base Case with No Ramp Metering.....	167
Figure 9-5 Average Ramp Travel Time – New Algorithm (Efficiency Only) .....	171

Figure 9-6 Congestion Maps – New Algorithm (Efficiency Only) .....	172
Figure 9-7 Adequacy of Ramp Storage – New Algorithm (Efficiency Only) .....	173
Figure 9-8 Average Ramp Travel Time – New Algorithm (Efficiency + Equity) .....	176
Figure 9-9 Average Ramp Travel Time – ALINEA Algorithm .....	186
Figure 9-10 Congestion Maps – ALINEA Algorithm .....	187
Figure 9-11 Adequacy of Ramp Storage – ALINEA Algorithm .....	188
Figure 9-12 Comparison of Ramp Travel Time Equity .....	191
Figure 10-1 Map of Study Area .....	194
Figure 10-2 Assumed Variation in Demand During the Peak Period .....	199
Figure 10-3 Comparison of Algorithm Results for Average Ramp Travel Time .....	210

# List of Appendices

APPENDIX A – Design of Ramp Metering Systems.....	232
APPENDIX B – Ramp Metering Mechanisms.....	241
APPENDIX C – Introduction to Bayesian Networks .....	252
APPENDIX D – Bayesian Network Example .....	264
APPENDIX E – Inference Techniques for Belief Updating in Bayesian Networks .....	268
APPENDIX F – Macroscopic Models of Traffic Flow .....	277
APPENDIX G – The Freeway Traffic Model .....	282
APPENDIX H – Development of Flow Breakdown Model.....	291
APPENDIX I – Introduction to Utility Theory.....	303
APPENDIX J – Assessment of Utility for Ramp Control .....	313
APPENDIX K – The VISSIM Test Network .....	326
APPENDIX L – Options to Improve VISSIM On-Ramp Merge Behaviour.....	334
APPENDIX M – The Ramp Signal Controller in VISSIM .....	338
APPENDIX N – Number of Particles Needed for Prediction Mode .....	352
APPENDIX O – On-Line Performance of the Freeway Traffic Model .....	354
APPENDIX P – Development & Application of the Ottawa Model.....	365

# Abbreviations

CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
DBDN	Dynamic Bayesian Decision Network
DBN	Dynamic Bayesian Network
DIC	Deviance Information Criterion
DOT	Department of Transportation
EU	Expected Utility
FTM	Freeway Traffic Model
GIS	Geographic Information System
HCM	Highway Capacity Manual
ITS	Intelligent Transportation System
LOS	Level of Service
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
MOE	Measure of Effectiveness
OD	Origin-Destination
RMSE	Root Mean Squared Error
TDM	Transportation Demand Management
VAP	Vehicle Actuated Programming (used in VISSIM)
VMS	Variable Message Sign
vph	Vehicles per hour
vphpl	Vehicles per hour per lane

# 1 INTRODUCTION

In many urban centres, traffic congestion is paralyzing the transportation network, impacting economic vitality and the quality of life enjoyed by residents. According to Transport Canada (2006), recurrent congestion in urban areas costs Canadians between \$2.3 billion and \$3.7 billion per year (in 2002 dollars) – a cost attributed to increased travel time, fuel consumption, and greenhouse gas emissions. While infrastructure projects are often implemented to increase road capacity and improve traffic flow, such projects can have significant environmental and social impacts. Moreover, the cost of road expansion is prohibitive. In today's fiscal environment, funding constraints limit the scope of what can realistically be accomplished.

As travel demand continues to increase, traffic congestion will only become more severe unless solutions can be found – solutions that create healthy, sustainable communities. To effectively meet the travel needs of urban residents, a balanced approach is required, in which efforts to manage travel demand are complemented with strategic road network improvements to address capacity deficiencies. There is also a need for innovative measures at the traffic operations level to improve the safety and efficiency of existing transportation infrastructure. Such measures not only have the potential to enhance mobility, but can also achieve important socio-economic objectives, including a reduction in vehicle emissions.

Of the various advanced traffic control systems under development, ramp metering in particular has proven to be an effective strategy for improving the operational performance of urban freeways – a strategy that can be implemented at lower cost and with less environmental impact than more traditional road-building initiatives. By carefully controlling the demand allowed to enter the freeway, a reduction in system-wide travel time can be achieved, despite the introduction of ramp delays.

While many ramp metering strategies have been developed, opportunities for improvement continue to exist, particularly with regards to equity. Ramp meters are often perceived to be unfair since some drivers may experience considerable ramp delays, while other access the freeway unimpeded. Such perceptions can influence public

support. Without support, a control strategy, no matter how effective, cannot be considered successful.

This thesis describes a new ramp control algorithm which balances efficiency and equity objectives, building on previous research to improve traffic operations in urban areas. To capture the uncertainty inherent in freeway systems, the algorithm has been implemented using a dynamic Bayesian decision network which incorporates a probabilistic model of freeway flow. In developing the algorithm, various techniques for real-time inference using freeway sensor data were investigated. Based on this review, a particle filter was selected for updating probabilities within the Bayesian network, allowing the algorithm to track the current freeway state and estimate its future performance under different ramp metering scenarios. The relative merit of each scenario is determined using a utility function which considers both efficiency and equity. While different options are available for solving the control problem, a pattern search technique was found to give reasonable results within the real-time constraints of the problem.

The algorithm has been implemented in MATLAB and integrated into a VISSIM micro-simulation model for assessing operational performance under realistic conditions of traffic flow. The following sections describe the new algorithm, provide a summary of key findings and present recommendations for future work. In particular, Section 2 provides background information on ramp metering, while Section 3 describes the limitations of current algorithms and highlights opportunities for improvement. A brief overview of dynamic Bayesian decision networks is provided in Section 4, followed by research objectives in Section 5 and the research methodology in Section 6. The new ramp control algorithm is introduced in Section 7. Section 8 describes the off-line testing that was carried out to confirm the reliability of the Freeway Traffic Model which forms the foundation of the algorithm, while Section 9 presents the results of on-line tests conducted using a conceptual freeway simulated in VISSIM. Section 10 describes the performance of the algorithm in a more realistic model of Highway 417 in Ottawa. Finally, conclusions and recommendations can be found in Sections 11 and 12 respectively.

## 2 BACKGROUND

### 2.1 Ramp Metering Defined

One of the major challenges facing urban areas is traffic congestion. Given the cost of building new roads, increasing emphasis is being placed on using existing infrastructure more efficiently, by applying new technologies and control measures to better manage the flow of traffic. One control strategy that has received considerable attention is freeway ramp metering.

In simplest terms, ramp metering involves the use of traffic signals on freeway entrance ramps to control vehicle access to the freeway system. Vehicles are released onto the freeway in groups of one or two in accordance with the specified metering rate. Timing plans may be fixed or traffic-responsive, implemented in isolation or coordinated over an entire system.

Ramp meters are generally installed to accomplish one of two primary functions.

The first is to reduce turbulence in on-ramp merge areas. A platoon of vehicles entering the freeway (often caused by an upstream traffic signal) can cause turbulence in the merge area as vehicles on the freeway change lanes or slow down to accommodate the merging traffic. Such localized congestion can be minimized through the use of ramp meters; by releasing vehicles in a controlled manner, ramp platoons are broken up, reducing turbulence. Of course, with less turbulence, the potential for collisions is also reduced, leading to a reduction in non-recurrent congestion. In more sophisticated implementations, the control algorithm may use sensor data to predict the occurrence of gaps in mainline flow, allowing vehicles to be released based on the freeway's ability to accept merging traffic. Metering systems whose primary function is to reduce turbulence in on-ramp merge areas produce minimal ramp delays, since the metering rate is set equal to the ramp demand.

The second, more common, use of ramp meters is to address congestion bottlenecks. In such applications, the metering system is designed to control on-ramp flow such that mainline capacity is not exceeded, or if flow breakdown does occur, the magnitude and

duration of congestion are reduced. By limiting vehicle access to the freeway at peak times through the creation of on-ramp queues, ramp meters effectively distribute the entrance demand over a longer time period for which excess capacity is available. The result? A system-wide reduction in travel time, despite the introduction of ramp delays. Ramp metering achieves this savings via four main mechanisms (Banks 2000). With reduced congestion, vehicles destined to exits upstream of the bottleneck are no longer delayed by mainline queues. At the same time, by preventing traffic breakdown, flows through the bottleneck can be sustained at higher levels (i.e. the free-flow capacity, rather than the somewhat lower congested capacity that occurs when flow breaks down). Ramp delays also encourage traffic diversion (to alternative routes, modes, and travel times), reducing freeway demand during periods of peak congestion. Finally, by minimizing freeway back-ups, ramp meters have the potential to reduce collisions and related traffic impacts.

Since their first introduction in the late 1950's and early 1960's, ramp meters have been deployed throughout North America. Based on 2007 data, it is estimated that there are nearly 4,200 ramp meters operating in some 20 metropolitan areas within the United States (U.S. DOT 2008). Cities with major ramp metering programs include Los Angeles, Minneapolis, San Diego, San Francisco, Seattle, Phoenix, Milwaukee, Portland, and Houston. In Canada, ramp metering is much less common (Abdulhai and Kattan 2004), however, ramp meters have been installed on a limited basis in Toronto, Montreal, Quebec City, and Vancouver.<sup>1</sup>

Currently, most ramp metering applications are designed to reduce the number of vehicles entering the freeway during peak periods to minimize mainline congestion (U.S. DOT 2006b). Nonetheless, simply by virtue of breaking up vehicle platoons, such systems may also benefit from reduced turbulence in the merge area, including a reduction in merge-related collisions.

---

<sup>1</sup> Based on information from: U.S. DOT 1995; Abdulhai and Kattan 2004; Transports Québec 2007; MTO 2010; ITS Canada, no date; and Delcan et al. 2006.

Appendix A provides an introduction to the design of ramp metering systems, including a discussion of key design parameters. Additional information on the mechanisms by which ramp metering improves freeway performance can be found in Appendix B.

## **2.2 Benefits of Ramp Metering**

Both simulation and field studies have demonstrated the many benefits of ramp metering: reduced freeway congestion, higher travel speeds, and lower system-wide travel time, even with ramp delays (see for example Papageorgiou et al. 1997; Cambridge Systematics 2001; U.S. DOT 1995, 2006b; Jacob and Abdulhai 2005; O'Brien 2000; Papamichail et al. 2010; Zhang et al. 2001). Environmental benefits attributed to ramp metering are also reported in the literature, including a reduction in fuel consumption and vehicle emissions. In terms of safety benefits, ramp meters have been shown to reduce freeway collisions, particularly rear-end and side-swipe collisions associated with stop-and-go traffic and on-ramp merging. With fewer collisions, non-recurrent congestion is also reduced, further improving traffic flow.

Table 2-1 illustrates some typical results from actual field installations. Although the data is somewhat dated, it provides a good indication of the types of benefits that can be achieved through the introduction of ramp metering. Note that the statistics provided in Table 2-1 are not necessarily comparable due to differences in measurement techniques between different jurisdictions (in particular, it is not known whether the operational benefits refer strictly to the freeway mainline, or also account for ramp delays). In general, the actual results obtained for a given corridor are likely to be highly site-specific, depending on geometric characteristics, traffic patterns, and driver behaviour.

**Table 2-1 Typical Ramp Metering Results**

<b>Ramp Metering Installation</b>	<b>Observed Operational Benefits</b>	<b>Observed Safety Benefits</b>
Portland, OR	173% increase in average travel speed	43% reduction in peak period collisions
Minneapolis, MN	16% increase in average peak hour travel speed and 25% increase in peak period volume	24% reduction in peak period collisions
Seattle, WA	52% reduction in average travel time and 74% increase in traffic volume	39% reduction in the collision rate
Denver, CO	57% increase in average peak period travel speed and 37% decrease in average travel time	50% reduction in rear-end and side-swipe collisions
Detroit, MI	8% increase in average travel speed and 14% increase in traffic volume	50% reduction in total collisions and 71% reduction in injury collisions
Long Island, NY	9% increase in average travel speed	15% reduction in the collision rate

Source: U.S. DOT (2006b), which cites U.S. DOT (1995) as the original source of the data

### 2.3 Negative Impacts

While ramp meters have proven to be an effective traffic management tool, a number of negative impacts also exist which must be carefully planned for and mitigated as part of the implementation and operation of any ramp metering system. The main issues can be summarized as follows:

- **Impacts to adjacent roads** – If ramp storage space is insufficient, ramp queues may spill back onto adjacent roads, blocking upstream signals and negatively impacting traffic operations on these facilities. To prevent queue spillback, many ramp metering systems include a mechanism to over-ride the metering rate if excessive queue lengths are detected. More sophisticated approaches include the development of ramp metering algorithms which incorporate queue constraints directly (Kotsialos and Papageorgiou 2004), or the use of specialized control routines for regulating ramp queues more precisely (Gordon 1996; Sun and Horowitz 2006).
- **Induced demand** – Any measure which improves the operational performance of the freeway system has the potential to generate induced demand. Such demand

may offset some of the benefits of the operational improvements, causing the system to function less effectively than intended.

- **Preferential Treatment of Long-Distance Trips** – As demonstrated by Levinson (2002), ramp metering tends to benefit longer trips, while shorter (local) trips are often disadvantaged. Likewise, Yin et al. (2004) found that ramp metering generally favours longer trips, although the results are highly site-specific. Not only does such preferential treatment for certain trip types have important equity implications, but by providing an incentive for longer trips, ramp metering risks triggering additional demand for long-distance freeway travel.
- **Urban sprawl** – Since ramp metering tends to favour longer trips, there is a concern that such systems may contribute to urban sprawl. At the periphery of the urban area, ramp meters are often unnecessary; it is not until closer to the urban centre that traffic volumes warrant ramp control. Under such a system, residents who access the freeway outside the metered zone benefit from low mainline congestion, without experiencing the ramp delays encountered elsewhere in the city. As a result, the accessibility of outlying communities is enhanced, encouraging urban sprawl. To the extent that such development generates additional freeway travel, the performance of the system may be compromised, triggering the need for further operational improvements.
- **Inequitable distribution of delays, leading to public opposition** – Ramp metering works by imposing delays at freeway entrance ramps. For some drivers, the delay is relatively short, for others, the delay may be much longer, depending on the ramp used to access the freeway and the time the trip commences. The equity imbalance becomes even more severe if only certain ramps are metered, forcing some drivers to wait, while others proceed unhindered, receiving all the benefits of reduced congestion without the cost of ramp delay. Such inequities foster public opposition to ramp metering, and may partially explain why ramp metering has not been deployed more widely as a control measure.

#### 2.4 Role of Traffic Diversion

In addition to the impacts noted above, ramp metering also has the potential to induce traffic diversion. As ramp delays increase, some drivers may be encouraged to use an alternative route to reach their destination. Whether such diversion is considered a positive or negative impact depends on the situation.

Some researchers argue that traffic diversion is one of the key benefits of ramp metering, since it promotes more efficient use of (presumably under-utilized) parallel routes. For example, Taylor et al. (1998) claim that “diverting traffic to alternative routes is the way in which ramp metering is typically most effective” (pg. 18), while the U.S. Department

of Transportation (1995) cites research claiming that “attractive and efficient alternative routes can be a key factor in the effectiveness of a ramp metering system” (pg. 20). In describing the operation of freeway-to-freeway ramp meters in Minnesota, Jacobson and Landsman (1994) claim that “MnDOT’s primary objective often has been to encourage route diversion away from onerous merges, especially where alternative routes have been identified” (pg. 49). Similarly, in their description of ramp metering benefits, Hellinga and Van Aerde (1995) note that ramp control can be used “to encourage spatial, temporal, and modal diversions to other roads, times, and modes having lower marginal systems costs” (pg. 75).

In contrast, a number of people cite traffic diversion as one of the negative impacts of ramp metering (Levinson et al. 2006; Wu et al. In Press; U.S. DOT 2006b). This viewpoint stems from concerns of increased traffic congestion and collisions on roads adjacent to the freeway which serve as diversion routes. Depending on the number and distribution of ramps subject to metering, increased cut-through traffic (particularly in residential areas) may also be a concern as drivers seek to avoid ramp delays.

Clearly, whether traffic diversion is viewed positively or negatively depends on the circumstances. In general, traffic diversion from the freeway is:

- Positive if parallel routes have excess capacity and can be used to route traffic around bottlenecks/incidents
- Negative if it causes unwanted cut-through traffic or generates congestion on alternative routes

This mindset is reflected in the following extract from the U.S. Department of Transportation’s Ramp Metering Status Report: “A well designed and operated ramp metering system improves operations and does not cause excessive diversion to adjacent streets” (U.S. DOT 1995, pg. 21). The key word is excessive – diversion is acceptable, even desirable if it can be accommodated by the adjacent road network; however, once diversion becomes excessive, mitigation measures are warranted.

While the above discussion focuses on traffic diversion from the freeway due to ramp delays, Papageorgiou et al. (1997) allude to a different type of diversion also associated

with ramp metering, in particular, that ramp metering is increasingly seen as a “means to control unwanted deviations of traffic to urban parts of the network” (pg. 97). In other words, by improving freeway operations through the use of ramp metering, drivers who may have previously diverted from the freeway due to mainline congestion are encouraged to divert back, thus improving operations on parallel routes. This behaviour was observed by Papageorgiou et al. (1997) in the case of the Boulevard Périphérique in Paris. Likewise, an evaluation of San Jose’s ramp metering system showed that, even with ramp delays, some drivers were attracted to the freeway due to the decrease in mainline congestion and associated reduction in travel time (U.S. DOT 1995).

The *Ramp Management and Control Handbook* published by the U.S. Department of Transportation (2006b) also highlights the benefits of ramp metering in terms of diverting traffic from unwanted areas:

*... ramp metering may be used to reduce traffic that cuts through neighborhoods or sensitive areas. If traffic is avoiding freeway congestion by driving through these areas to access a downstream ramp, the downstream ramp can be metered. If this ramp feeds the bottleneck that causes the freeway congestion, the problem can be attacked on two fronts. First, ramp metering can improve the flow on the mainline, thereby reducing the need for traffic to cut through the neighborhood or sensitive area. Second, the ramp meter will add a delay to the cut-through trip, again reducing the incentive to cut through the area of concern. In this case, ramp delays may not be a major concern. If the ramp traffic during the metered time is primarily traffic diverting from upstream ramps and the ramp has enough storage, long delays may be advantageous in meeting the objective.*  
(Section 5.3.1)

Diversion of traffic to the freeway in many ways can be considered induced demand. Such diversion may mitigate the benefits of ramp metering to some extent if the diverted traffic is destined through the bottleneck. However, if much of the diverted traffic is destined to exits upstream of the bottleneck, the impacts to freeway operation will be negligible.

In the literature, there is some disagreement over the importance and extent of traffic diversion associated with ramp metering. Evaluation of the INFORM system in Long Island, New York found that 15% of drivers frequently use another road to avoid waiting at a ramp meter, while another 27% do so occasionally (Smith and Perez 1992). The U.S. Department of Transportation’s 1995 Ramp Metering Status Report cites case studies from Portland, Los Angeles, Denver, Seattle, and Detroit. In each of these ramp metering

installations, no significant diversion of traffic from the freeway to adjacent surface streets was observed. According to the report, factors that influence the extent of diversion include: trip length, queue length, entry delay, and the availability (and attractiveness) of alternative routes (U.S. DOT 1995). In a user preference survey conducted in Portland, Oregon, Alkadri (1998) found that only 13% of freeway trips are short enough to be realistically targeted for diversion, suggesting an upper limit on the potential magnitude of freeway diversion within the Portland area.

The above discussion has focused on ramp metering and its ability to trigger traffic diversion through ramp delays (diversion from the freeway) or reduced freeway congestion (diversion to the freeway). In some cases, traffic diversion is considered a control strategy to be integrated with ramp metering. In this context, diversion refers to drivers already using the freeway who divert to an alternative route to avoid a bottleneck or incident. To be effective, this type of diversion requires the use of traveler information systems to communicate relevant information to drivers, such as the presence of downstream congestion or the travel time on alternative routes. However, while it may be possible to use such information to encourage diversion, it is difficult to control who responds. In general, the response rate depends on the quality of the information provided (and drivers' confidence in the information), the availability of alternative routes, and driver familiarity with the road network.

The provision of traveler information can also play an important role in diverting traffic destined to the freeway, for example, by communicating information on current ramp wait times at key decision points (i.e. intersections with alternative routes as drivers approach the freeway). This is particularly important if the ramp metering rate varies significantly (i.e. as part of a traffic responsive system), and is difficult for drivers to predict in advance. The more diversion is relied upon to achieve ramp metering objectives, the more important it is to provide travelers with real-time information.

Section 3.2.2 provides additional information on traffic diversion and its potential role in freeway control applications.

## 2.5 Public Perceptions

As noted in Section 2.1, there are nearly 4,200 ramp meters operating throughout the United States. However, despite their wide deployment, ramp meters remain controversial, limiting their use in freeway control applications. While ramp metering systems may produce many benefits from an efficiency perspective, they are often perceived to be inequitable. Some drivers are delayed at entrance ramps, others access the freeway with little or no restriction. Complicating the situation is the fact that people perceive different types of delay differently, placing more (or less) value on ramp delay versus delay incurred on the freeway mainline (Levinson et al. 2006). Some drivers are willing to wait at a ramp meter in return for improved freeway conditions; others prefer higher mainline congestion, even if it means a longer trip overall.

As noted by the U.S. Department of Transportation in their 1995 Ramp Metering Status Report, “to the public, ramp meters are often seen as a constraint on a roadway normally associated with a high degree of freedom” (pg. 21). Given this view, what prompts public acceptance of ramp metering? Levinson et al. (2006) hypothesize that there are two main rationales for accepting ramp control:

- Delayed gratification (drivers must wait at the ramp meter in order to benefit when travelling on the freeway)
- Social dilemma (drivers experience some delay at the ramp so that the collective delay is lower)

For many people, unrestricted freeway access is viewed as a right, and it is difficult to convince drivers that the introduction of ramp control will lead to an overall benefit in freeway performance, even though some drivers may be worse off. For this reason, public education campaigns are viewed as crucial to the successful deployment of a ramp metering system (U.S. DOT 2006b, U.S. DOT 1995). Interestingly, traffic signals may also be viewed as a restriction to roadway travel, but over time, have become almost universally accepted.

There is some evidence that drivers familiar with ramp metering do recognize the benefits of ramp control, and may be willing to tolerate some extent of ramp delay in return for improved mainline operations. In some cities, such as Los Angeles, additional

ramp meters have been requested by the public (U.S. DOT 1995). A user preference survey of freeway drivers in Portland, Oregon found that most respondents accept ramp metering as a beneficial traffic control technique, with the majority indicating that they would accept some form of freeway control if it improved the quality of their commute and reduced their overall travel time (Alkadri 1998). Only 2% would not tolerate ramp metering delays of any length, 87% would accept delays of up to five minutes, while 11% would tolerate delays exceeding five minutes.

As part of the evaluation of the INFORM traffic management system in Long Island, New York, a survey of households in the INFORM corridor was carried out to gauge public perceptions of the system. It was found that approximately 40% of residents considered ramp metering to be a good idea, while another 40% viewed it to be a poor idea. The remaining 20% had no opinion (Smith and Perez 1992).

While the reasons for driver dissatisfaction with ramp metering are varied, ramp delay is perhaps the greatest source of discontent, particularly if the delay is believed to be excessive, or disproportionate to the perceived benefits. In Minneapolis-St. Paul, public complaints prompted the Minnesota Department of Transportation to temporarily shut down the ramp metering system to assess its effectiveness (Cambridge Systematics 2001). As part of the study, traveler surveys and focus groups were conducted to assess drivers' perceptions of the ramp metering system before and after the shut-down experiment. Respondents generally believed that overall traffic conditions had deteriorated with the shut-down of the ramp meters. However, despite this increased appreciation for ramp metering, many respondents in both the pre- and post- shut-down survey advocated modifications to the system to address concerns related to ramp wait times, the period of ramp meter operation, and the number of ramps subject to control. In particular, there was a desire to reduce ramp wait times, shorten the hours of ramp meter operation, and limit the number of meters to areas of high traffic congestion. Interestingly, only 20% of respondents before and after the shut-down supported a complete removal of the system.

It seems clear that support for ramp metering does exist, however, the importance of properly designing the system to reflect public concerns cannot be overstated. Although

measures to limit ramp delays below a certain threshold or similar actions to address complaints may impact the effectiveness of the system from a strictly operational perspective, trade-offs are often necessary to achieve public support; a system that does not meet the needs of the public cannot be considered successful.

## **2.6 Ramp Metering Algorithms**

Ramp metering algorithms can be characterized in a number of different ways. Most commonly, algorithms are classified according to the type of ramp metering operation:

- Fixed time vs. traffic responsive
- Local vs. coordinated

Algorithms can also be classified based on the specific objectives and assumptions incorporated into the control logic:

- Is mainline congestion permitted?
- How are ramp queue/delay constraints addressed?
- Does traffic diversion occur and if so, is diversion to be optimized as part of the control problem?

Some algorithms are predictive, and are able to prevent congestion from occurring; others operate in a strictly responsive mode, reacting to congestion as it arises.

Solutions to the ramp metering problem have encompassed a diverse array of control and optimization methods, including: fuzzy logic, neural networks, machine learning, linear programming techniques, and more. Since most ramp metering algorithms seek to minimize congestion and delay, the most common objective functions involve minimizing total travel time (including ramp delays) or maximizing system outputs.

Under the framework proposed by Zhang and Levinson (2004a), ramp metering strategies can be classified according to the control method used in the algorithm and the threshold values adopted by the agency:

- ***Control method***
  - Based on flow or density? While early strategies focused on traffic flow, there is an increasing trend towards controlling density, which may be a better indicator of flow breakdown
  - Feedback vs. feed-forward? Under feedback control, metering rates are adjusted based on detected differences between the observed flow/density and the corresponding threshold value. Under feed-forward control, potential discrepancies are estimated in advance, and action is taken accordingly
  - Linear vs. non-linear? A number of methods are available to determine the appropriate control parameters once a deviation from the critical threshold has been detected. While some non-linear controllers (i.e. neural network controllers, fuzzy rules) have been shown to outperform their linear counterparts, linear controllers are often simpler and less costly to implement
- ***Threshold values***
  - What is the capacity/critical occupancy of each freeway section? Since flow breakdown is probabilistic, the thresholds used in the algorithm will depend on whether the local freeway authority is risk-adverse or risk-seeking

By de-composing the algorithm into distinct components, researchers can examine the impact of different threshold values or control methods, facilitating the analysis and comparison of ramp metering strategies. Such analysis and comparison is essential:

*Simulation studies can show whether one strategy outperforms another, but do not shed light on how and why that is the case ... If we do not pursue answers to the how and why questions, successful simulation, even field evaluation results, do not necessarily imply that the underlying theory is superior (Zhang and Levinson 2004a, pg. 871).*

While the number of algorithms reported in the literature is extensive, only some have been implemented in the field. In their review of ramp metering algorithms, Zhang and Levinson (2004a) distinguish between theoretical approaches and practical approaches.

Theoretical approaches may give “optimal” results but have not seen widespread use, having proven difficult to implement in practice. Such approaches typically require a detailed model of the system to predict how traffic patterns will evolve over time in response to the control action. Some theoretical approaches suffer from computational limitations which makes them difficult to implement in real-time. Others rely on data that is not readily available (such as dynamic OD patterns). Still others may incorporate

unrealistic assumptions, for example, diversion rates based on system-optimal considerations that do not reflect individual decision-making.

In contrast, practical approaches can be readily implemented in the field using available data. Algorithms in this category are often based on simplified heuristic approaches that may not yield optimal results. However, while such algorithms may lack the sophistication of the more theoretical approaches, they have proven their effectiveness in real-world applications. Examples include:

- Minnesota Zone Algorithm (Xin et al. 2004)
- Washington State DOT Fuzzy Logic Algorithm (Taylor and Jacobson 1998)
- Denver, Colorado Helper Algorithm (U.S. DOT 2006b)
- System-Wide Area Ramp Metering (SWARM) Algorithm (U.S. DOT 2006b)
- ALINEA Algorithm (Papageorgiou et al. 1991)

In terms of effectiveness, it has been found that the most efficient coordinated ramp metering algorithms are those that meter as few ramps as possible, with metering restricted to those ramps immediately upstream of the bottleneck. In their evaluation of the Advanced Motorway Optimal Control (AMOC) strategy, Kotsialos and Papageorgiou (2004) found that the most efficient scenario is the one with no ramp queue constraints, since the algorithm is able to limit the ramps under control to those closest to the bottleneck. With ramp queue constraints, more ramps must be added to the metering scheme, decreasing efficiency. In this latter scenario, mainline congestion was not completely eliminated, since it was found to be more efficient to have some mainline congestion than meter additional ramps where a high proportion of drivers are destined to exits upstream of the bottleneck location.

Likewise, Banks (2000) shows that the most efficient ramp metering strategy is to initiate metering at the ramps closest to the critical section. In cases where the metering scheme must be extended beyond the first few ramps, overall system delay tends to increase. The control logic developed by Zhang and Levinson (2004a) employs a similar philosophy; only those on-ramps nearest to the potential bottleneck are metered to ensure that flows remain strictly below capacity, eliminating mainline queues. They conclude that this feature of the control logic “explains why some local metering algorithms, and

coordinated algorithms that specifically target bottlenecks are successful – they are really close to the most-efficient metering logic” (pg. 883).

## **2.7 Limitations of Existing Methodologies**

The number of ramp metering algorithms reported in the literature is impressive. However, despite our current level of knowledge, opportunities for improvement continue to exist. From a review of recent literature, the key short-comings of existing ramp control strategies include:

- Inadequate consideration of equity
- Focus on freeway operations / lack of integration with the arterial network
- Failure to capture traffic diversion (both to and from the freeway)
- Failure to account for uncertainty

While some approaches have tried to address these issues in isolation, focusing for example on traffic diversion or network integration, more work remains to be done. In particular, there is a need to develop effective algorithms which address all of the identified short-comings under a single framework. Movement towards this objective may occur gradually through incremental improvements to existing algorithms, or in a sudden leap forward as new approaches are developed.

Section 3 provides a more detailed discussion of the limitations of existing ramp metering strategies, and outlines various opportunities for improving algorithm performance. As part of this assessment, results from the literature review are summarized, confirming the above issues and related research needs.

## **3 EXPLORATION OF RESEARCH NEEDS**

### **3.1 Introduction**

The primary goal of ramp metering is to minimize congestion and delay by controlling the amount of traffic entering the freeway system. A number of algorithms have been developed to improve the efficiency of ramp metering as a control strategy, however, several key issues have not been fully addressed, and there remains considerable scope for improvement.

This section describes three crucial areas where existing ramp metering strategies could be improved upon: the treatment of equity, traffic diversion, and integrated control. For each area, an overview of key issues is provided to set the context, along with a discussion of the current state of practice and associated deficiencies. However, rather than focus on problems, the deficiencies are presented as opportunities – opportunities that were ultimately used to define the research objectives described in Section 5.

### **3.2 Opportunities for Improvement**

#### **3.2.1 Equity**

Equity concerns are often cited as one of the main reasons for public opposition to ramp metering. Generally, the most efficient ramp metering strategy is the one that only meters those ramps closest to the bottleneck. Such a strategy is also likely to be the least equitable; the more ramps included in the metering scheme, the more equitable the distribution of ramp delay.

While equity considerations are clearly important, traditionally, the main focus has been efficiency. In his assessment of how different professions measure efficiency in transportation, Levinson (2003) criticizes the “the general focus on systematic efficiency” which ignores the “equity effects on individual welfare”, and recommends broadening the evaluation process to include consideration of equity (pg. 153). To date, little effort has been made to quantify the equity impacts associated with ramp metering, let alone study the underlying mechanisms which cause inequity, or develop improved ramp metering

strategies which address equity concerns (Yin et al. 2004). A review of the literature on equity and ramp metering certainly bears this out, as demonstrated later in this section (once the concept of equity has been more formally introduced).

Given the above discussion, there is considerable opportunity to improve ramp metering algorithms to include a more explicit accounting of equity. The following sections provide an introduction to the equity issue – how equity is defined and measured, how it can be incorporated into ramp metering algorithms, and future research directions.

### ***Equity Defined***

Most people have an intuitive sense of what is equitable and what is not. The concept of equity brings to mind notions of fairness and justice. In some cases, equity is associated with equality; an equitable health system implies that all individuals have equal access to high-quality treatment and care. However, equity is also subjective; one person's notion of what is 'fair' may not mirror that of another, due to different value systems and beliefs, or simply different perspectives. As Levinson (2002) notes, "the concept of equity is highly subjective and changes with the individual concerned" (pg. 179).

Equity issues in transportation tend to focus on how specific projects or policies will impact different groups of society, creating both winners and losers in terms of mobility, accessibility, environmental, and economic concerns (Levinson 2002). The *Online TDM Encyclopedia* produced by the Victoria Transport Policy Institute distinguishes between four general types of equity that are relevant to transportation. An excerpt is provided below.

### Definitions of Transportation Equity

Equity impacts can be difficult to evaluate, in part because the word "equity" has several meanings, each with different implications. There are four general types of equity related to transportation:

1. **Egalitarianism** - This refers to treating everybody the same, regardless of who they are. Egalitarianism implies that everybody should receive the same quality of services, pay the same price, and bear the same costs. In practice, this can be arbitrary and unfair, because it depends on how impacts are measured, and does not take into account differences in abilities and needs.
2. **Horizontal Equity** (also called "fairness") - This is concerned with the fairness of impact allocation between individuals and groups considered comparable in ability and need. Horizontal equity implies that consumers should "get what they pay for and pay for what they get," unless a subsidy is specifically justified.
3. **Vertical Equity With Regard to Income and Social Class** - This focuses on the allocation of costs between income and social classes. According to this definition, transport is most equitable if it provides the greatest benefit at the least cost to disadvantaged groups, therefore compensating for overall social inequity.
4. **Vertical Equity With Regard to Mobility Need and Ability** - This is a measure of how well an individual's transportation needs are met compared with others in their community. It assumes that everyone should enjoy at least a basic level of access, even if people with special needs require extra resources and subsidies.

Source: Adapted from VTPI (2011)

In the case of ramp metering, most discussions of equity are really dealing with equality (egalitarianism), the argument being that all drivers should enjoy a similar quality of service no matter where they live, what route they take, or when their trip commences. Horizontal equity is more concerned with the fairness of the distribution of impacts imposed on different categories of trip-makers. In this context, certain user groups may justify preferential treatment on the basis of their geographic location, trip length, or accessibility to the freeway system. While such fairness issues may have broader social implications, from a freeway operations perspective, all drivers are generally considered equal; even if origin-destination patterns are known (which is difficult to achieve under real-time conditions), it is difficult to justify favouring one group over another without a strong rationale for doing so. Moreover, from the perspective of public acceptance, concerns typically centre around equality of treatment; no driver wants to be delayed for significantly longer than any other driver using the freeway system.

There is also the challenge of addressing equity. When defined in terms of equality, equity is much easier to deal with from a system implementation and operations point of view. Given the range of drivers using any one ramp, it is much more difficult to provide preferential treatment to certain groups of users to satisfy notions of fairness. Where specific measures to address fairness are considered to be warranted, they are likely to be implemented at the policy level (i.e. by providing ramp meter by-pass lanes or deciding not to meter the ramp at all), not the operations level.

In the discussion which follows, any reference to equity can generally be taken to refer to the equality definition of equity (which, for the most part, is consistent with the approach taken by other researchers).

As alluded to above, equity has both temporal and spatial dimensions. In the context of ramp metering, temporal equity measures the difference in travel conditions among drivers who travel on the same route, but arrive at the entrance ramp at different times, while spatial equity measures the difference among drivers who arrive at different entrance ramps at the same time (Yin et al. 2004, citing work by Levinson et al.). According to the above definitions, perfect equity would imply that all drivers experience the same conditions whenever/wherever they access the freeway. This definition suggests that inequity among freeway users is not merely caused by ramp metering, but also other factors, such as accessibility – a dimension of social equity (Yin et al. 2004).

It has been suggested that the tendency to overlook equity in the development and implementation of ramp control strategies can be attributed, at least partially, to the difficulty in defining and measuring equity in a way that captures drivers' experiences and beliefs (Yin et al. 2004). Several definitions of what constitutes an equitable freeway system have been proposed, but none have seen widespread adoption. Some definitions are based on the assumption that the existing system (without ramp metering) is equitable; the installation of ramp meters causes some drivers to be worse off and others to be better off, introducing inequity to the system. While such an approach allows other complicating factors to be disregarded, it ignores existing inequities in the system, making it difficult to determine whether ramp metering has reduced inequity, or exacerbated it.

Other definitions take existing freeway congestion into account, and therefore imply that the existing system is not perfectly equitable. For example, Zhang and Levinson (2005) define an equitable transportation system as one with equal delay per distance travelled. Similarly, Kotsialos and Papageorgiou (2004) consider the freeway system to be equitable if vehicles entering the freeway from any ramp, at any point in time, have the same travel time, where travel time is defined as the ramp delay plus the time required to drive a fixed distance on the mainstream.

In practical terms, it may be sufficient to address equity issues by ensuring that all facilities operate at a pre-defined level of service or better. As long as this objective is satisfied, it can be argued that the system is equitable – no matter what route is selected or when, drivers are “guaranteed” a certain quality of service.

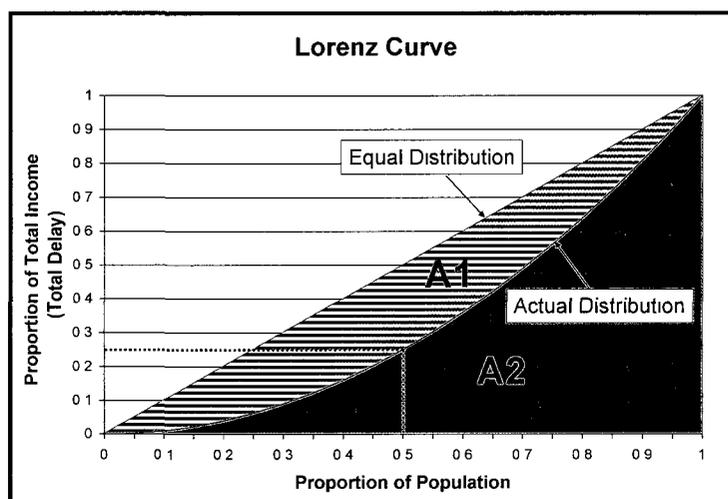
### ***Measuring Equity***

To assess equity impacts in a meaningful way, it is important to move beyond broad definitions of equity and develop quantitative measures which express the level of equity (or inequity) in a system as experienced by key stakeholders. In the case of ramp metering, equity is often measured using techniques for assessing income distribution. The most commonly used techniques are the Lorenz curve and the Gini co-efficient (Levinson 2002; Yin et al. 2004).

The Lorenz curve gives the proportion of the population which receives a given proportion of the total income (or in the case of ramp metering, total delay). For example, in Figure 3-1, the bottom 50% of the population receives only 25% of the total income (or delay). Note that the most desirable part of the curve from an individual perspective depends on what is shown on the y-axis – in the case of income, it is better to be on the right part of the graph with high income; in the case of delay, the best outcomes (i.e. lower delay) are found on the left part of the graph.

The Gini-coefficient can be calculated from the Lorenz curve by dividing area A1 by area A1 plus A2. A co-efficient of zero indicates perfect equality, while a co-efficient of one indicates perfect inequality (i.e. one driver gets all the delay). While the Gini co-efficient provides a useful measure of (in)equity, in practice, a state of perfect equity may be

difficult to achieve. As a result, rather than focus on the absolute value of the Gini coefficient, it is often more insightful to use the Gini co-efficient to examine how equity may change as a result of a given policy or action (Levinson 2002).



**Figure 3-1 Lorenz Curve & Gini-Coefficient**

Other measures that can be used to assess equity include the entropy statistic and the redundancy statistic (Levinson 2002). Kotsialos and Papageorgiou (2004) used the variance in travel time to measure the equity of a given ramp metering strategy, while Meng and Khoo (2010) define an equity index (described below). While some measures are applied at an origin-destination level, others focus strictly on ramp delays, ignoring other aspects of the trip.

Despite the wealth of literature on ramp metering, only a few researchers have tried to quantify how ramp metering impacts equity. In one notable example, Yin et al. (2004) examined the spatial distribution of the travel time savings associated with three ramp metering algorithms (ALINEA, BOTTLENECK, and ZONE) by plotting the ratio of the travel time savings for each OD pair (calculated as the travel time without metering divided by the travel time with metering). The ratio of travel time savings was also used to develop the Gini co-efficient for each algorithm as a measure of system-wide equity. To isolate the effects of ramp metering, the 'no control' case was assumed to be perfectly equitable.

From the analysis, it was found that up to 30% of OD pairs saw an increase in travel time with the introduction of ramp metering, with different strategies leading to different distributions of gains and losses. While ramp metering was generally found to favour longer trips, certain ramps near the middle of the study area suffered the greatest increase in travel time, suggesting that the results are site-specific. It was also found that the degree of inequity is impacted by the level of travel demand – the more congestion, the greater the inequity associated with ramp metering. The results of the analysis led the authors to conclude that:

- Different algorithms may lead to different levels of inequity
- The equity performance of a given algorithm is difficult to ascertain and highly site-specific due to the absence of a built-in mechanism for addressing equity – the only way to assess the relative performance of different algorithms is through simulation
- Although efficiency and equity are partially competitive objectives, in some cases, it may be possible to develop an efficient control strategy that does not seriously compromise equity

The findings by Yin et al. (2004) are generally consistent with Levinson's (2002) analysis of the Twin Cities ramp metering shut-off experiment. Similar to Yin et al., Levinson examined the distribution of travel delay over space, and concluded that the system becomes more equitable when ramp meters are removed. The results also confirm that ramp metering benefits longer trips, and that shorter trips are often disadvantaged.

### ***The Trade-off Between Equity and Efficiency***

As discussed in Section 2.6, the most efficient ramp metering strategy generally involves metering only the ramps closest to the freeway bottleneck. While this approach tends to produce the greatest reduction in network travel time, such a strategy is also inherently inequitable, since some drivers are subject to extensive ramp delays, while others access the freeway unimpeded, yet still benefit from reduced mainline congestion. In essence, as more ramps are added to the control strategy, more drivers are delayed who exit the freeway upstream of the congestion. Since these drivers do not contribute to mainline congestion, overall travel time increases, reducing efficiency. At the same time, delay is

spread over more ramps, so the impact to any one driver is less, creating a more equitable system.

Several studies have confirmed that the most efficient ramp metering scheme is also the least equitable one, implying that a trade-off must be made between efficiency and equity. For example, Zhang and Levinson (2004a) developed a new control logic which meters the ramps closest to the freeway bottleneck to eliminate mainline queues. The algorithm uses off-ramp exit percentages to determine ramp metering rates rather than relying on real-time origin-destination data. From a review of the algorithm, the authors conclude that the most efficient control strategy is also the least equitable one:

*It is interesting, though not very surprising, that the findings suggest the most efficient ramp control logic is also the least equitable one. To achieve efficiency goals, we must meter the least number of on-ramps in order to provide free-flow conditions for all commuters on the freeway mainline, a majority of whom access the freeway through other on-ramps with less restricted metering rates. With efficiency as the sole criterion, we may have done an engineering job very well. However, such a strategy is not politically palatable and may lack public acceptance. Minimized travel time for the system as a whole is a good thing. However, if that is achieved by helping some drivers at the expense of others, there is also a serious equity issue that should be considered. Future studies should pursue a mechanism balancing efficiency and equity of ramp meters (pg. 885-886).*

Kotsialos and Papageorgiou (2004) reached a similar conclusion in their analysis of the Advanced Motorway Optimal Control (AMOC) strategy for freeway network-wide ramp metering. AMOC is based on a validated macroscopic traffic flow model of the freeway network which explicitly captures the impact of mainline queues. Formulated as a dynamic optimal control problem, the algorithm attempts to minimize the total time spent in the system by all vehicles, including delays at ramp meters. The trade-off between efficiency and equity is addressed implicitly through consideration of available ramp storage space.

From the results obtained by applying the algorithm to the Amsterdam ring-road, it was found that the most efficient scenario is the one with no ramp queue constraints, since the algorithm is able to restrict metering to those ramps closest to the bottleneck. In contrast, the most equitable control scenario is the one with the least ramp storage space, since this constraint forces the algorithm to include more ramps in the metering scheme. Imposing maximum queue constraints can be seen as a way of distributing delay to the drivers

using the various on-ramps. The least equitable scenario (i.e. the scenario with the highest variance in travel time) is the one with no ramp control. This implies that the system is not equitable to begin with due to the spatial distribution of traffic congestion, and it is therefore inappropriate to use existing conditions as a benchmark for what is considered equitable; any evaluation of equity should include consideration of both ramp delay and mainline congestion.

While a trade-off between efficiency and equity may be inevitable, it is possible to develop ramp metering strategies that still meet efficiency objectives, but do so in a way that is at least somewhat equitable and fair. Yin et al. (2004) say it best: “While it is very difficult, if not impossible, to find a ramp metering strategy that reduces overall delay without causing inequality, it remains feasible to address inequity in ramp metering, by defining it within an appropriate context and incorporating the definition and measurement into the design, selection and implementation of ramp metering strategies” (pg. 497).

### ***Incorporating Equity into Ramp Metering Applications***

The success or failure of a ramp metering project in many ways depends on how well equity issues are addressed. Yin et al. (2004) outline a number of practical approaches that have been developed over time to mitigate the inequity inherent in ramp metering. However, the majority of such approaches address equity issues indirectly, and do not necessarily guarantee optimal results.

- Maximum queue constraints force ramp metering algorithms to include additional ramps in the metering scheme, resulting in a more equitable distribution of delay. Examples include Kotsialos and Papageorgiou’s (2004) network-wide ramp metering strategy and the Seattle bottleneck algorithm.
- Similarly, constraints on the maximum level of ramp delay ensure that all drivers enjoy a similar quality of service at freeway on-ramps wherever/whenever they access the highway, with no drivers subject to excessive waits. For example, the modified Minnesota algorithm limits the maximum ramp delay to less than four minutes.
- Restrictions on the minimum ramp metering rate also serve to limit ramp delays. To accommodate minimum metering rates, more ramps must be added to the metering scheme, which again serves to distribute ramp delays more evenly.

- There are examples where the metering system itself was constrained to address equity concerns. In Detroit, the initial ramp metering system was limited to the outbound direction only to alleviate concerns that suburban residents would be able to travel downtown with no ramp delays while freeway access for downtown residents was restricted. If the ultimate goal is to expand the system to both directions of travel once the benefits of ramp metering are established (as was the case with Detroit), such an approach may serve to aid in deployment, but does not address any equity concerns that may arise once the full system is operational.
- In the approach proposed by Atta-Armah (1994), the metering rate is adjusted to ensure that all arterial facilities operate at a pre-defined level of service or better. This approach is based on the assumption that ramp queue constraints are unnecessary, since drivers will begin to divert once the queue reaches a certain length. The system is considered equitable if all facilities impacted by traffic diversion operate at an acceptable level of service.

"A theoretical way to consider equity in ramp control has not been previously studied, but some practical equity considerations have evolved implicitly over time in real-world ramp control strategies ... A more systematic way to consider equity in ramp metering probably requires a change in the objective function itself."

Zhang and Levinson 2004a, pg. 885

While the above approaches certainly have value, their effectiveness is unclear. As noted by Yin et al. (2004): "all of these practical equity considerations can balance efficiency and equity to some extent. However, ... their impact is difficult to determine in advance and the compromising process is achieved implicitly" (pg. 498). With the exception of the studies referenced in this section, few researchers have attempted to examine how equity can be incorporated more directly into ramp control algorithms; efficiency is the driving motivation.

To improve the balance between efficiency and equity, Zhang and Levinson (2005) propose a new objective for ramp metering – minimizing the weighted travel time. This objective is based on the observation that drivers typically perceive ramp delays as more onerous than mainline congestion. If all travel time (free flow, congested, stop and go) is counted equally, there is no particular incentive to reduce ramp delays. However, by weighting travel time appropriately, the attitudes and expectations of drivers can be more adequately addressed, resulting in a more equitable system.

Rather than develop a new ramp metering algorithm, Zhang and Levinson (2005) modified an existing “efficiency-maximizing” algorithm to include equity considerations. Different versions of the algorithm were tested using simulation techniques, and the weighted travel time was calculated to determine the optimal strategy. To estimate the weighted travel time, a weighting function was developed, reflecting drivers’ perceptions, which weights longer delays more heavily.

The modified algorithm developed by Zhang and Levinson (2005) identifies active bottlenecks in real-time. When a bottleneck is detected, a predetermined number of on-ramps upstream of the bottleneck are metered to prevent mainline congestion (the number of ramps subject to control is a function of the equity coordination factor “ $X$ ”). Under the coordination scheme, the individual ramp delay at all on-ramps in the same coordination group is the same (achieved by equalizing the ratio of metering rates to ramp demand). Using simulation to test the various control strategies, it was found that equity is improved at the expense of efficiency. As more ramps are metered, the overall delay increases, however, the average delay per driver within the same coordination group declines.

Unlike the approach proposed by Zhang and Levinson (2005), the algorithm developed by Bellemans et al. (2006a) includes a constant weighting term which puts more or less emphasis on ramp delay. By choosing different weighting factors, the algorithm can be tuned to reflect public opinion regarding the trade-off between ramp queues and mainline congestion. However, since longer delays are not weighted more heavily, there is no guarantee that the resulting control solution will result in an equitable distribution of delay.

Yin et al. (2004) also examined ways to include equity in system-wide control strategies. They conclude that adding constraints (i.e. queue length restrictions) to existing strategies can be easily done, but may not necessarily improve system-wide equity. Incorporating equity measures such as the Gini co-efficient in the objective function is likewise not recommended, since doing so would require either solving a bi-objective problem, or combining measurements of efficiency and equity in one equation – both challenging tasks. Instead, the authors propose a new objective function which would maximize the

sum of the transformation of the travel time savings ratio for each origin-destination pair (calculated as the travel time without metering divided by the travel time with metering). The function includes an “inequity aversion” parameter which can be used to specify the relative importance of efficiency vs. equity in the optimization problem. An important property of this function is that, holding the average ratio constant (i.e. constant efficiency), the maximum value is obtained when all the ratios are equal.

More recently, Meng and Khoo (2010) developed a multi-objective optimization model for ramp control incorporating a modified cell transmission model of traffic flow. To account for equity as well as efficiency objectives, the authors define an equity index, which is calculated as the ratio of the minimum and maximum average ramp delay experienced at a group of on-ramps over a certain horizon. As this index goes to one, all ramps within the group experience a similar level of delay, resulting in a more equitable system. While the index as originally defined is concerned with spatial equity, the authors also demonstrate how to expand the index to address temporal equity as well.

To solve the multi-objective optimization problem, Meng and Khoo employ a hybrid non-sorting genetic algorithm, resulting in a set of Pareto-optimal ramp metering solutions. From this set, the system operator has the flexibility of choosing a preferred solution which offers an acceptable trade-off between efficiency and equity in accordance with local practice. Test results for a section of freeway in California suggest that the Pareto-solutions obtained from this approach are more equitable than the system optimal solution, but less equitable than the no-metering scenario. Likewise, the Pareto-solutions were found to be more efficient than the no-metering scenario, but less efficient than the system optimal solution. Such results underscore the trade-off that exists between efficiency and equity. According to Meng and Khoo, the proposed approach is most appropriate for off-line decision-making; whether the equity index could be used in real-time control remains unclear.

A final way to address the equity impacts associated with a particular ramp metering project is through compensation (Yin et al. 2004). Rather than adjust the ramp metering algorithm to ensure a more equitable distribution of delay at the expense of efficiency, it may be possible to compensate those individuals who are negatively impacted by the

introduction of ramp control. As an example, an agency may choose to improve transit service in certain areas to compensate for increased delay in accessing the freeway network. While such an improvement does little to benefit the freeway drivers impacted by ramp delays, it may address certain equity concerns at a regional level by ensuring that all areas of the city receive some benefit, either through the project directly or related compensation.

As noted by Levinson (2002), “solutions to equity problems include ideas such as bundling improvements, so that not only is there a net benefit (when all projects are considered together), but the number of winners exceeds the number of losers by a significant amount” (pg. 185). This implies that it is not only the distribution of benefits and losses that is important, but also the number of winners and losers. To gain public support, ramp metering algorithms must be adjusted to ensure no one is disadvantaged by a disproportionate amount. Equally important is the adoption of complementary projects, so that overall, a majority of residents are better off as a result of the proposed improvements.

### ***Equity – Key Issues***

When dealing with transportation and the movement of people and goods, there are a number of issues that may have a bearing on how equity is measured and addressed:

- One of the most common ways to measure equity is to examine the change in travel conditions between origin-destination pairs. When doing so, it is important to consider the number of trips represented by each OD pair. However, in practice, such information is often difficult to estimate without extensive data collection.
- In addition to creating ramp delays, ramp metering may also cause traffic diversion. Drivers who choose to divert are presumably worse off than in the case with no ramp metering, otherwise, they would have already diverted to an alternative route. As a result, any traffic diversion triggered by ramp metering may have equity implications that should be captured in the analysis. To assess the impacts of diversion, information is needed on the extent of diversion along each alternate route, as well as the corresponding travel time. For large networks with complex OD patterns, such information can be difficult to estimate.
- Typically, equity impact assessments are limited to the freeway network; implications for the arterial network are ignored. However, traffic diversion or

queue spillback onto adjacent roads can negatively affect the users of these facilities. Is it fair to penalize arterial roads in order to improve freeway operations? A comprehensive assessment of equity would consider the impacts to users of both facilities. However, such an endeavour brings an added level of complexity as the study area is expanded to include both freeway and arterial links.

- As discussed in Section 2.3, ramp metering tends to benefit longer trips at the expense of shorter ones. While this could be perceived as inequitable, it is also in many cases a deliberate design objective. According to Alkadri (1998), one of the roles of ramp metering is to “lower demand on the freeway by diverting some short-trip drivers to parallel surface streets that may have excess unused capacity” (pg. 75). Chaudhary and Messer (2002) also allude to this role: “A secondary objective of ramp metering is to introduce controlled delay (cost) to vehicles wishing to enter the freeway and, as a result, reduce the incentive to use the freeway for short trips during peak hour” (pg. 80). The U.S. Department of Transportation (1995) claims that “in concept, freeways are not intended to serve very short trips, and diverting some trips may even be desirable if there are alternate routes that are under-utilized” (pg. 20). It is unclear if a system designed specifically to target certain trip types can ever be considered equitable, at least without some form of compensation.
- In general, the more ramps included in the metering scheme, the more equitable the distribution of ramp delay. However, if a decision is made to meter ramps far from the bottleneck location, drivers using those ramps may not understand the rationale for the ramp controls, since the freeway immediately downstream from the ramps is uncongested. This in turn may lead to public opposition, despite a system which is in fact more equitable.
- Equity impacts will vary by direction of travel and time of day. Drivers commuting to work may experience low ramp delay during the morning trip, but high ramp delay during the reverse trip in the afternoon. Whether these delays are considered separately or as part of a combined trip could have a significant bearing on the equity performance of the system.
- There is evidence that drivers view freeway and ramp delays differently, which may influence perceptions of equity among the public. For example, Levinson et al. (2006) tested user preference for congested freeway travel vs. ramp delay using a computer administered stated preference survey (CASP) and virtual experience stated preference (VESP) survey involving a driving simulator. The results were statistically analyzed using a binary logit model to examine how different socio-economic, demographic, and personality traits impact travel preferences. In general, the CASP subjects showed a preference for freeway congestion over ramp delays, even if this meant overall longer travel times, while the VESP subjects showed a preference for ramp delays, accompanied by an improvement in freeway conditions. The authors hypothesize that the variation in results between the two groups can be attributed to differences in the test

methodologies. Although it is impossible to confirm which method is more correct based on the available data, the results show that people value different types of delay differently.

According to Levinson et al. (2006), the influence of personality on travel behaviour has not been adequately researched, although studies have shown that driving behaviour correlates with personality, and, in other contexts, personality scores have been correlated with willingness to wait. Clearly, more research is needed in this area to ensure that ramp metering strategies are consistent with driver expectations.

- To assess the impact of a particular project or action, equity must be measured relative to an appropriate baseline. However, in the case of ramp metering, the choice of a baseline can be problematic. In the early design stages of a project, it is often appropriate to measure equity relative to existing conditions. However, as discussed earlier, some researchers prefer to consider the existing situation as perfectly equitable in order to isolate the effects of ramp metering, while others feel it is more appropriate to include existing inequities in the analysis in order to gauge the true effect of ramp metering. Efforts to incorporate equity considerations directly into the ramp metering algorithm require a different baseline. As traffic patterns adjust to the new system and growth in demand occurs, it is difficult, if not impossible, to estimate how the system would evolve without ramp metering. In this situation, it is more meaningful to compare the equity of the system against the ideal situation, and optimize the system performance accordingly.

### ***Equity – Research Needs***

Based on the above discussion, a number of issues must be resolved if equity considerations are to be more fully addressed in the development and implementation of ramp metering strategies. Some of the key research questions that warrant further investigation are identified below:

- What is the best way to define/measure equity as demand patterns and travel behaviour change over time? How can the equity implications of any traffic diversion associated with ramp metering be accounted for?
- Should equity considerations be incorporated more directly into the metering strategy, and what is the best way of doing so?
- How should people's perceptions of travel time be addressed?

### 3.2.2 Diversion

Another limitation of existing ramp metering algorithms relates to traffic diversion. The majority of ramp metering algorithms do not consider traffic diversion. Of those that do, most focus on either:

- Diverting drivers off the freeway and around congested bottlenecks by providing information via variable message signs or other ITS-based systems
- Diverting drivers who would otherwise enter the freeway to a parallel arterial route by letting the ramp queue grow to its equilibrium length

In the case of the former, the control problem is expanded to include both ramp meters and dynamic traffic diversion in the optimization of freeway performance. To exploit the potential benefits of traffic diversion, algorithms have been developed which try to estimate system-optimal diversion rates (i.e. Jacob and Abdulhai 2005, Kotsialos et al. 2002). While such algorithms may provide useful information, there is no way to ensure this level of diversion happens in practice. Recognizing that actual diversion patterns may differ significantly from optimal, Wu and Chang (1999) compute a compliance rate based on observed exit ramp flows which is applied to the control solution. However, while this may improve accuracy to a certain extent, the algorithm continues to calculate an optimal diversion flow rate which may not be achieved in reality, since there is no guarantee that the same compliance rate will hold under different conditions.<sup>2</sup>

To overcome such limitations, Karimi et al. (2004) developed an algorithm which uses a logit model to predict how drivers will actually respond to the message displayed on a variable message sign. Control parameters include not only the ramp metering rates for the freeway corridor, but also the message displayed on each VMS. In essence, the algorithm computes the “optimal” route travel times to display on the VMS which will trigger an optimal level of traffic diversion, with the constraint that the “optimal” travel times equal the actual travel times as closely as possible. The objective function to be

---

<sup>2</sup> It is also unclear how the compliance rate is calculated by Wu and Chang (1999), since it is necessary to measure both the actual off-ramp flow, and the off-ramp flow that would occur in the absence of diversion. The latter cannot be measured directly, but must be estimated from knowledge of ‘typical’ OD patterns.

minimized consists of a weighted combination of the total time spent in the system, the prediction error, and the control variance.

The use of a logit model to predict driver behaviour is an improvement over other approaches, however, the algorithm is not well-suited to networks with complex origin-destination patterns or numerous alternative routes, due to the limited extent of information that can be displayed on a VMS. In-vehicle traveler information systems could certainly be used to provide more customized information, however, it is difficult to imagine a system which displays erroneous travel times in order to induce drivers into diverting a certain way to improve system effectiveness.

Logit models can also be used to predict traffic diversion at on-ramps as a function of the control parameters. To apply such models, real-time origin-destination data is needed, however, even where such data is available, its accuracy is often questionable.

Alternatively, in the approach adopted by Stephanedes and Kwon (1993), an incremental approach to route choice was adopted in which drivers make decisions at each potential diversion point without consideration of their ultimate destination. Although the approach was shown to be effective, its applicability to larger networks is questionable – certainly drivers may re-evaluate their route choice throughout the trip, however, such decisions invariably include consideration of where they ultimately want to be, and the conditions likely to be encountered along the way.

Other approaches used to model traffic diversion include time-series analysis based on previous observations, and equilibrium traffic assignment techniques. The former fails to capture the relationship between traffic control and route choice, limiting its utility for predicting traffic outcomes under various control scenarios. The latter requires time-varying origin-destination data that may be difficult to obtain, ignores short-term fluctuations in traffic flow, and incorporates potentially unrealistic assumptions regarding the information available to drivers and their ability to choose the shortest route (Stephanedes and Kwon 1993).

As the above discussion implies, there is significant opportunity to improve the prediction of traffic diversion associated with freeway control. By incorporating such approaches directly into ramp metering algorithms, a more realistic representation of

driver behaviour can be reflected in the control strategy, facilitating integration with arterial traffic management systems for improved network performance.

### 3.2.3 Integrated Operations

Interaction between the freeway and arterial network occurs primarily via two mechanisms:

- Traffic control
  - Controls at freeway on-ramps may causing queuing which spills back onto arterial roads
  - Controls on the arterial network may impact the demand entering/exiting the freeway (i.e. by creating vehicle platoons which affect on-ramp merging operations, by metering traffic destined to the freeway, or by causing off-ramp queues which spill back onto the freeway).
- Traffic diversion
  - Traffic diversion is often related to traffic control. Control measures influence the relative operating conditions on alternative routes, which in turn influences route choice behaviour. For example, drivers may divert from the arterial network in response to improved freeway performance due to ramp metering. Alternatively, drivers may divert to the arterial network to avoid ramp delays. As diversion occurs, changes in traffic control may be warranted to accommodate vehicles along the diversion route (or control the extent of diversion), further influencing diversion patterns.
  - Diversion may also occur in response to incidents. By providing appropriate information (i.e. via variable message signs, radio announcements, or in-vehicle information systems), drivers can be encouraged to divert around bottleneck locations, allowing system capacity to be used as efficiently as possible.

Because of this interaction, integration is essential to ensure that all transportation elements work together to enhance mobility. Integration aims at achieving greater system-wide efficiency, so that changes aimed at improving traffic operations in one part of the network do not jeopardize performance elsewhere.

Without traffic diversion, integration is relatively straight-forward to achieve. Changes in ramp or intersection signal timing can be introduced to achieve specific objectives.

Traffic diversion complicates the situation considerably, yet at the same time offers a significant opportunity for enhancing network efficiency.

In urban areas, drivers generally have a choice of travel routes, involving a variety of arterial and freeway facilities. While many travel decisions are made based on past experience, drivers also respond to real-time conditions, altering their travel choices en-route to avoid congestion and delay. Planning for such diversion is essential to minimize system impacts; through integration of arterial and freeway control, appropriate mitigation measures can be implemented. Of much greater value is the potential to influence routing decisions to achieve a more efficient distribution of traffic in response to real-time events. In such applications, the effectiveness of integrated control depends in large part on the ability to predict the traffic diversion resulting from control activities.

Intuitively, freeway-arterial integration implies coordination of traffic control measures to reduce system-wide congestion and delay. Chu et al. (2004) distinguish between advanced and integrated control techniques: “An advanced control algorithm, such as an adaptive ramp-metering algorithm, responds to traffic change dynamically through communication with field devices on a real-time basis. Integrated control combines and coordinates different ITS components” (pg. 78). In practice, there are different levels of integration possible, with different operational characteristics and associated network performance under recurrent and non-recurrent congestion. Possible integration strategies include:<sup>3</sup>

- **Integration based on arterial feedback** – Under this strategy, ramp metering rates are adjusted to reflect impacts to the arterial network, such as queue spillback or congestion caused by freeway traffic diversion.
- **Integration based on freeway feedback** – Under this strategy, signal timing on the arterial network is adjusted to accommodate traffic diversion triggered by ramp delays or freeway incidents. Signal timing may also be adjusted to reduce queuing on freeway off-ramps, or control on-ramp flow through metering at upstream signals.

---

<sup>3</sup> Note that the U.S. Department of Transportation’s *Coordinated Freeway and Arterial Operations Handbook* (2006a) uses a slightly different classification scheme for coordination of traffic signals within freeway-arterial systems: local coordination, area-wide integration, diversion strategies (for responding to incidents), and congestion strategies. These strategies generally fall within the broad categories described below.

- **Integration based on both freeway and arterial feedback** – Under this strategy, ramp metering rates and arterial signal timing are adjusted based on data from traffic sensors located on both the freeway and arterial network. For example, in the event of traffic diversion to the arterial network, signal timing at arterial road intersections could be modified to better accommodate the diverted traffic, while ramp metering rates could be adjusted to address any arterial congestion that materializes. Both systems operate autonomously, but include a mechanism to account for operating conditions on the other.
- **Concurrent optimization** – Under this strategy, ramp metering rates and arterial signal timing are determined concurrently as part of a fully integrated control algorithm.

The first three integration strategies are generally reactive, responding to traffic conditions as they arise. Integration is achieved through sharing of information, however, the arterial and freeway systems continue to operate independently. In more advanced applications, predictive capabilities may be added, allowing operational issues to be identified before they actually arise. However, unless ramp metering and intersection signal timing are optimized concurrently, there is no guarantee that system-wide optimality will be achieved. Indeed, as long as the two systems operate independently, the best that can be done is to optimize the performance of each system separately, and adjust the control parameters as necessary to accommodate constraints imposed by the other.

Only a few ramp control strategies have attempted to take the traffic conditions on parallel routes into account when determining the control parameters. Atta-Armah (1994) developed a pre-timed ramp control strategy in which “optimized” ramp metering rates were adjusted manually to satisfy arterial level of service constraints in a type of off-line integration. Chu et al. (2004) and Tian et al. (2002) applied simulation approaches to evaluate different traffic management strategies with varying levels of integration. In both cases, integration between the arterial and freeway network was achieved by adjusting the traffic signal timing on the arterial network to accommodate traffic diversion in the event of a freeway incident.

In a more comprehensive (but theoretical) treatment of the problem, Wu and Chang (1999) present an integrated control model and heuristic solution algorithm which concurrently optimizes off-ramp diversion flow rates, ramp metering rates, and

intersection signal timing on surface streets. In another example, Van Katwijk and Van Koningsbruggen (2002) explore the viability of using agent technology to coordinate traffic control instruments, including ramp metering and variable message signs. In terms of real-world applications, MacCarley et al. (2002), McLean et al. (1998), and Wang (2003) describe various efforts to implement integrated traffic control in Irvine, California, Glasgow, Scotland, and Beijing, China, respectively.

While the benefits of freeway-arterial integration are significant, more research is needed to design effective integration strategies and implement them successfully under real-world conditions. As noted by Wu and Chang (1999):

*... integrated real-time control for freeway corridor systems is one of the most promising strategies for developing advanced traffic management systems. However, studies on this subject are still in their infancy, and many critical issues regarding both modeling and solution strategies remain to be explored (pg. 14).*

Van Katwijk and Van Koningsbruggen (2002) concur, noting that “there is a need for more coordination of both the traffic control measures and the supporting information processing facilities” (pg. 456).

Integration of freeway-arterial control will ultimately require more than simply new control algorithms. Barriers to implementation include technology inter-operability issues, as well as institutional barriers related to the ability of different organizations to work together towards a common objective, taking a broad systems view rather than a narrow jurisdictional one. There is also a need to set clear and unambiguous policy goals (Van Katwijk and Van Koningsbruggen 2002). When the objectives of traffic management instruments conflict, prioritization is needed to determine what compromises may be necessary.

### **3.2.4 Summary**

The preceding sections have provided an overview of three key opportunities for improving ramp metering algorithms:

1. Enhancing algorithms to include a more systematic consideration of equity;
2. Implementing approaches to better capture traffic diversion; and

3. Devising new and innovative ways to integrate freeway and arterial traffic control.

For each opportunity, information was presented highlighting the current state of practice, specific challenges to be overcome, and major factors influencing future research directions.

The opportunities presented above provide the basis for the current research. Clearly, one project cannot address every limitation. The goal is to explore new ways of addressing equity, diversion, and system integration in the management of freeway traffic, advancing the state of knowledge in the pursuit of increasingly more effective control strategies.

After much deliberation, a decision was made to focus primarily on the equity aspects of ramp metering, while at the same time developing a control framework that is sufficiently flexible to address traffic diversion and integration objectives as potential future enhancements. This focus reflects both the importance of equity and the considerable scope for improvement that exists given the traditional emphasis on efficiency only. Without adequate consideration of equity, any efficiency gains may be irrelevant if the system is abandoned due to public opposition.

Based on the above discussion, a research statement was formulated. Section 5 presents the research statement and outlines guiding principles for the development of a new ramp metering algorithm.

## 4 BAYESIAN NETWORKS AS A FRAMEWORK FOR RAMP METERING

### 4.1 Overview

To successfully develop a new ramp control algorithm, a new analytical framework is needed which is capable of addressing the limitations of existing algorithms described in the previous section, particularly as related to the treatment of equity. One such framework that was explored was dynamic Bayesian decision networks. The following sections provide a brief introduction to Bayesian networks and their application in the transportation field.

### 4.2 Introduction to Dynamic Bayesian Decision Networks

A Bayesian network is a graphical model which captures the uncertainty inherent in real-life systems. Also known as probabilistic graphical models or belief networks, Bayesian networks can include both discrete and continuous variables. They can model systems of any size and complexity, incorporating both static and dynamic processes involving linear and non-linear relationships.

Within a Bayesian network, random variables are represented by nodes, while direct dependencies between variables are represented by arcs. The relationships between connected nodes are quantified through the use of conditional probability distributions: given a particular instantiation of the parent nodes, what is the probability of the child node taking on each of its possible values? If some of the relationships in the model are temporal, the network is referred to as a Dynamic Bayesian Network (DBN). Such networks allow us to model stochastic processes as the system changes over time.

Bayesian networks provide a framework for reasoning under uncertainty (Korb and Nicholson 2004). As new evidence is received, it is used to update beliefs about the system in a process known as probabilistic inference. More specifically, the evidence is used to calculate the posterior probability distribution for each node in the Bayesian network for applications involving predictive or diagnostic reasoning. The posterior

probability distribution is simply the probability after incorporating evidence; the prior probability distribution can thus be thought of as one's belief before evidence is received.

Since Bayesian networks are designed to model uncertainty, they are particularly well-suited for decision applications where impacts cannot be predicted precisely.

In decision theory, utility represents the level of satisfaction derived from the outcome of a particular course of action, taking risk and uncertainty into account. Where outcomes are uncertain, it is assumed that people will act to maximize their expected utility. Expected utility reflects the probability of each possible outcome,  $i$ , as well as its corresponding utility:

$$\text{Expected Utility} = \sum_i \text{Probability}_i \times \text{Utility}_i$$

Thus, for a given option, the expected utility is calculated by multiplying the probability of each potential outcome by the utility of the outcome, and summing the results. Outcomes may be defined based on a single attribute, or may incorporate multiple attributes which reflect different objectives. In Bayesian networks, a multi-attribute utility function can be represented as a utility node. The parents of a utility node represent the system attributes on which the utility is based. Given a probability distribution for these attributes, the expected utility is easily computed.

To be used in decision problems, a Bayesian network must also include action (or decision) nodes. Such nodes represent specific actions or policies that can be carried out on or by the system, impacting the system state and the associated utility.

A Bayesian network containing utility and action nodes is called a decision network or influence diagram. Using such networks, it is possible to compute the optimal action (or sequence of actions) which will maximize the expected utility. In the ramp metering control problem, decision nodes represent the ramp metering rates to be applied at different on-ramps, while utility nodes represent preferences for various system outcomes as related to the control objectives.

A more detailed overview of dynamic Bayesian decision networks can be found in Appendix C. A simple example of a Bayesian network is presented in Appendix D while Appendix E provides a discussion of inference techniques.

### **4.3 Bayesian Networks in Transportation**

While Bayesian networks are gaining widespread popularity in a diverse range of applications, only a few examples exist involving the use of Bayesian networks in transportation. Huang et al. (1994) applied Bayesian networks to analyze traffic scenes. In the BAT (Bayesian Automated Taxi) project, Forbes et al. (1995) explored the use of dynamic probabilistic networks to create autonomous vehicles which can interact with traffic.

Examples of Bayesian networks with potential application to freeway traffic management are rare. In one notable example, Kwon and Murphy (2000) developed a coupled Hidden Markov model (a special type of dynamic Bayesian network) to predict freeway congestion using sensor data. In the model, the observed speed is assumed to be a noisy representation of the unknown freeway state (either free-flow or congested). Model parameters were “learned” by applying various inference techniques to the test data. Although the predictive power of the model was found to be limited, several enhancements were planned to improve accuracy, including expansion of the freeway state variable to include additional modes of operation.

The probabilistic traffic model developed by Foo (2006) employs a more general Bayesian network representation to model speed, occupancy, and flow. The model assumes that the traffic state on a particular freeway segment depends on the previous traffic state on the segment, as well as the previous state on all facilities immediately upstream and downstream of the segment, including on-ramps, off-ramps, and express-collector transfers. Probabilities were developed by simple counting within the dataset. Similar to Kwon and Murphy (2000), the performance of the model was found to deteriorate significantly as the time horizon increases, limiting its utility for prediction.

The models developed by Kwon and Murphy (2000) and Foo (2006) were produced strictly from sensor data, and do not incorporate physical models of traffic flow. Since the

relationship between freeway control and freeway state is not considered, such models are not well-suited to control applications.

In contrast, several authors have applied Bayesian-based filtering techniques to improve the accuracy of traffic parameters (such as the critical density) for use in freeway control (Ozbay et al. 2006; Kosmatopoulos et al. 2006). There is also a growing body of research involving the use of filtering techniques for traffic state estimation, using real-time sensor data to update the traffic estimates from macroscopic models (Sun et al. 2003; Bellemans et al. 2006b; Wang and Papageorgiou 2005; Mihaylova et al. 2007; Wang et al. 2009; Hegyi et al. 2006; Yongjun et al. 2009; Ngoduy 2008; Sun and Horowitz 2006). While such efforts are usually not formulated in Bayesian network terms, the resulting stochastic traffic model can generally be considered a special form of dynamic Bayesian network. Initial research efforts have tended to focus on variants of the Kalman filter for updating the traffic state, however, the use of particle filters is also gaining attention. Unlike the standard Kalman filter, particle filters do not require linearization of the state and observation models, and are not constrained to Gaussian distributions for the measurement and process noise.

Despite the recent interest in filtering techniques for traffic state estimation, in many ways, the research is still in early stages. Of the examples cited above, many rely on artificial data, or involve limited freeway sections (with few or no ramps). Only Wang et al. (2009) provide real-data test results for an extended freeway section with multiple on- and off-ramps.

Although the development of stochastic models incorporating real-time information represents a significant step forward in terms of addressing uncertainty in freeway traffic behaviour, none of the models reported in the literature fully capture the probabilistic nature of flow breakdown. In addition, while most models were developed with control applications in mind, few have actually been applied to the ramp metering problem. In one noteworthy exception, Bellemans et al. (2006b) used an extended Kalman filter to track changes in the traffic system over time. This information was then fed into a predictive control algorithm to determine optimal ramp metering rates. While this approach certainly has merit, the control action is decided externally, and as a result, any

uncertainty which might influence the control solution is ignored. In another example, Sun and Horowitz (2006) developed a switching traffic responsive ramp-metering controller which employs a mixture Kalman filter to estimate the freeway state.

No examples could be found involving the use of dynamic Bayesian decision networks in ramp metering applications. Given the benefits of Bayesian networks and their ability to capture uncertainty, there is much value to be gained in exploring their potential application to the ramp metering problem. Unlike standard filtering approaches to traffic state estimation, the more general Bayesian network formulation works well for tracking *and* prediction, and also supports the development of control solutions which address competing objectives – a key requirement for the new algorithm.

## 5 RESEARCH OBJECTIVES

### 5.1 Research Statement

Over the years, ramp metering has been the subject of considerable research activity. While many advances have been made, not all deficiencies have been resolved, and there remain several opportunities for improvement. From the review provided in Section 3, one of the key deficiencies of current algorithms is the lack of consideration given to equity. Virtually all ramp metering algorithms have been developed to maximize efficiency, yet equity considerations can have a major impact on the success of the ramp metering initiative. If the system is perceived to be inequitable, the resulting public opposition may be difficult to overcome, regardless of the efficiency gains. Clearly, it is not enough to measure the success of an algorithm by how well it improves traffic flow. It must also meet the needs of system users, who may value ramp delays differently than delays due to freeway congestion, and who may be willing to trade-off some of the potential gains in operational performance for a system that operates more fairly.

To date, few attempts have been made to consider equity explicitly in the control problem; the approaches explored thus far have yielded important insights, but have failed to produce a definitive means for including equity in real-time control applications:

- Actions such as limiting the maximum queue length or imposing constraints on the ramp delay have been shown to improve equity, however, there is no guarantee that the solution is truly optimal.
- Zhang and Levinson's approach (2005), while innovative, relies on a non-linear weighting function for travel time which may be difficult to determine in practice – a fact acknowledged by the authors. Moreover, in the control strategy presented, it was assumed that freeway congestion can be eliminated through ramp metering, which is not always possible in real-world networks. The strategy also relies on a fixed ramp coordination scheme which cannot be modified to respond to changes in freeway conditions, and whose effectiveness can only be evaluated after observing the impacts on freeway performance.
- In the approach adopted by Bellemans et al. (2006a), a constant weighting factor is used to adjust the emphasis placed on ramp delay in the objective function. However, since longer delays are not weighted more heavily, there is no guarantee that the resulting solution will be more equitable.

- The proposal by Yin et al. (2004) to develop a new objective function for ramp metering also has merit. However, the approach relies on a concave transformation function, and it is not clear how such a function would perform in practice, or if a suitable function even exists. Moreover, the approach requires detailed origin-destination data, and is intended for implementation as part of a system-wide coordinated optimal control strategy which have traditionally not seen widespread adoption in the field.
- The multi-objective optimization approach developed by Meng and Khoo (2010) provides good insight into the trade-off between equity and efficiency, but results in a set of Pareto-optimal solutions, rather than a single solution which can be implemented by the ramp controller. Given the formulation of the problem and the computational requirements, the authors acknowledge that the approach is not appropriate for real-time control. It is also important to note that the equity index proposed by Meng and Khoo (2010) only considers the ramps with the highest and lowest average ramp delay. By focusing strictly on the extremes, the equity implications associated with the distribution of ramp delay are ignored.

From the above discussion, it is clear that:

1. Equity is an important consideration in ramp control applications
2. Most ramp metering algorithms focus only on efficiency, and consider equity in an ad hoc way, if at all
3. There has only been limited research on incorporating equity into ramp metering algorithms
4. There is a need for a new control strategy which is capable of trading off equity and efficiency objectives within the same framework

Given the importance of equity in ramp control applications, and the significant potential to improve on current practice, the following research objective was defined.

**Research Objective:**

*To develop a new ramp metering algorithm for freeway corridors which better meets the needs of system operators and users by considering both equity and efficiency in the control problem.*

To ensure the new algorithm operates as effectively as possible in achieving the above objective, several key requirements were identified. In particular, it was felt that the proposed approach should:

- Utilize **real-time data** from traffic sensors to develop a **coordinated** control plan
- Be sufficiently **flexible** to support the future development of **integrated control applications** for improved coordination of freeway and arterial operations

Moreover, it was also felt to be important to develop an algorithm which explicitly accounts for uncertainty. Since traffic flow is a random process, the impact of any control action can only be assessed in terms of probabilities; an algorithm which takes such probabilities into account is more likely to achieve an outcome which meets the control objectives while respecting the risk-tolerance level of the system operator. While existing algorithms may address uncertainty to a limited extent, for example, by applying a conservative assumption for the freeway capacity, few algorithms have modelled uncertainty in any systematic way.

From a review of current practice, dynamic Bayesian decision networks were identified as an appropriate framework for developing the new algorithm to satisfy the above requirements. Bayesian networks are ideally suited for dealing with uncertainty, allowing control solutions to be developed which reflect the stochastic nature of traffic flow and associated risk. As evidence is received, beliefs about the system are updated, providing a means to track network performance over time. By incorporating utility nodes which represent preferences for different outcomes, an optimal control strategy can be selected which balances competing objectives in accordance with local attitudes and values – all within a single framework. Bayesian networks are also flexible; once the initial algorithm has been developed, it is a relatively simple matter to expand the network to capture different aspects of the control problem, such as integration with the arterial network or traffic diversion.

Accordingly, the new ramp metering algorithm was developed as a dynamic Bayesian decision network. The effectiveness of the algorithm was tested using a micro-simulation approach. The performance of the algorithm was compared against the “no control” case, as well as the popular ALINEA algorithm. Specific details of the new algorithm and its underlying structure are presented in Section 7.

## 5.2 Guiding Principles

In developing the new algorithm, a number of principles were identified to guide the research effort. In particular, the algorithm:

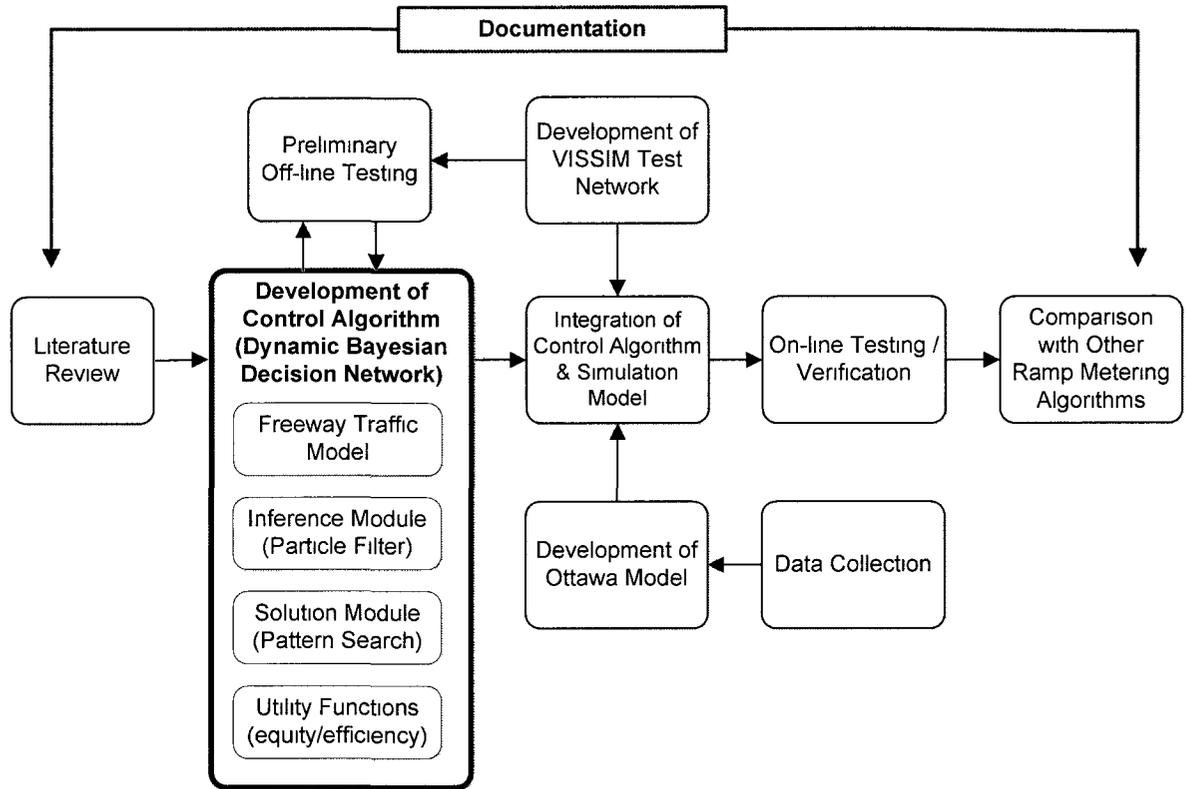
- Should be theoretically sound but application-driven; practical considerations regarding agency requirements, data availability, and cost should be paramount
- Should be designed with ultimate field implementation in mind
- Should avoid unnecessary complexity
- Should be easy to calibrate and use. To the greatest extent possible, a “black box” approach should be avoided
- Should be scalable to any network size
- Should make use of sensor data available in real-time
- Should be based on technology that is either currently available, or anticipated to be available within the next 5 to 10 years
- Should incorporate realistic assumptions for driver behaviour – For example, an algorithm that calculates the optimal diversion rate for a congested freeway is of limited value if there is no way to ensure this diversion rate actually occurs

The above principles played a key role in all phases of the research, particularly the up-front tasks involving the design of the algorithm and its various components. The research methodology is described in the following section.

## 6 RESEARCH METHODOLOGY

### 6.1 General Framework

The research methodology is illustrated in Figure 6-1. In general, the research involved three main tasks: conducting the literature review, developing the control algorithm, and implementing the algorithm in a simulation environment for testing and evaluation. Results from the literature review fed into the development of the various components of the algorithm, including the Freeway Traffic Model and associated utility functions, as well as the modules for carrying out probabilistic inference and finding the optimal solution to the control problem. Following an iterative process of off-line testing and refinement, the algorithm was integrated with the simulation software VISSIM for on-line testing and verification. Initially, a simple test network was used for fine-tuning the algorithm and evaluating its performance in relation to the research objectives, in particular, its ability to trade-off equity and efficiency within the same framework. Comparison with the popular ALINEA ramp control algorithm was also carried out to ensure reasonable performance under efficiency criteria alone. Once satisfactory results were achieved, the algorithm was applied in a more realistic model of Ottawa, Ontario developed specifically for this research. This final validation/verification phase was essential for establishing the overall merit of the new algorithm, and provides a basis for moving forward with future work.



**Figure 6-1 Research Methodology**

## 6.2 Literature Review

To gain an appreciation of freeway traffic control in general and ramp metering in particular, a literature review was carried out. This literature review provided insight into the features and limitations of existing ramp control strategies, and provided a basis for developing the research objectives outlined in Section 5. As part of the initial phase of work, a number of research databases were consulted, including:

- ASCE – American Society of Civil Engineers
- CISTI – Canada Institute for Scientific and Technical Information
- Ei Engineering Village (allows for combined searching of the Compendex, Inspec, & GEOBASE databases)
- IEEE – Institute of Electrical and Electronics Engineers
- ITE – Institute of Transportation Engineers
- ScienceDirect
- TAC – Transportation Association of Canada
- TRB – Transportation Research Board
- TRIS / NTL

In addition, a general internet search was carried out, and specific publications were examined for relevant articles, including: Transportation Quarterly, Transportation Research Parts A to F, Transportation Research Record, ITE Journal, etc. For the most part, the search was limited to articles published within the last 10 to 15 years, with greater emphasis given to more recent publications. Search terms were initially quite general (i.e. “ramp\* AND meter\*”) in order to generate a large list of potentially relevant articles, with the list narrowed and refined as the literature review proceeded.

Given the wealth of information available, the initial review was structured to provide a good introduction to the different aspects of ramp metering. As the review progressed, certain issues were identified as being of particular interest (i.e. equity), and increasing attention was focused on these areas. Rather than reviewing every control algorithm reported in the literature (a time-consuming task!), an effort was made to gain an appreciation of the different types of algorithms available, their main benefits, and potential opportunities for improvement. Particular attention was given to algorithms implemented in the field, to identify characteristics essential for successful real-world deployment.

Subsequent to the initial literature review, more detailed searches were carried out on specific topics of interest, including:

- Dynamic Bayesian decision networks
- Particle filters
- Utility functions
- Traffic state estimation
- Bayesian model development

Findings from the literature review are predominantly found in Sections 2, 3, and 4, which provide a background on ramp metering and set the context for the study. However, given the scope of the research exercise, additional information on specific topics can be found throughout this document. For example, Section 7.4.1 contains a review of the current literature on the prediction of flow breakdown. Overall, the literature review provided the foundation for developing the new ramp control algorithm and its various components. An overview of the new algorithm and its key features is presented in the following section.

# 7 THE RAMP METERING ALGORITHM

## 7.1 General Algorithm Structure

The general structure of the new ramp metering algorithm is presented in Figure 7-1. As illustrated, the algorithm consists of four main modules:

- The **Freeway Traffic Model** is used to model freeway operations by representing key relationships between the different variables which characterize the freeway system. The Freeway Traffic Model has been formulated as a Bayesian network, expanded to include utility and decision nodes for solving the control problem.
- The **Inference Module** is applied in conjunction with the **Freeway Traffic Model** to accomplish one of two main tasks: tracking and prediction. In tracking mode, beliefs about the current freeway state are continuously updated as new information is received from traffic sensors. In prediction mode, the model is used to predict future conditions under different ramp metering scenarios.
- The **Demand Prediction Module** uses data from traffic sensors to develop estimates of future ramp and mainline demand over the prediction horizon. This information is used as input to the Freeway Traffic Model when applied in prediction mode.
- The **Solution Module** is used to determine the “optimal” ramp metering rates by applying the Freeway Traffic Model in prediction mode for various ramp metering scenarios and finding the solution with the maximum utility.

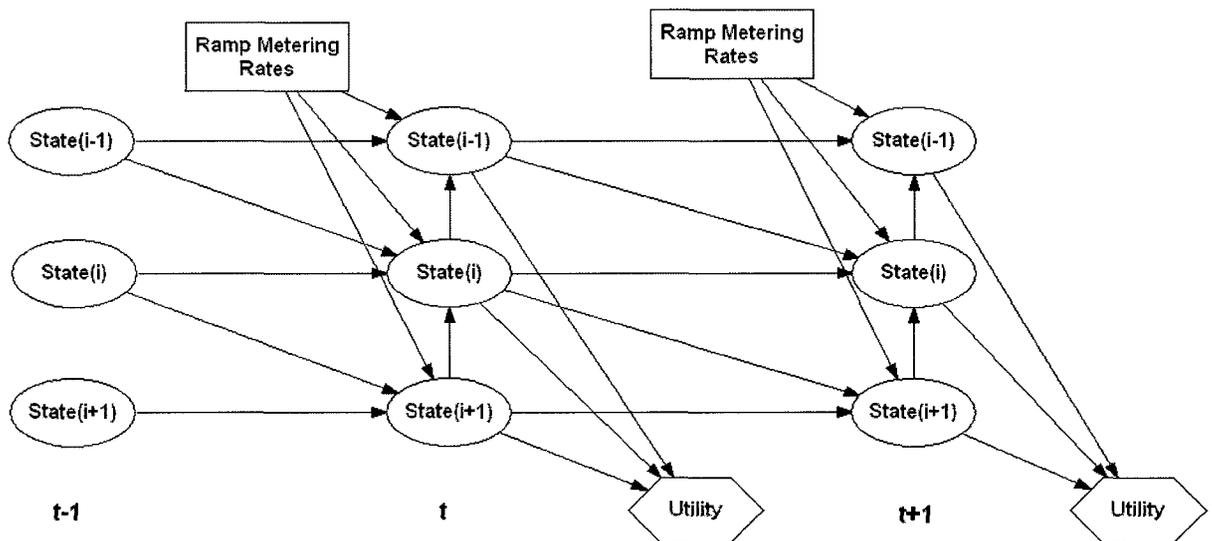
In essence, the algorithm uses the Freeway Traffic Model in tracking mode to estimate the current freeway state, with inference carried out using a standard particle filter. This estimate can then be used as a starting point for predicting future freeway conditions over a specified time period as required by the solution module in determining the optimal control parameters. The solution module calls the Freeway Traffic Model in prediction mode to determine the expected freeway performance under different ramp metering schemes. Based on this expected performance, utility estimates are produced which allow the solution module to refine the optimal solution. This iterative process continues until convergence has been achieved, or time constraints have been exceeded. The resulting “optimal” ramp metering rates are then modified as necessary to reflect operational constraints and implemented over the next control interval, impacting freeway operations. Such impacts are reflected in the freeway sensor data as the process is repeated.



The importance of equity versus efficiency is established via the “weighting terms” in the multi-attribute utility function, which can be tailored to reflect the objectives of the transportation agency or users of the system. Drivers’ preferences for ramp delay versus freeway congestion can also be captured in the utility function by applying different weights to the various attributes that define freeway efficiency. Additional information on the development of the utility function can be found in Section 7.6.

### 7.3 Bayesian Network Formulation

The new ramp metering algorithm has been implemented as a dynamic Bayesian decision network. Accordingly, a Bayesian network was developed illustrating the relationships between the key variables of interest. To model freeway operations, the Bayesian network depicts the freeway as a series of segments, with each segment described by a unique set of nodes representing its physical and operational characteristics. Figure 7-2 illustrates the interaction between adjacent freeway segments over time, simplified so that all variables representing the state of segment  $i$  at time  $t$  are collapsed into a single node. Decision and utility nodes are also shown.



**Figure 7-2 Simplified Bayesian Network of the Freeway Control Problem**

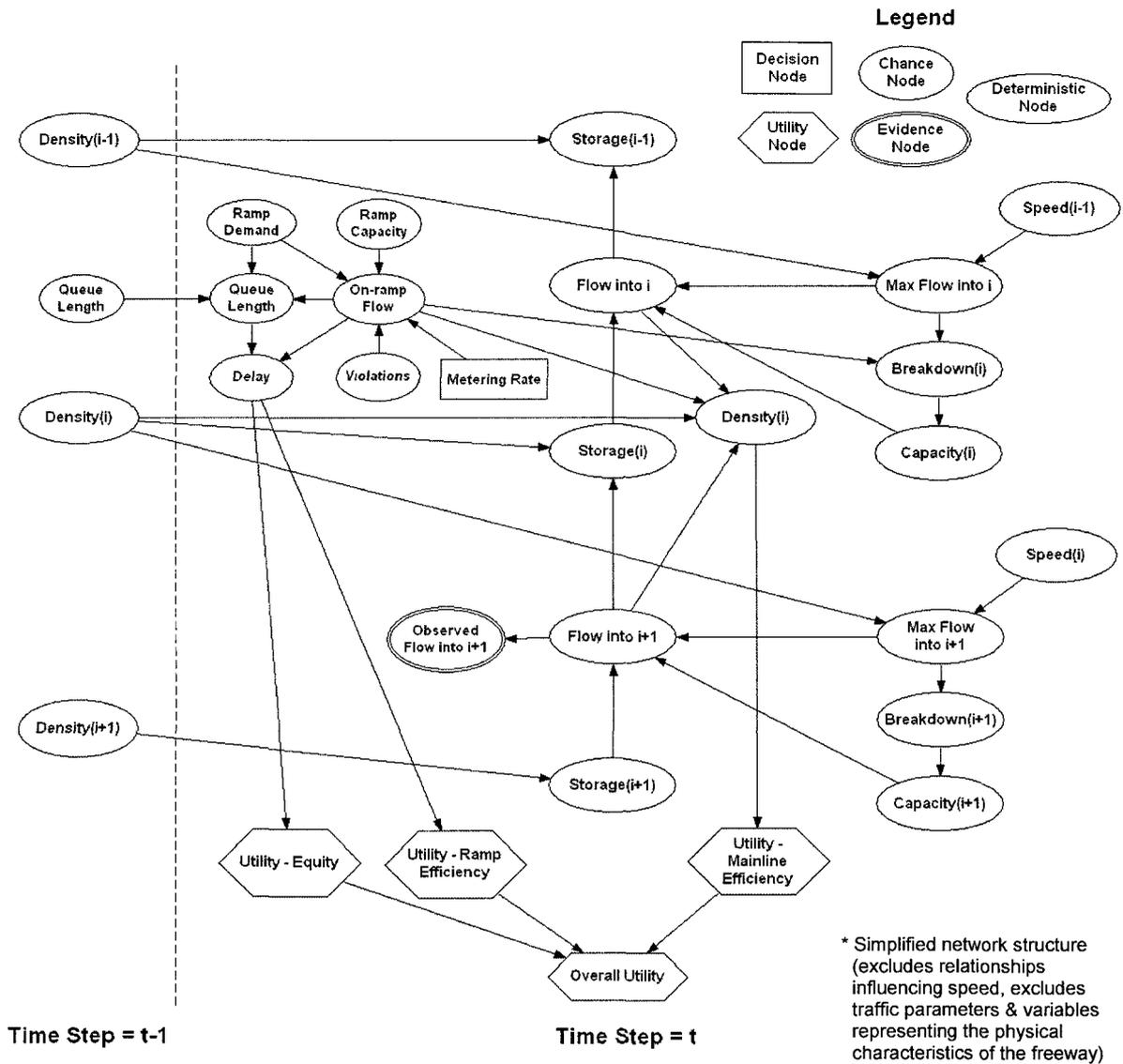
In reviewing Figure 7-2, discrete time steps are shown along the horizontal axis, while freeway segments are shown along the vertical axis. This dynamic representation of the

freeway system provides a means to model the state of each freeway segment from one time step to the next. As evidenced by the network linkages, it is assumed that only segments immediately upstream and downstream of a given segment have a direct impact on the segment's current operating state.

Figure 7-3 provides a more detailed representation of the Bayesian network for a single freeway segment *i* with an on-ramp, including key linkages to upstream and downstream segments. For legibility, the network has been simplified, although all main relationships are shown.

Together, the variables and relationships illustrated in Figure 7-3 make up a macroscopic model of freeway flow. Within the context of the new algorithm, this macroscopic model is known as the Freeway Traffic Model. As suggested by the network structure, the density of a freeway segment depends on the previous segment density, adjusted for the number of vehicles entering and departing during the current time step. The flow from one segment to another is determined by the speed and density in the upstream segment, the capacity entering the downstream segment, and the available storage space in the downstream segment. In turn, the storage space is influenced by the flow out of the downstream segment, while the capacity depends on whether or not flow breaks down, which is a function of the mainline and ramp approach flows.

The Freeway Traffic Model forms the basis for the Bayesian network structure shown in Figure 7-3, with utility and action nodes introduced to model the impact of different ramp metering rates on network performance. A more detailed explanation of the variables and relationships encoded in the Bayesian network can be found in Section 7-4, which provides additional information on the structure and key features of the Freeway Traffic Model.



**Figure 7-3 Bayesian Network Illustrating the Nodes Interacting with Segment ‘i’**

In interpreting Figures 7-2 and 7-3, the following comments should be noted:

- The Bayesian network diagram in Figure 7-3 includes both chance and deterministic nodes. A chance node has a probabilistic relationship with its parents; its value is expressed as a probability distribution over a range of possible values. In contrast, the value of a deterministic node can be uniquely defined as a function of its parents’ values. If the parent values are known, the value of a deterministic node can be computed with certainty. In the case of the Bayesian network model, many of the nodes are deterministic, and can be expressed as a function of their parents. Traffic parameters (such as the capacity or queue density) are treated as chance nodes, since their value is uncertain, and can change over time. Probabilistic terms have also been added to certain traffic relationships,

reflecting the stochastic nature of traffic flow, and the limitations of current models.

- The utility nodes in the Bayesian network are used in conjunction with the action (or decision) nodes to determine the optimal course of action. More information on the utility functions used in the ramp metering algorithm can be found in Section 7.6.
- Evidence nodes in the Bayesian network represent traffic sensors. Due to space limitations, many of the evidence nodes have been excluded from Figure 7-3; the one evidence node that is shown is for illustrative purposes only, and may not be reflective of the actual sensor position along the freeway.

In developing the control algorithm, it was generally assumed that traffic sensors would be present on all on- and off-ramps for determining the volume of traffic entering and exiting the freeway. An additional sensor was assumed to be located at the entrance to each on-ramp for assessing ramp demand and monitoring queue spillback. On the freeway mainline, it was assumed that sensors would be located near the terminus of the on- and off-ramp speed change lanes, the former to provide evidence on traffic breakdown at ramp merges, the latter to develop off-ramp exit percentages, and also to provide additional evidence for belief updating. All of the sensors must be capable of providing traffic flow data; it was assumed that the mainline (and queue spillback) sensors would also be capable of detecting vehicle speed. Occupancy data is not used in the current version of the model although it could conceivably be added in the future.

Increasing the number of traffic sensors will improve the accuracy of the freeway model. However, from a cost perspective, it is unrealistic to install traffic sensors indiscriminately. Moreover, not all sensors may be functioning at a given point in time. As a result, it is important that the network model function reliably with only a limited number of sensors in operation.

- Evidence nodes provide the basis for belief updating in Bayesian networks. Most commonly, evidence is used to refine probabilities once the system model has been applied. However, evidence can also serve as input to the system model. To do so, the evidence nodes must be structured as root nodes, rather than leaf nodes as is traditionally the case. Such “evidence reversal” is recommended as a way to improve probabilistic inference in dynamic Bayesian networks (Murphy 2002; Kanazawa et al. 1995). By structuring the system inputs as root nodes, there is no need to estimate the inputs and then calculate the likelihood of the evidence. For the Freeway Traffic Model, “input” nodes include the on-ramp demand and ramp flow entering/exiting the freeway, as well as the mainline demand approaching the study area.
- In the Bayesian network representation, it is assumed that traffic evidence is available at every time step, and the network is updated accordingly. In reality, the update interval may be considerably shorter than the evidence interval. It is

common for traffic sensors to provide data every 20 to 30 seconds. However, a shorter interval (in the order of 10 seconds) is needed for updating the traffic model, since the model is based on the assumption that no vehicle can pass more than one segment boundary in a single time step (refer to Section 7.4). Since all probabilistic updating is done after the evidence is received, the network model is always one sensor reading behind reality.

- The Bayesian network shown in Figures 7-2 and 7-3 is based on the assumption that the current state of a freeway segment is dependent only on conditions in the current and previous time step. However, the model used to predict flow breakdown (described in Section 7.4.1) requires flow rates from several preceding time steps. Although this historical dependency is not illustrated in the Bayesian network, it nonetheless forms a key component of the breakdown model.
- The above diagrams do not distinguish between the two modes of operation in which the Bayesian network will be used. In tracking mode, the network is used to track operations on the freeway network, refining probabilities based on available evidence. In predictive mode, the network is used to predict future operating conditions under different ramp metering scenarios in order to determine which scenario is optimal. In essence, the traffic states estimated in tracking mode become the starting point for the projections developed in prediction mode.

While the two operating modes rely on the same underlying model formulation, differences do exist. In tracking mode, evidence from the preceding observation interval is used to update the freeway model. All inputs are known, including the ramp demand and mainline flow entering the study area. In prediction mode, evidence from preceding intervals is used to forecast ramp and mainline demand over the prediction horizon. Since such forecasts are subject to error, the corresponding operational predictions are also more uncertain. This is particularly true since there is no evidence collected over the prediction horizon to refine the traffic state probabilities.

The Bayesian network representations presented in this section are intended for illustrative purposes only; the actual freeway model implemented in the ramp metering algorithm includes additional variables and relationships, which in some cases may impact the linkages shown (for example, the utility is actually a function of freeway speed, not density as indicated in Figure 7-3). Nonetheless, Figures 7-2 and 7-3 provide a reasonable depiction of the basic network structure, and the main interactions between key variables.

#### 7.4 The Freeway Traffic Model

In the new ramp metering algorithm, control decisions are based on traffic projections for different ramp metering scenarios, allowing the algorithm to respond to problems before they arise. For such predictive control to be effective, the algorithm must be capable of predicting the onset of flow breakdown. The algorithm must also be capable of tracking mainline congestion over time in order to trade-off competing objectives relating to freeway performance.

Given these requirements, a macroscopic model of freeway flow is needed which can be readily incorporated into the Bayesian network comprising the ramp metering algorithm. While various macroscopic models have been developed, two feature prominently in the literature for ramp metering control: the METANET model and variants of the cell transmission model (refer to Appendix F). Based on a review of the various options, a modified version of the cell transmission model was developed for inclusion in the ramp metering algorithm.

The cell transmission model was first proposed by Daganzo (1994). As originally defined, the cell transmission model is a relatively simple first order model which captures the forward propagation of traffic flow and the backward propagation of congestion. The model is discrete in both space and time. To apply the model, the highway is divided into homogenous cells such that no vehicle can cross more than one cell boundary during a single time step. The model attempts to predict the number of vehicles  $n$  in each freeway cell  $j$  at time  $k$  by applying the conservation of vehicles principle at each time step:

$$n_j(k+1) = n_j(k) + q_j(k) - q_{j+1}(k)$$

The flow  $q$  into each cell is defined as the minimum of: the flow that can be supplied by the upstream cell, the flow that can be absorbed by the cell given its available storage, and the maximum observable flow (i.e. the flow capacity). Thus, the number of vehicles in a given cell  $j$  is equal to the previous number of vehicles in the cell, plus the number of vehicles arriving in the current time step, minus the number of vehicles departing.

Several research teams have modified the cell transmission model for specific applications. The asymmetric cell transmission model developed by Gomes and Horowitz (2006) includes a modified treatment for on-ramp merges. Sun et al. (2003) used the cell transmission model as the basis for a switching state-space model with two discrete modes of operation, congested and free-flow. The ‘compositional traffic model’ developed by Mihaylova et al. (2007) is also noteworthy. Whereas Daganzo modelled forward and backward traffic waves using deterministic relationships, Mihaylova et al. defined stochastic sending and receiving functions, and also extended the model to include consideration of the speed dynamics within each cell.

For the current research, a new variant of the cell transmission model was developed which explicitly captures the probabilistic nature of flow breakdown. In many ways, it is similar to the ‘compositional traffic model’ developed by Mihaylova et al. (2007) in that it includes stochastic relationships and models changes in both density and speed. It also incorporates elements of the switching state-space model developed by Sun et al. (2003) in that discrete modes of operation are considered separately. However, unlike Sun et al., both congested and uncongested cells can be present on the same freeway section as traffic conditions evolve over time.

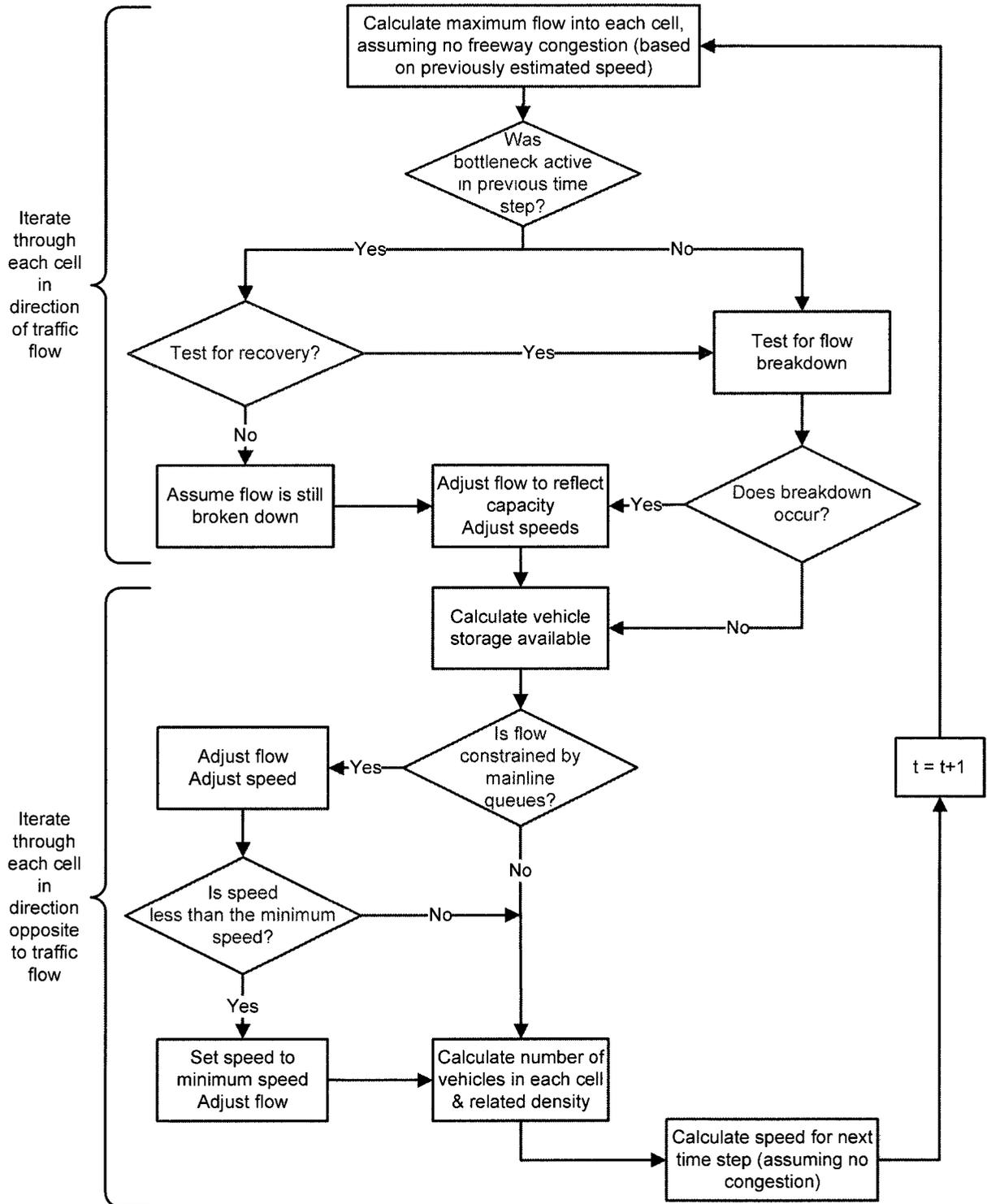
Within this thesis, the proposed extension of the cell transmission model is denoted as the Probabilistic Freeway Traffic Model, or simply the Freeway Traffic Model (FTM). A flow chart of the model is presented in Figure 7-4, while a more complete description, including all relevant equations, can be found in Appendix G.

The Freeway Traffic Model as proposed is a relatively simplistic model designed to achieve a balance between accuracy and ease of implementation/use. The model is able to simulate both the forward wave of traffic flow which occurs under uncongested conditions, and the backward wave of queued vehicles which occurs when flow breaks down. To apply the model, the flow entering a given freeway cell is restricted to the lesser of the following two values:

- The flow contributed by the upstream cell (if breakdown does not occur), or the queue discharge flow (if breakdown does occur)

- The available storage space, assuming a maximum storage capacity based on the expected queue density (as influenced by shockwave behaviour)

To better capture the breakdown phenomenon, the traffic model includes a new relationship which predicts the probability of flow breakdown at on-ramp merges as a function of ramp and mainline flow. By including a more explicit representation of flow breakdown in the model formulation, the factors influencing breakdown can be considered more fully, supporting the development of control solutions. Details of the probability model are provided in Section 7.4.1.

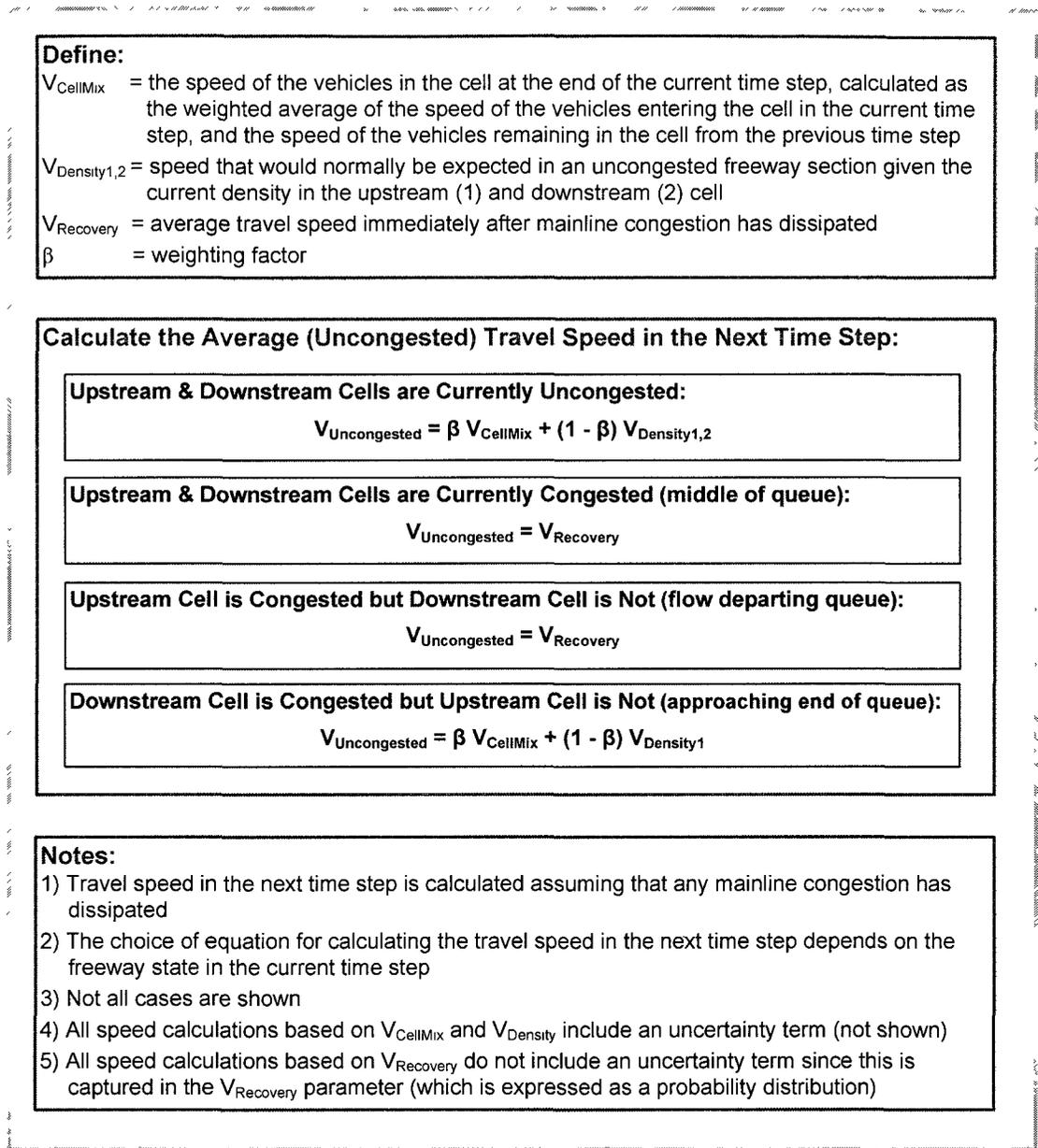


**Figure 7-4 Flow Chart of the Probabilistic Freeway Traffic Model**

The flow chart in Figure 7-4 provides an overview of the main calculations in the Freeway Traffic Model and how they fit together. As depicted, the model starts by computing the maximum flow which is able to enter a given freeway cell under the assumption of no mainline congestion. This maximum flow is a function of both the number of vehicles upstream of the cell, and the average (uncongested) travel speed. Based on the maximum flow entering the cell, the probability of breakdown is computed; whether or not breakdown actually occurs is determined by drawing a sample from the uniform distribution. In the event that breakdown is predicted, the speed and flow variables are adjusted accordingly to reflect queue discharge from the active bottleneck.

Next, the available storage capacity in the freeway cell is computed. The available storage is based on the maximum number of vehicles allowed in the cell (accounting for shockwave effects), minus the number of vehicles that were present in the cell during the previous time step, adjusted for the number of vehicles that departed during the current time step. If the remaining storage is insufficient to accommodate the projected incoming flow, the flow and speed are again adjusted, such that the speed does not fall below a certain minimum threshold. Once the flow rates have been determined, the number of vehicles in each cell is calculated by taking the previous vehicle estimate and adjusting for the number of vehicles entering and departing during the current time step. As a final task, the average travel speed is computed for the next time step assuming that all mainline congestion has dissipated (refer to Figure 7-5), and the process is repeated. Additional steps within the model deal with the issue of predicting on- and off-ramp flows, testing for freeway recovery, updating the parameter estimate for mainline capacity, and calculating ramp queues/delay.

The model is applied by iterating through each freeway cell in sequence; first in the direction of traffic flow to assess the maximum flow contributed by upstream cells and probability of flow breakdown, then in the direction opposite to traffic flow to assess storage constraints associated with queue spillback.



**Figure 7-5 Calculation of the Uncongested Travel Speed**

To run the Freeway Traffic Model, several parameters must be specified. A summary of these parameters is provided in Table 7-1. In addition, the model also requires information on the physical characteristics of the freeway, including:

- The length of each freeway segment and the corresponding number of lanes;
- The ramp storage capacity; and
- The length of auxiliary speed change lanes for on-ramp merging.

**Table 7-1 Summary of Model Parameters**

Parameter	Description	Assumed Value <sup>1</sup>
Speed-density relationship for uncongested conditions	<ul style="list-style-type: none"> <li>Estimates the average travel speed in uncongested freeway segments as a function of the cell density</li> <li>Corresponds to <math>V_{\text{Density}}</math> in Figure 7-5</li> <li>Intercept term represents the free-flow speed</li> <li>Within the Freeway Traffic Model, the speed-density relationship feeds into the calculation of the uncongested travel speed in the next time step. As an intermediate calculation, uncertainty is not considered, and all underlying parameters are assumed to be fixed (however, uncertainty in the final speed estimate is considered – see below)</li> </ul>	$V = V_{\text{ff}} + \alpha K$ where: $V$ = average speed $V_{\text{ff}}$ = free-flow speed = 112 km/hr $\alpha$ = -0.537 $K$ = average density
$\beta$ parameter (used in speed model)	<ul style="list-style-type: none"> <li>For a given freeway segment, the uncongested travel speed in the next time step is estimated as a weighted function of the speed of the vehicles in the cell at the end of the current time step, and a density-based speed</li> <li><math>\beta</math> represents the weighting factor, and is applied as defined in Figure 7-5</li> </ul>	Normal Distribution Mean: 0.25 Std Dev: 0.03
Queue discharge speed	<ul style="list-style-type: none"> <li>Represents the average speed in the cell immediately downstream of the mainline queue (inside the bottleneck)</li> <li>Within this cell, the speed varies. As vehicles accelerate away from the queue, the speed is low and the density is high. As distance from the queue increases, vehicle speeds also increase, with a corresponding drop in density. Thus, the average speed for the cell as a whole will change depending on the cell length</li> <li>Cells located further downstream are assumed to be unaffected by the queue (i.e. speeds have returned to normal)</li> </ul>	Normal Distribution Mean: 84 km/hr Std Dev: 5 km/hr
Recovery speed	<ul style="list-style-type: none"> <li>Provides an estimate of the uncongested travel speed in the event that a congested freeway segment is able to recover in the next time step (since the segment is initially congested, it is not appropriate to use the current density to estimate the uncongested travel speed)</li> <li>Given the uncertainty involved, the speed distribution essentially covers the full range of travel speeds that are typically observed under non-failure conditions</li> </ul>	Normal Distribution Mean: 95 km/hr Std Dev: 5 km/hr
Minimum speed	<ul style="list-style-type: none"> <li>To prevent the average speed within a queue from dropping to zero, a minimum speed variable was introduced</li> <li>This variable ensures that there is always some vehicle movement between cells in each time step</li> </ul>	Normal Distribution Mean: 20 km/hr Std Dev: 5 km/hr
Queue density	<ul style="list-style-type: none"> <li>Represents the expected density within the traffic queue (as distinct from the jam density which is the maximum density observed)</li> <li>At any point in time, a freeway segment within the queue may have a density that is substantially higher or lower than the average queue density depending on the movement of traffic shockwaves. This is reflected in the relatively high standard deviation for this variable</li> </ul>	Normal Distribution Mean: 58 veh/km/ln Std Dev: 8 veh/km/ln

Parameter	Description	Assumed Value <sup>1</sup>
Critical density	<ul style="list-style-type: none"> <li>The critical density is used to determine when freeway congestion has dissipated (refer to Table 7-2 below)</li> <li>For the most part, the critical density is determined by the model in real-time based on the density actually observed when flow breaks down</li> <li>One exception exists: overlapping mainline queues may obscure individual bottlenecks along the corridor which only become evident again when the queues begin to dissipate. The critical density for such bottlenecks cannot always be reliably determined from the density at flow breakdown due to the confounding influence of the overlapping queues. In such cases, an assumed value for the critical density is applied. Since such situations are rare, the value used for the critical density is unlikely to have a major impact on the model performance</li> </ul>	Normal Distribution Mean: 28 veh/km/ln Std Dev: 1 veh/km/ln
Queue discharge flow (congested capacity)	<ul style="list-style-type: none"> <li>The mean capacity is initialized to 2300 vphpl, but this value is updated by the model based on evidence received from traffic sensors (refer to Table 7-2 below for more information)</li> </ul>	Normal Distribution Mean: 2300 vphpl Std Dev: 150 vphpl
On-ramp capacity	<ul style="list-style-type: none"> <li>In the current version of the model, it is assumed that the on-ramp capacity is adequate to accommodate the demand</li> <li>As a result, the on-ramp capacity was assigned an arbitrarily high capacity value</li> <li>Typically, capacity is based on peak 15-minute flows. However, in the case of the Freeway Traffic Model, capacity is applied over a much shorter time interval (i.e. 10 seconds). Within such a short interval, high bursts of demand are possible, resulting in hourly flows that may be much higher than the standard ramp capacities commonly cited</li> </ul>	3500 vph
Off-ramp capacity	<ul style="list-style-type: none"> <li>Not used in the current version of the model</li> <li>It is assumed that the off-ramp capacity is adequate to accommodate the demand</li> </ul>	N/A
Noise term for speed estimate	<ul style="list-style-type: none"> <li>Used to account for uncertainty/error in the speed calculation</li> </ul>	Normal Distribution Mean: 0 Std Dev: 3 km/hr
Noise term for maximum flow into segment	<ul style="list-style-type: none"> <li>Used to account for uncertainty/error in the maximum flow calculation</li> </ul>	Normal Distribution Mean: 0 Std Dev: 100 vphpl
Noise term for speed measurements	<ul style="list-style-type: none"> <li>Used to account for error in the sensor reading</li> <li>Since the speed observed at the cell boundary may be different than the average speed of the vehicles entering the cell, a relatively high standard deviation was assumed (refer to Table 7-2)</li> </ul>	Normal Distribution Mean: 0 Std Dev: 10 km/hr
Noise term for flow measurements	<ul style="list-style-type: none"> <li>Used to account for error in the sensor reading</li> <li>When using sensor data to estimate model inputs, a smaller noise term was applied (refer to Table 7-2)</li> </ul>	Normal Distribution Mean: 0 Std Dev: 100 vphpl

<sup>1</sup> For use with the VISSIM test network described in Section 8.2.

Of the parameters listed above, many can be estimated directly from traffic data. Others may be influenced by more abstract considerations. In all cases, careful judgement must be applied.

In developing the Freeway Traffic Model, a number of issues were encountered, many of which were resolved, others which were deferred for future work. The following table provides a detailed discussion of these issues, along with the various measures that were adopted to improve the performance of the model for ramp control applications.

**Table 7-2 Model Features, Assumptions & Key Issues**

<p><b>a) Parameter uncertainty</b></p>
<p>Within the Freeway Traffic Model, most parameters are treated as stochastic, allowing a more explicit representation of the problem and the associated sources of uncertainty.</p> <p>Parameters such as the queue density or queue discharge flow are not just uncertain in the sense that their value is unknown, but also due to the fact that their value may change – randomly from one time step to another, as well as in response to changing external factors such as weather.</p> <p>As a result, the parameters are assumed to be dynamic; at each time step, the parameter is sampled from a Gaussian distribution. This is done to avoid the problems associated with static parameters, which, once initially chosen, cannot be modified later on, even if the initial estimate turns out to be poor, or conditions change over time. Depending on the particles selected during the inference process, there is also a risk that the parameter distribution reflected in the particles will deteriorate to the point that only a few unique values are represented. By allowing the parameter values to change, such issues can be alleviated.</p> <p>During the inference process, the various model parameters are updated as new information is received. However, these updated estimates are not used to adjust the underlying probability distribution for each parameter, which is assumed to remain constant from one time step to the next. This ensures that the parameter values remain realistic, but also means that in sampling from the distribution, fewer particles are likely to take on a value which lies close to the true value, since the distribution has not been refined to reflect the most recent conditions. Moreover, since the distribution is essentially static, it may be necessary to adopt a wider distribution (with more uncertainty) to ensure it captures the full range of values likely to be encountered under changing conditions.</p> <p>One exception to the above is the probability distribution for the congested capacity (queue discharge flow), which <i>is</i> updated over time based on evidence from traffic sensors (additional details provided below).</p>
<p><b>b) Estimation of queue discharge flows (congested capacity)</b></p>
<p>During the model testing process, it was found that the model is extremely sensitive to the probability distribution assumed for the congested capacity. If the distribution is centred about an average capacity that is too high or too low, the model has difficulty predicting the magnitude and duration of mainline queuing.</p> <p>It was further discovered that the capacity depends on the conditions at the start of the bottleneck. If turbulence from on-ramp merging is high, the capacity tends to be lower. If merging activity is reduced (i.e. due to ramp metering), the capacity is slightly higher. It is not known whether these findings also apply to real-world networks, or were simply a by-product of the simulation. Nonetheless, there would appear to be value in allowing the model to update the probability distribution for the congested capacity based on data received from traffic sensors.</p> <p>Accordingly, in each time step, for each particle, the mean value of the probability distribution is updated based on the average of the capacity values stored from the previous three updates. If, in one of these previous updates, breakdown occurred, the stored capacity is the value actually used in the model calculations. If breakdown did not occur, the value stored is simply the current mean capacity for the update interval in question. During the process of belief updating, only those particles that are consistent with the sensor data are carried forward, and as a result, the selected particles should have an updated capacity distribution which more closely reflects the observed data (at least for the sections where flow has broken down). To ensure the updated mean capacity is realistic, upper and lower limits on the mean capacity are</p>

imposed. The averaging process was adopted to smooth out extreme values for defining the capacity distribution, since the capacity flow can fluctuate significantly from one sensor reading to the next.

In the course of conducting additional tests, it was found that the particles selected during the inference process do not always provide a good estimate for the congested capacity. Since the model has difficulty predicting shockwave behaviour, only a few particles may be consistent with the evidence as the freeway becomes increasingly congested, and these particles may not provide a representative sample of the congested capacity. Of the various options explored to address this situation, the most effective by far was the use of evidence reversal to estimate the congested capacity directly using data from traffic sensors (for those cells experiencing flow breakdown). Since it is not appropriate to apply the same data as both input to the model and evidence for developing particle weights, the following rule was applied: If all of the particles predict flow breakdown, then evidence reversal can be used to estimate the congested capacity for the affected segment. Otherwise, the sensor data is used as a standard leaf node in the Bayesian network for updating probabilities during the inference process.

The use of evidence reversal was found to be so effective that the following unexpected behaviour was observed: As the number of particles increased, the accuracy of the model decreased. It is hypothesized that this behaviour relates to the criterion for evidence reversal. With more particles, it is less likely that all of the particles will agree that flow has broken down, and therefore less likely that evidence reversal can be applied.

By using evidence from traffic sensors to update the probability distribution for the congested capacity, the model performance was found to improve significantly, particularly when combined with evidence reversal. At the same time, the number of particles used to represent the freeway state could be reduced, supporting model run-time objectives for real-time inference. While acceptable results for any one freeway scenario could be achieved by simply fine-tuning the capacity assumptions (without adopting the changes described above), these assumptions are not applicable under all conditions, and the results for other freeway scenarios would deteriorate. Instead, the approach adopted allows the model to work well under a wide range of operating states, including those likely to be encountered during ramp metering.

The use of evidence reversal for estimating capacity flows has certain implications, primarily related to the number of particles for representing the freeway state. Under the methodology adopted, evidence reversal is most effective if the number of particles is kept low. As a consequence, it was found that, at times, too few particles are being carried forward, particularly when the freeway is congested and the model has difficulty estimating the queue density under shockwave conditions. For the number of particles employed, it was found that better results could be achieved by increasing the sensor error (so that more particles are selected as being consistent with the evidence) and decreasing the level of uncertainty associated with certain parameters (so that more particles are concentrated around the most likely values). From a capacity perspective, the use of evidence reversal fully compensates for the above assumptions; since the capacity is estimated new each time step using sensor data, it is not particularly dependent on which particles are carried forward, or the level of parameter uncertainty.

Obviously, if evidence reversal were not employed, or the distribution for the congested capacity was assumed to be static, the above assumptions would need to be revised. In particular, the number of particles would need to be increased substantially, so that a larger, and therefore more representative sample is carried forward in each time step, mitigating the risk of obtaining a poor estimate for the posterior distribution. It would also likely be necessary to increase the uncertainty associated with the congested capacity in order to better predict rare events (which is less critical when using sensor data to estimate the capacity directly), and decrease the sensor noise to ensure that only particles more closely matching the evidence are carried forward. In both cases, additional particles would be required.

Possible modifications to explore in the future include:

- Applying the same capacity distribution to each particle, rather than updating the distribution for each particle individually as is currently the case
- Developing an approach for applying evidence reversal which would not require all particles to agree that flow has broken down (thereby allowing the number of particles to be increased, addressing other limitations of the model)
- Expanding evidence reversal to apply where flow has not broken down (however, this would require a fundamental shift in the model formulation)

### c) Dissipation of mainline queues (testing for recovery)

One of the major challenges in using the Freeway Traffic Model is predicting the end of freeway congestion – particularly in cases of temporary flow breakdown. Originally, the intent was to assume that breakdown conditions persist from one time step to the next if the segment density is greater than a pre-defined critical density. In practice, this criterion was found to be problematic, since the critical density is uncertain, and can only be known once flow breaks down. As a result, rather than try to estimate a critical density in advance, the density of the segment at flow breakdown was stored and used as a basis for determining when congestion had dissipated.

In cases where the probability of flow breakdown is low, but flow breaks down nonetheless, the critical density may be lower than typically observed. In such situations, there is a risk that congestion may quickly dissipate, yet the density increases due to higher flow rates entering the segment. Since the density is greater than the density when flow broke down, the model assumes the segment is still congested, when in fact it is not. A similar error may occur whenever the density at recovery is even slightly higher than the density at breakdown. To overcome this issue, the congestion criterion was modified as follows: if the segment density is 15% greater than the critical density observed when breakdown first occurs, the segment is assumed to be congested. However, if the density does not meet this criterion, it is assumed that congestion may have dissipated, and the probability of congestion relationship is re-applied to determine whether in fact this is the case.

Note that the same probability of breakdown relationship is applied regardless of whether the segment was congested in the previous time step or not. In reality, the probability of flow breakdown may be impacted by any pre-existing mainline queues.

### d) Speed estimation

Speed estimation within the model involves several steps. At the end of each time step, the travel speed in the next time step is estimated assuming that any freeway congestion has dissipated. During the next time step, this speed is then adjusted to reflect the actual freeway conditions predicted by the model.

- If flow breakdown is predicted at a particular cell boundary, the speed in the downstream cell is set equal to the queue discharge speed, while the speed in the upstream cell is calculated based on the anticipated capacity flow and previous cell density using the relationship  $flow = speed \times density$ .
- If mainline queues are spilling back onto upstream cells, restricting flow, then the speed is again calculated using the speed-flow-density relationship, with the flow based on the available storage capacity in the downstream cell, and the density based on the previous cell density. If the resultant speed is less than the minimum speed, then the minimum speed is adopted as the new speed (and the flow is adjusted accordingly)

Note that in both cases above, the adjusted speeds are used to describe the current freeway state, but are not used in any subsequent calculations (except for assessing the likelihood of the observed evidence). However, the initial uncongested travel speed is used to estimate the maximum flow between cells under conditions of no freeway congestion, and therefore plays a key role in the model calculations.

Calculation of the uncongested travel speed is illustrated in Figure 7-5. In freely moving traffic, this speed is based on two components: a density-based speed, and the previous speed of the vehicles entering/remaining in the cell. For the density-based speed, the density reflects the current state of the freeway segment, as well as the conditions anticipated downstream.

The Freeway Traffic Model does not directly capture the uncertainty inherent in the intermediate speed calculations and associated parameters used to evaluate the uncongested travel speed. Instead, such uncertainty is captured in the random noise variable which is added to the final speed estimate. Ideally, however, for a Bayesian network representation of the problem, it is preferable to treat each source of uncertainty separately.

In some respects, it could be argued that one of the weakest components of the Freeway Traffic Model is the method used to estimate the travel speed. Rather than apply a single equation which captures all flow regimes, different relationships are applied depending on the type of traffic flow. This results in a more

complex model formulation, with many parameters to tune. On the other hand, the proposed approach recognizes the inherent difficulty in developing a single speed equation which encompasses the unique characteristics of the various operating modes. In this respect, the proposed methodology is both straightforward and elegant; a speed is estimated for uncongested conditions, and then modified if the flow type turns out to be different. In the end, it is probably fair to conclude that the speed estimation approach used within the Freeway Traffic Model is not necessarily better or worse than that used elsewhere, merely an alternative method that fits well within the probabilistic framework employed. However, other speed models could certainly be added to the Freeway Traffic Model, if desired.

#### **e) Location of cell boundaries**

In the Freeway Traffic Model, the potential for flow breakdown is assessed at cell boundaries based on the number of vehicles wishing to cross the boundary during a single time step. As a result, the location of the cell boundary can have a significant impact on the performance of the model. This is not a major issue with known bottlenecks, since the cell boundary can be placed at the start of the observed queue. However, in some instances, breakdown has been observed to occur randomly over a given freeway section (Banks 2002). In other cases involving incidents, the location of breakdown is impossible to determine in advance. In such situations, the model will perform less accurately. The problem can be mitigated to a certain extent by minimizing the cell length.

#### **f) Model initialization**

To use the Freeway Traffic Model, an estimate of the initial freeway state (speed and density per cell) must be developed. If the model is set to begin tracking freeway operations overnight during the lightest hour of demand, it should be possible to obtain a relatively accurate estimate of the initial freeway state using the available sensor data. When volumes are low, the extent of error is also likely to be low. It becomes more challenging if the model begins tracking freeway operations when flows are heavier. To obtain as accurate an estimate as possible, the model should be initialized when the freeway is uncongested, before the onset of any mainline queues. Under such conditions, sensor readings are more likely to be applicable over a wider freeway section, leading to more accurate estimates of cell density and speed.

For the current model, the initial density was assumed to follow a normal distribution with a mean of 13 veh/km/lane, and a standard deviation of 3 veh/km/lane. Likewise, the initial speed was assumed to follow a normal distribution with a mean of 105 km/hr, and a standard deviation of 2 km/hr. The level of uncertainty associated with the initial freeway state is reflected in the standard deviation. As information is received from traffic sensors, the freeway state is gradually refined to better reflect the observed data.

In developing the initial density and speed assumptions, conditions along the entire freeway corridor were considered. However, more localized assumptions could also be developed using data from the nearest traffic sensors.

While the initial density and speed assumptions were found to work well in all of the off-line test scenarios that were investigated, it is anticipated that initialization of the Freeway Traffic Model may prove more challenging in real-world networks where conditions are more variable. Several options exist for improving the estimate of the initial freeway state:

- Several model runs could be carried out simultaneously, each incorporating different assumptions for the initial freeway state. After a few minutes, the model which best matches the observed data would be chosen to continue tracking freeway performance, and the other models would be dropped. This approach would require additional computing power during the initialization phase, however, once initialization was complete, this computing power could be re-allocated to solving the control problem (assuming the model is initialized prior to the start of ramp metering).
- Similarly, the number of particles used in the first few model updates could be increased to better represent the uncertainty surrounding the initial freeway state. After several updates, the number of particles could be reduced (again freeing up computing power for use in ramp metering control).

Previously collected data from the past several minutes may provide a useful starting point for initializing the model, particularly if the time required for updating is significantly less than the time between sensor readings. However, at some point, the model must “catch up” to the real-time observations.

### g) Working with sensor data

In their assessment of detection technologies for Intelligent Transportation Systems, Klein and Kelley (1996) define specifications for tactical decision-making (i.e. decisions made by control systems in response to real-time data). According to these specifications, traffic volume detectors used in freeway management should have a maximum error of no greater than  $\pm 2.5\%$  at 500 vphpl. For speed, the allowable error is  $\pm 1.6$  km/hr. In both cases, a 20 second data collection interval is assumed to apply. To meet this specification, the authors conclude that the sensor accuracy must be close to 100% in practical terms:

*For the postulated 20-second data collection interval for tactical control, an error of 0.07 in vehicle count for every 2.8 vehicles is implied at a flow of 500 vehicles per hour. Practically, this means that all vehicles must be detected during each 20-second interval. While no detector guaranteed 100-percent detection accuracy, some did perform with less than 1-percent error. (Klein and Kelley 1996, Section 11).*

Results from the literature confirm that vehicle detection error rates in the range of 1% to 2% are feasible, although the results are highly dependent on the type of technology employed. Martin and Feng (2003) provide a good overview of the different types of intrusive and non-intrusive detectors that exist, and their corresponding accuracy as determined from real-world trials. A summary of more recent test results is provided in Middleton et al. (2007).

In evaluating sensor accuracy, there are numerous ways to present the results. Often, data is aggregated over a 15-minute period or longer. However, doing so risks masking errors which cancel out. In addition, to develop a probability distribution for the sensor noise, the error must be expressed in terms of the frequency of occurrence (i.e. error of less than  $\pm x\%$ , 90% of the time), or alternatively, the mean and standard deviation of the error observations. This implies that multiple observations are needed (from the same and different sensors), covering the range of conditions likely to be encountered in the field.

In developing such a distribution, it is important that the observations correspond to the data collection interval being used for freeway control. Consider a traffic sensor with 2% error. For a flow rate of 1000 vphpl, this translates into an error of 20 vehicles. For a 20 second data collection interval, there are 180 sensor observations per hour. Thus, for the majority of observations, the error will be zero; only in a minority of cases will a miscount occur, and in these cases, the percentage error will be greater than 2%.

While a normal distribution is often used to approximate the sensor error (as was done in this research), it may not be the most appropriate for the conditions described above. According to Mihaylova et al. (2007), errors in flow measurement may be better represented as a Poisson distribution (actually, two distributions, one to estimate the number of false detections, and the other the number of missed vehicles). This is entirely feasible within the particle filtering framework used for inference, since one of the main advantages of particle filters is that they do not require Gaussian distributions to work. Several tests were carried out using a Poisson distribution to represent the flow sensor error, however, the results were generally inferior. Different hypotheses were developed to explain this behaviour, but further testing is required.

Within the Freeway Traffic Model, sensor data is used as both input to the model (through evidence reversal), and also as standard evidence for developing particle weights during the inference process. In testing the model, the sensor error assumed for the former was slightly less than assumed for the latter, even though both are derived from same underlying source. In the case of the latter, the sensor error has a major impact on which particles are considered to be consistent with the evidence. If the sensor error is too small, not enough particles will be selected. Thus, in some situations, it may be appropriate to select a larger sensor error than exists in reality. This is particularly relevant when working in simulation, where the actual sensor error is zero.

For model testing purposes, the sensor noise was assumed to follow a normal distribution as documented in Table 7-1 (for standard evidence) and Section 7.4.2 (for model inputs). In the case of flow evidence, the sensor noise was assumed to have a mean of zero and a standard deviation of 100 vphpl (or 0.56 veh/20 seconds). Given the characteristics of the normal distribution, this implies that 95% of the time, the error will be less than  $\pm 1.1$  vehicle for each 20 second observation. For a flow rate of 2000 vphpl (11.1 veh/20 seconds), the percentage error is thus expected to be less than 10%, 95% of the time. However, on average, the error will be zero. In the above formulation, the error is not dependent on the magnitude of the observed

flow. During peak conditions, the demand is relatively stable, and such an assumption is considered to be reasonable. However, a distribution based on the percentage error is also feasible.

Working with small time intervals poses a challenge when dealing with vehicles which have only partially crossed the cell boundary at the end of the data collection interval. Whether such vehicles are counted or not as part of the current observation can increase or decrease the flow rate significantly (a one vehicle difference over 20 seconds equates to 180 vph when expressed as an hourly flow). In the Freeway Traffic Model, fractional flows are allowed. While the original sensor data is expressed in whole vehicles, the added sensor error is not, given the assumption of a (continuous) Gaussian distribution.

#### **h) Inconsistency between point & segment data**

The speed estimates produced by the Freeway Traffic Model do not necessarily correspond to the speed at the cell boundary as estimated from traffic sensors. The Freeway Traffic Model predicts the average speed of traffic moving into the downstream segment over a 10-second time step. However, if traffic is accelerating away from a queue, or is decelerating as it joins the end of a queue, the speed at the cell boundary may not reflect the average travel speed. To account for this situation, adjustments were made to the probability distribution representing the evidence. By increasing the level of uncertainty associated with the speed observations, only those particles with significantly different speeds are discarded. This approach ensures that particles are not discarded in cases where the average speed is slightly different from the speed at the cell boundary, but also limits the usefulness of the speed evidence.

#### **i) Dealing with broken sensors**

In order to work correctly, the Freeway Traffic Model requires information on the ramp and mainline demand entering the freeway network. If the sensors capturing this data are broken, the model will not work. However, the same is not true for intermediate traffic sensors on the freeway network which provide speed and flow information at segment boundaries. In its current form, the model will continue to operate if any of these sensors are broken, although admittedly, with less evidence available to refine the probability distributions, the estimated traffic state may be less accurate. While the model is robust in its ability to accommodate broken sensors, it is not able to determine if a sensor is malfunctioning. If such is the case, the model will try to replicate the incorrect data, deviating from real-world conditions. A feature to detect incorrect sensor data could certainly be added to the model for field applications, but this was not pursued as part of the current undertaking.

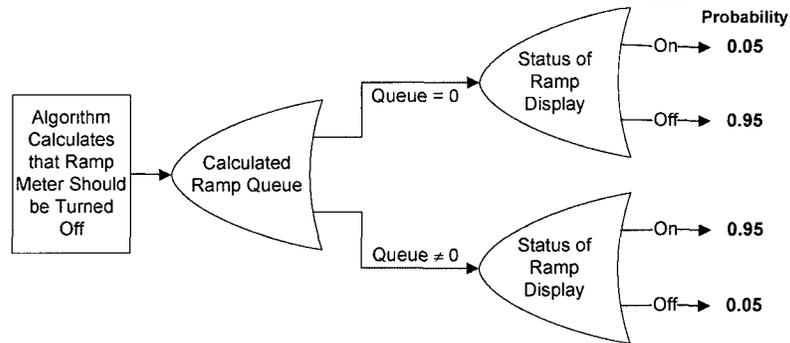
#### **j) Evidence used for belief updating**

The Freeway Traffic Model utilizes two main sources of evidence for tracking mainline operations: speed and flow data measured at strategically placed traffic sensors. While such sensors may also provide occupancy data, occupancy at a specific location is not necessarily a good indicator of freeway density over a wider section. For this reason, occupancy has not been used in the current version of the model, however, it could conceivably be used in future versions as an indicator of flow breakdown.

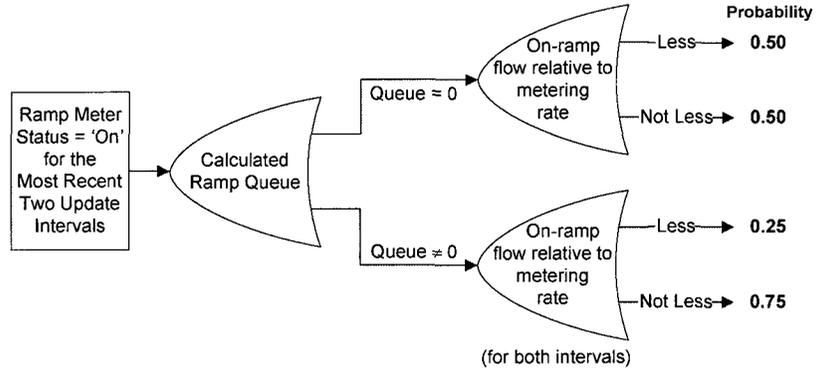
Evidence regarding on-ramp queues is also used in the Freeway Traffic Model. With just one sensor located back of the ramp meter (at the entrance to the ramp), only limited evidence is available for updating the probability distribution for the current ramp queue. Two situations are considered:

- **Evidence to assess a ramp queue of zero**

The first piece of evidence looks at the status of the ramp meter in the case where the algorithm recommends turning the meter off. Regardless of what the algorithm computes, the ramp controller is only able to turn off the ramp display if no vehicles are detected at the meter stop bar (otherwise, all vehicles queued at the meter would be released simultaneously, potentially creating a safety hazard and negatively impacting traffic flow). Thus, if the controller is unable to turn off the meter despite instructions to the contrary, there is good evidence that a ramp queue exists. In contrast, if the controller is able to execute the algorithm's instructions, there is good evidence that there is no ramp queue. The assumed probabilities are illustrated below.

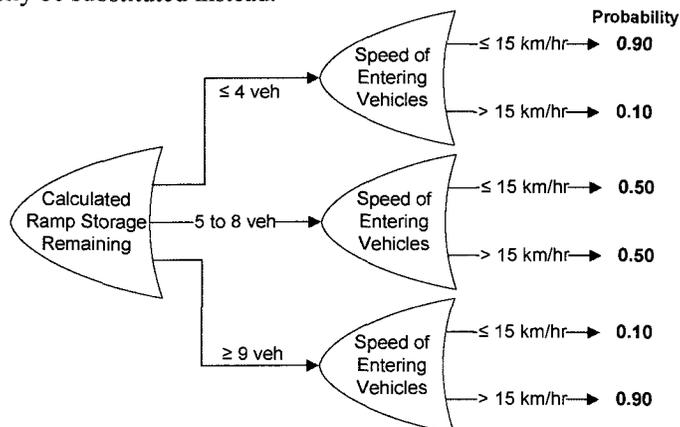


The second piece of evidence applies to the case where the ramp meter has been turned on for at least two model updates. In this situation, if the ramp flow onto the freeway is less than the ramp metering rate, it is reasonable to conclude that the ramp queue has more or less dissipated. Conversely, if a ramp queue exists, it is unlikely that the ramp flow will be less than the volume permitted onto the freeway. The assumed probabilities are shown below. In the case of no ramp queue being predicted, the demand may be less than the metering rate, or it could also be equal to the metering rate. As a result, the same probability is given to either event. Note that, given the very short time intervals involved (which may impact the measured flow rate), the conditions must be met in each of two successive intervals to be fully satisfied.



**Evidence to assess queue spillback onto the arterial network**

Queue spillback is assessed using data from the traffic sensor located at the entrance to the ramp. If the average speed falls below a certain threshold (in this case, 15 km/hr), it can roughly be assumed that the ramp queue has exceeded the available storage space. The corresponding probability assumptions are illustrated below. Note that for the traffic sensor data to be applied, at least one vehicle must have entered the ramp over the observation interval. Moreover, while a speed-based criterion for queue spillback was employed in the current version of the mode, an occupancy-based criterion could easily be substituted instead.



Using the probability distributions presented above, the performance of the Freeway Traffic Model was generally found to be acceptable for the test network under investigation (refer to Section 8.2). As a result, no further effort was spent fine-tuning the evidence assumptions. However, it is recommended that other options for incorporating new/existing evidence into the inference process be explored as part of future work to identify whether further improvements in model performance may be possible. Should a similar approach be adopted, the ramp queue evidence will need to be tailored to the specific corridor where the model is being applied.

#### **k) Estimating on-ramp flows**

In tracking mode, on-ramp flows are estimated directly from sensor data. In prediction mode, the flow onto the freeway is a function of the ramp demand, the ramp metering rate, and the number of ramp meter violations.

On-ramp flows are assumed to enter the freeway at the cell boundary downstream of the ramp. In most cases, this boundary is expected to be located near the end of the ramp speed change lane to capture the onset of flow breakdown. However, the traffic sensors used to estimate on-ramp flows are located on the ramp itself, near the start of the speed change lane. Initially, any time lag associated with travel in the speed change lane was ignored and it was assumed that all ramp vehicles detected during a given time step would enter the freeway and cross the mainline segment boundary within the same interval. Any errors in this assumption were handled by increasing the level of uncertainty associated with the on-ramp flow. While this approach was found to yield acceptable results in terms of freeway operations, ramp queue estimates would often deviate significantly from their observed value.

Accordingly, a new variable was introduced, representing the number of vehicles “stored” in the speed change lane who have not yet completed the lane change manoeuvre. In any given time step, the number of ramp vehicles actually entering the freeway is a function of the number of vehicles stored in the speed change lane at the end of the previous time step, the number of vehicles entering the speed change lane in the current time step, and the proportion of those same vehicles which are also able to exit the speed change lane during the current time step. The number of stored vehicles is calculated based on the average vehicle spacing, assuming a merge speed of 90 km/hr, relative to the effective merge distance (taken as 75% of the length of the speed change lane), and adjusted to include a random component. Because any errors in the speed change lane volume tends to be carried forward into subsequent time steps, the accuracy of the freeway predictions was found to degrade slightly under this new assumption, however, the ramp queue estimates were substantially improved (since the uncertainty term for the ramp flow observations could be reduced to more realistic levels).

The use of a speed change lane variable also addresses the situation where heavy freeway congestion generates queues in the speed change lane as vehicles wait for an opportunity to merge with mainline traffic. As long as the queue does not extend beyond the ramp sensor, in theory, the model should be able to handle it. [Previously, it was assumed that all of the ramp vehicles would be able to force their way onto the freeway during the current time step regardless of mainline queues. An alternative solution would be to apply a merge parameter which reflects the tendency of ramp and mainline vehicles to enter a congested merge in a relatively fixed ratio (Cassidy and Ahn 2005)].

#### **l) Estimating ramp demand under conditions of queue spillback**

When the vehicle queue on the ramp spills back beyond the ramp sensor, the sensor readings no longer provide an accurate estimate of the true ramp demand, making it difficult to track the length of the queue over time. To estimate the ramp demand under such conditions, previous sensor readings taken before the on-set of queue spillback are used to develop an average demand rate, which is assumed to persist as long as queue spillback is observed (unless the most recent observations suggest that a higher rate would be appropriate).

Within the model, queue spillback is defined as occurring whenever the speed at the ramp sensor drops below 15 km/hr, although in practice this value may be dependent on the position of the sensor relative to the start of the ramp. As an alternative, an occupancy-based criterion could also be used to indicate when ramp storage has been exceeded.

### m) Estimation of ramp queues

Within the Freeway Traffic Model, calculation of the ramp queue is relatively straightforward: in each time step where ramp metering is operational, the vehicles entering the ramp are added to any pre-existing queue, while the vehicles exiting the ramp (after passing the ramp meter) are subtracted. In tracking mode, both values are estimated from traffic sensors, implying that any ramp queue estimates should be relatively accurate. However, in practice, the estimated queue length was sometimes observed to deviate significantly from the actual queue length, hindering the performance of the ramp control algorithm.

With no evidence to assess the accuracy of the ramp queues, the particles selected in each time step are determined primarily based on their consistency with the observed speed and flow data for the freeway mainline. Even if a representative sample of particles is carried forward in each time step (which does not always occur for reasons discussed elsewhere), there is risk that the ramp queue estimates will deteriorate over time, particularly if the assumed sensor error used in developing the flow inputs is relatively large.

To improve the accuracy of the ramp queue estimates, the Freeway Traffic Model was updated to include limited ramp queue evidence for the two extreme conditions where the ramp queue goes to zero or spills back onto the arterial network (refer to the discussion above). In addition, adjustments were made to the model which would allow the on-ramp sensor error to be reduced (also described above). As an added measure, an over-ride mechanism was introduced to re-set the ramp queue under certain circumstances, when it is deemed probable that the estimated queue is wrong.

- If the ramp display is turned off in the simulation (indicating that no vehicles are detected at the ramp meter), but all of the particles tracking freeway performance suggest that a ramp queue exists, then, for all of the particles, the queue is re-set to zero.
- If the speed at the entrance to the ramp indicates that queue spillback has occurred, but all of the particles show space for 9 or more vehicles remaining on the ramp, then, for all of the particles, the queue is re-set such that between 5 and 8 vehicles of storage remains (as determined by sampling from the uniform distribution).
- If the speed at the entrance to the ramp indicates that queue spillback has not occurred, but all of the particles show space for fewer than 4 vehicles remaining on the ramp, then, for all of the particles, the queue is re-set such that between 5 and 8 vehicles of storage remains (as determined by sampling from the uniform distribution).

The latter adjustments are particularly useful for adjusting the queue estimate after a period of spillback has occurred, since any estimate of the queue length during spillback conditions is subject to considerable uncertainty given the lack of reliable sensor data available for assessing the ramp demand.

In a simulation environment, the traffic sensors are always 100% accurate, and as a result, setting the sensor error to zero will eliminate any queue prediction errors. However, in real-world networks, the sensor error will generally not be zero. Moreover, it is often necessary to assume some level of sensor error for sufficient particles to be selected during the inference process. While the procedures presented above were found to improve the accuracy of the ramp queue estimates, this remains a weakness of model which may be worth addressing in future work.

### n) Current limitations & opportunities for improvement

In its present form, the Freeway Traffic Model **only predicts flow breakdown at freeway merges**. If breakdown is also expected to be an issue at off-ramps or lane drops within the corridor, the model may need to be refined to better capture the breakdown phenomenon at these locations.

There is also **opportunity to expand the Freeway Traffic Model to predict collisions and traffic incidents**. Given the Bayesian formulation of the problem, it should be relatively straight-forward to incorporate a Bayesian-based incident detection module which detects freeway incidents given speed and flow evidence from traffic sensors. Even if no incident detection module is added, adjustments will still be needed to ensure the model is sensitive to the occurrence of flow breakdown due to freeway incidents wherever they may happen along the corridor.

In the current version of the model, **it is assumed that the ramp capacity is adequate to accommodate the**

**observed demand.** In the case of on-ramps, a parameter is provided for the ramp capacity, however, the capacity value has been set artificially high so that capacity constraints never come into effect. In theory, modifying this parameter to a more appropriate value would allow for more realistic modelling of on-ramp behaviour. In the case of off-ramps, the model currently has no mechanism to deal with queues which spread back onto the freeway mainline due to insufficient intersection capacity at the ramp-arterial junction. Should such behaviour be observed in reality, adjustments to the model will be needed.

Another potential future enhancement involves the relationship used to estimate the average travel speed in uncongested conditions as a function of the cell density. Currently, **the speed relationship does not account for the impact of merging and diverging behaviour.** However, turbulence at ramp junctions can influence vehicle speeds in the vicinity of the ramp. By adjusting the speed model to account for such impacts, model performance may be enhanced.

Other possible areas of improvement include:

- **Modelling of freeway shockwaves** – The current version of the Freeway Traffic Model employs a relatively simplistic approach to modelling traffic queues. Essentially, the model samples from the probability distribution for the queue density, which captures the range of conditions likely to be encountered in the queue as the density increases and decreases due to passing shockwaves. If evidence is available, the density estimates are refined to more closely match the observed conditions; if evidence is not available, the sampling process results in a cell density which is approximately equal to the average density within the queue. In either event, it is assumed that the density is relatively uniform throughout the freeway cell. In reality, this may not be the case depending on the shockwave position, impacting the flow of vehicles both into and out of the cell. The more effective ramp metering is at reducing freeway congestion, the less important it is to be able to accurately model shockwave behaviour. Nonetheless, development of an improved methodology for modelling traffic queues could potentially improve the overall performance of the Freeway Traffic Model for ramp metering and other applications.
- **Testing for freeway recovery** – To predict when freeway congestion has dissipated, the Freeway Traffic Model reassesses the probability of flow breakdown whenever the cell density approaches the critical density observed when congestion first appeared. This requires some judgement in establishing the density threshold which may impact the operation of the model. Too high and the model may fail to predict queue formation (particularly in prediction mode); too low and the model may not test for recovery as congestion dissipates. Accordingly, opportunities for improving on the recovery test may be worth exploring as the model is further refined.
- **Updating of the congested capacity (queue discharge flow)** – While the approach adopted for updating the congested capacity was shown to improve the accuracy of the Freeway Traffic Model considerably, other options may also exist which may perform better in certain circumstances.
- **Estimation of the initial freeway state** – Within the Freeway Traffic Model, the initial freeway state is represented by a probability distribution derived from traffic data. By sampling from this distribution, the initial state associated with each particle can be derived. In the current model, the same distribution is applied across the entire corridor, although distributions could also be developed for specific freeway cells using more localized data. Moving forward, options should be explored for improving the accuracy of the initial distribution and/or refining it through an initialization process.
- **Prediction of future demand** – The ability of the Freeway Traffic Model to predict future traffic operations depends in large part on the accuracy of the demand inputs. The current demand prediction module (described in Section 7.4.2) uses flow observations from the previous 2 minutes to predict future demand over the prediction horizon. It is envisioned that more sophisticated demand estimation techniques could be employed to further improve the accuracy of the model.

### 7.4.1 Prediction of Flow Breakdown

One of the key features of the proposed Freeway Traffic Model is the ability to predict flow breakdown at on-ramp merges. Within the Bayesian network, the probability of flow breakdown is estimated directly; whether or not breakdown occurs influences flow characteristics in upstream and downstream segments.

In operational analysis, it is common practice to define a single value for the “capacity” of a highway section. If demand exceeds capacity, traffic flow breaks down. In reality, flow breakdown is a random phenomenon which occurs over a range of traffic flow rates. Any model to predict flow breakdown must therefore capture the various factors which influence the probability that breakdown will occur.

A number of researchers have attempted to model the stochastic nature of the breakdown phenomenon. In the work by Davis et al. (1990), statistical pattern recognition techniques were used to forecast bottleneck formation over the next minute. In their review of field data, Elefteriadou et al. (1995) conclude that breakdown at ramp-freeway junctions is a probabilistic event which is function of the occurrence and size of on-ramp vehicle clusters. Based on this observation, the authors developed a probability of breakdown model by assigning probabilities to the vehicle cluster size, presence of freeway traffic within the merge area, and driver response to merging activity. Of particular significance is the finding that while high ramp and mainline flows may not be the direct cause of breakdown, they do increase the probability of breakdown by virtue of their influence on vehicle clusters. The authors also highlight one of the key challenges in developing probabilistic models – the significant data requirements.

Persaud et al. (1998) also allude to the extensive data needed to develop breakdown models based on real-world evidence. Using data for two major freeways in Toronto, the authors developed probability of breakdown curves as a function of the 1-minute flows in the median lane of the freeway (where trucks are prohibited). The probabilities were estimated based on a simple counting approach. While the research was only exploratory, the authors conclude that the probability of breakdown curves are not necessarily transferable between sites unless differences in the underlying causes of breakdown can be accounted for. In later research, Persaud et al. (2001) developed a logistic regression

model to estimate the probability of flow breakdown as a function of the average 1-minute flows on the subject freeway.

Ozguven and Ozbay (2008) also use empirical data to develop ‘capacity distribution curves’ which give the probability of breakdown at different freeway flow rates. To develop the curves, the authors applied survival analysis, using nonparametric Bayesian estimation to estimate the survival function.

Taking a different approach, Evans et al. (2001) calculated the probability distribution for the time of breakdown by examining the probability of vehicle movements through different zones within the merge area and applying Markov Chains. Son et al. (2004) attempt to explain the breakdown phenomenon in terms of the random propagation of traffic disturbances, and apply a wave propagation model to derive the probability of breakdown on a simple one-lane freeway. Similar attempts have been made to apply other models from the physical sciences to explain traffic breakdown, and there is a significant body of research devoted to this area (see for example Kerner and Klenov (2006) or Kuhne et al. (2002) which use nucleation models to describe the breakdown phenomenon).

### ***Development of Breakdown Model – Data Generation & Assessment***

The development of a flow breakdown model for the new ramp metering algorithm draws on the work of Persaud et al. (2001) described above. Using data for the QEW freeway in Toronto, Persaud et al. developed a “probability-of-breakdown” function which gives the probability of flow breakdown as a function of the traffic volume immediately downstream of the ramp merge.

Using a similar technique, a probability-of-breakdown function was estimated for use in the Freeway Traffic Model. Since all testing and evaluation of the Freeway Traffic Model (and associated ramp metering algorithm) was conducted using a simulation environment, the probability-of-breakdown function was likewise developed based on simulation data. In developing the function, a simple VISSIM model was used, comprised of a generic three-lane freeway section with a single on-ramp. This configuration was selected for consistency with the larger VISSIM test network described

in Section 8.2. Other parameters (such as the ramp and mainline speed distribution, proportion of heavy vehicles, etc.) were similarly adopted. To assess the effect of the speed change lane on flow breakdown, various versions of the network were developed for speed change lanes varying in length from 100 m to 350 m.

To obtain the necessary data, a macro was written to carry out a series of simulation runs for different levels of ramp and mainline demand.<sup>4</sup> In some cases, congestion was observed, in others, free-flow conditions were maintained for the duration of the simulation, providing a rich dataset for assessing the conditions under which breakdown occurs. For each demand scenario, 15 one-hour simulations were conducted. To model fluctuations in demand, random variation was added to the demand inputs at 10 minute increments.

During the simulations, speed and flow data were collected at 20 second increments for developing the breakdown model. The data were collected at three locations:

- On the freeway ramp, ~25 m upstream of the ramp gore;
- On the freeway mainline, ~100 m upstream of the ramp gore; and
- On the freeway mainline, ~10 m downstream of the end of the speed change lane.

The ramp and mainline flow measurements taken upstream of the freeway merge were used to determine the flow characteristics contributing to breakdown conditions. The measurements taken near the end of the speed change lane were used to identify the onset of flow breakdown.<sup>5</sup>

Given the sensor arrangement described above, the flows used to develop the breakdown model are not collected where breakdown actually occurs. Instead, the flows contributing

---

<sup>4</sup> Ramp demand was varied between 300 and 1200 vph in 100 vph increments, while mainline demand was varied between 6000 and 8000 vph in 250 vph increments, resulting in a total of 90 unique combinations of ramp and mainline flow.

<sup>5</sup> By positioning the traffic sensors near the end of the speed change lane, any speed drops due to mainline congestion can be readily identified. Further downstream within the bottleneck, vehicles are accelerating, making the speed drop less apparent. Moreover, the effects of any breakdown are delayed due to the travel time involved, making it difficult to determine when queuing was initiated. On the other hand, if the sensors are located too far upstream within the speed change lane, flow breakdown may occur beyond the traffic sensors, again making it difficult to determine when queuing begins. Ideally, the traffic sensors would be located just downstream of the breakdown location. Since breakdown was observed to occur within the speed change lane, measurements collected at the end of speed change lane were considered to provide an appropriate basis for identifying the onset of congestion.

to breakdown are measured upstream of the ramp merge, primarily due to the difficulty in estimating exactly where breakdown will occur, and the need to distinguish between ramp and mainline data. Given the time required to travel from the measurement location to the location of flow breakdown, there is a risk that the flow measured during a given time step and the flow actually triggering breakdown may be slightly different. However, the impact of any discrepancy is likely to be small, since the final model uses an average of the previous two traffic flow measurements (i.e. 40 seconds worth of data) to predict the probability of breakdown occurrence.

In developing the model, it was assumed that breakdown has occurred whenever the speed of vehicles departing the merge dropped below 85 km/hr. If a drop in speed of more than 10% was observed in the preceding time step, this preceding time step was taken as the beginning of flow breakdown. Otherwise, the first time step with a speed of 85 km/hr or less was assumed to represent the start of mainline queuing. In terms of recovery, it was assumed that the freeway has returned to an uncongested state when the speed reaches 95 km/hr or higher.

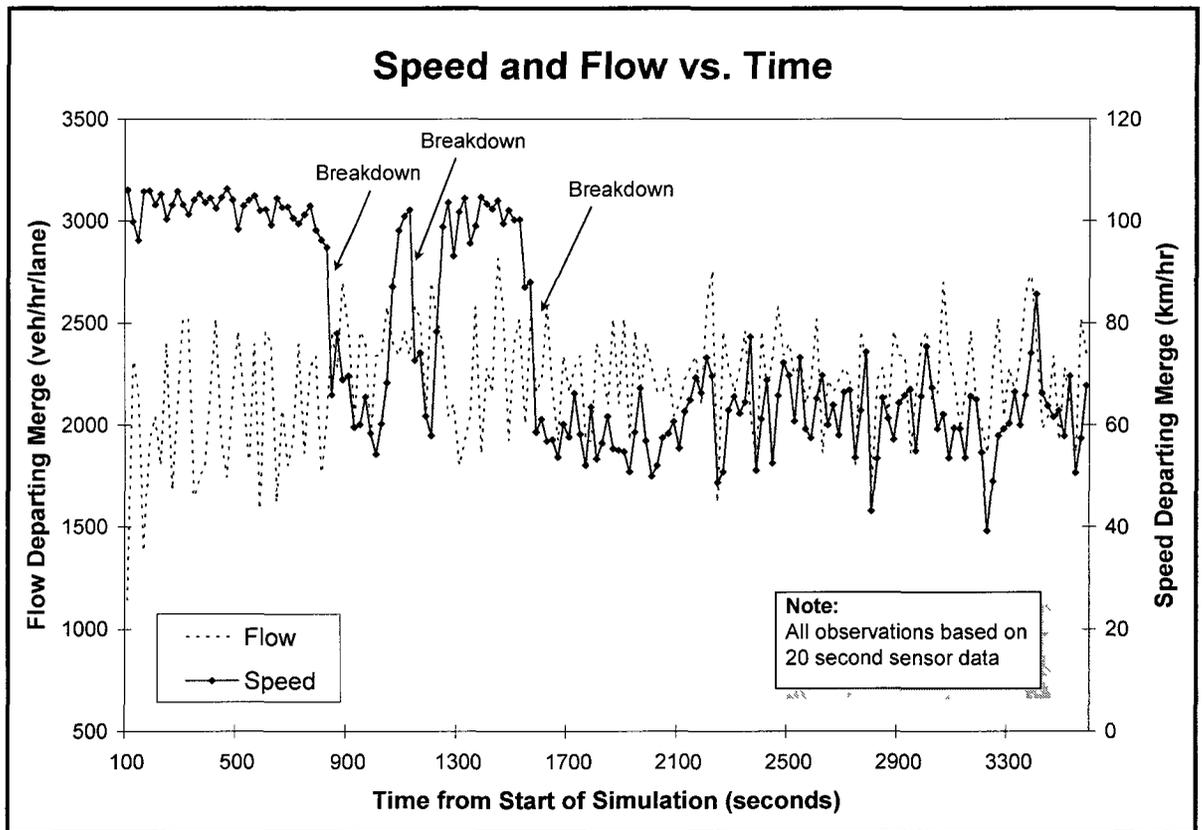
The probability of flow breakdown reflects the number of times a particular flow rate is observed and flow breaks down versus the total number of times the flow rate occurs.<sup>6</sup> Accordingly, to develop the prediction model, all flow measurements recorded when the freeway was in an uncongested state were flagged according to whether or not breakdown was initiated. All measurements recorded during freeway congestion were discarded from the analysis since it is meaningless to try to predict breakdown after it has already occurred, and any flow observations will be impacted by traffic queues.

Figure 7-6 illustrates a typical scenario run. The points of flow breakdown are highlighted. As this figure shows, some breakdown occurrences are only temporary; demand is reduced and the freeway quickly recovers. Although it is tempting to ignore these temporary breakdowns as simply the result of flow instability, the breakdown is real, and should be captured in the model. Of course, the model must also be able to capture the recovery process. Once breakdown has occurred, the probability of

---

<sup>6</sup> Note that this is not the same as the probability distribution for the breakdown flow. For example, at a flow rate of 2700 vphpl, the probability of flow breakdown might be 8%. However, of the flows causing breakdown, a flow rate of 2700 vphpl may represent 20% of the observations.

congestion during the next time step is affected. Within the traffic model, breakdown conditions are assumed to persist into the next time step if the density is above a certain critical value, indicating that a queue exists on the freeway mainline. If the density is less than this threshold value, the potential for flow breakdown is re-assessed based on the current demand, assuming the speed has recovered to a certain pre-specified value. In most cases, the probability of breakdown will be greater than 0%. As a result, in some cases, breakdown will occur, implying that the bottleneck is still active. In other cases, breakdown will not occur, implying that the freeway has successfully transitioned to an uncongested state.



**Figure 7-6 Typical Results from the VISSIM Test Network**

For each speed change lane configuration, upwards of 50,000 observations were used to develop the probability of breakdown function, based on a total of 1350 simulation runs (90 demand scenarios times 15 runs per scenario). As an example, for the 175 m speed change lane scenario, 83,283 valid flow observations were collected. Of these

observations, 4352 (or roughly 5%) were flagged as representing the on-set of flow breakdown.<sup>7</sup>

In developing the flow breakdown model, a decision was made to include a separate variable for ramp flow. This approach differs from that adopted by Persaud et al. (2001), which only considered the total flow in the merge section. This decision was made based on evidence cited by Banks (2000) and Elefteriadou et al. (1995), indicating that breakdown may be associated in some cases with the arrival of large clusters of vehicles from the ramp. Findings from Bertini and Malik (2004) likewise suggest that high on-ramp flows may trigger bottleneck activation.

In general, the vehicles involved in causing flow breakdown are the ones having just passed the breakdown location. To capture these vehicles, measurements covering a 40 second time period were incorporated in the model. Measurements over a shorter time period risk missing the vehicles involved in initiating breakdown; measurements over a longer time period risk masking flow variations which contribute to breakdown. For obvious reasons, it is important to identify the time of flow breakdown correctly.

Given the above, both ramp and mainline flow data were expressed in vehicles per 40 seconds for model development purposes. While these flow rates could have been converted to vehicles per hour, doing so would tend to obscure the observation interval actually used for predicting breakdown. Moreover, it is possible to observe short bursts of very high flow which could never be sustained over a full hour; expressing these flows on an equivalent hourly basis may be somewhat misleading. The use of 40 second data also fits well within the Bayesian framework used for estimating the breakdown probabilities.

A decision was also made to express the mainline flow data on a vehicle per lane basis. In general, it was thought to be easier to relate to traffic flows expressed this way, and also allows for easier comparison with ramp flows (which in this case are all from single lane entrances). Whether the flows are averaged over the number of lanes or expressed as an overall total, the resulting model should behave the same. It should be noted, however,

---

<sup>7</sup> Since measurements taken during congested flow are discarded, the number of valid observations tends to increase as the length of the speed change lane increases as these scenarios generally have less congestion. At the same time, the number of observations corresponding to flow breakdown decreases.

that the average flow per lane may be quite different from the actual lane flows depending on the lane utilization (something which is not considered in the Freeway Traffic Model). As a result, even though expressing the flow rates on a per lane basis makes it relatively easy to transfer the results to other freeway cross-sections, doing so may not be appropriate in all cases depending on the impact of cross-section (and the associated lane utilization) on breakdown occurrence.

### ***Development of Breakdown Model – Specifying Relationships***

Within the Bayesian network, the Breakdown Model is used to estimate the probability of flow breakdown as a function of the freeway geometry and ramp/mainline flow. Given this Bayesian context, it was considered appropriate to use Bayesian techniques to estimate the model parameters. We treat the learning problem as one of parameterizing a discrete binomial model: binomial since there are two possible outcomes (traffic flow has broken down, or it has not broken down) for each independent observation, and discrete since we assume that all variables can be discretized (in the case of traffic flow, this involves defining flow “bins” of arbitrary size).<sup>8</sup>

In developing the model, two approaches were explored. In the first, no local structure was assumed (i.e. the parameters were assumed to be independent across different instantiations of the parent variables), in the second, local structure was exploited by assuming that the relationships between variables could be represented by a logit formulation.

The discussion which follows provides a brief overview of each approach. The discussion is not intended to be theoretical, and no proofs are provided. For a more in-depth treatment of the subject, the reader is referred to Korb and Nicholson (2004), Neapolitan (2004), and Gelman et al. (2004), from which much of the following material was derived.

---

<sup>8</sup> A binomial distribution is a discrete probability distribution of the number of successes (i.e. flow breakdowns) in a sequence of  $n$  independent observations, where each observation yields success (i.e. breakdown) with probability  $\theta$ .

### Learning a Binomial Model with No Local Structure

In the Breakdown Model, assume that **BD** is a binomial variable with two possible values: Yes (indicating that flow breakdown has occurred) and No (indicating that flow breakdown has not occurred). To “learn” the model, we are interested in estimating the parameter value  $\Theta = \theta$ , where  $\theta$  is the probability  $P(\mathbf{BD} = \mathbf{Yes})$ .

According to Bayes’ Law, if we observe a breakdown occurrence:

$$P(\theta | \mathbf{Yes}) = \frac{P(\mathbf{Yes} | \theta)P(\theta)}{P(\mathbf{Yes})} = \beta P(\mathbf{Yes} | \theta)P(\theta)$$

Where  $\beta$  is the inverse of the probability of the evidence. Since  $P(\mathbf{Yes} | \theta) = \theta$ , the above equation simplifies to:

$$P(\theta | \mathbf{Yes}) = \beta \theta P(\theta)$$

Thus, we can estimate the value of  $\theta$  for a given instantiation of the parent variables (i.e. a given combination of ramp geometry and ramp/mainline flow) using evidence on the number of breakdown occurrences observed in the corresponding dataset. If, for a particular set of freeway conditions, the evidence  $\mathbf{e}$  consists of  $\mathbf{m}$  occurrences of flow breakdown out of a total dataset of  $\mathbf{n}$  observations, then our beliefs about the parameter  $\theta$  can be updated as follows:

$$P(\theta | \mathbf{e}) = \beta \theta^{\mathbf{m}} (1 - \theta)^{\mathbf{n} - \mathbf{m}} P(\theta)$$

The above relationship is based on the assumption that all breakdown occurrences are independently identically distributed (i.e. each observation is independent, and each observation is drawn from the same underlying probability distribution).

To apply the above equation in practice, an assumption must be made for the prior distribution over  $\Theta$ . Most commonly, the prior is restricted to the family of beta distributions. It can be shown that if  $\Theta$  has a beta distribution with hyperparameters  $\alpha_1$  and  $\alpha_2$ , that is:

$$\rho(\theta) = \mathbf{beta}(\theta; \alpha_1, \alpha_2)$$

[where  $\rho(\theta)$  is the density function of  $\Theta$ ]

then:

$$\rho(\theta | \mathbf{e}) = \text{beta}(\theta; \alpha_1 + \mathbf{m}, \alpha_2 + (\mathbf{n} - \mathbf{m})) \quad [\text{where } \rho(\theta | \mathbf{e}) \text{ is the density function of } \Theta \text{ conditional on the evidence } \mathbf{e}]$$

Thus, the posterior estimate of the model parameter  $\Theta$  also follows a beta distribution, but with modified hyperparameters.<sup>9</sup>

The expected value  $E(\Theta)$  represents our estimate of the parameter  $\theta$  for the next observation (i.e. our estimate of the probability of flow breakdown). For a general beta distribution  $\mathbf{B}$ , the expected value of the distribution can be calculated as a function of the hyperparameters as follows:

$$E_{\mathbf{B}(a,b)} = \frac{\mathbf{a}}{\mathbf{a} + \mathbf{b}}$$

where  $\mathbf{a}$  and  $\mathbf{b}$  are the hyperparameters. Thus, the posterior estimate of  $\theta$  given the observed evidence  $\mathbf{e}$  can be calculated as:

$$P(\text{BD} = \text{Yes} | \mathbf{e}) = E(\Theta | \mathbf{e}) = \frac{\alpha_1 + \mathbf{m}}{\alpha_1 + \alpha_2 + \mathbf{n}}$$

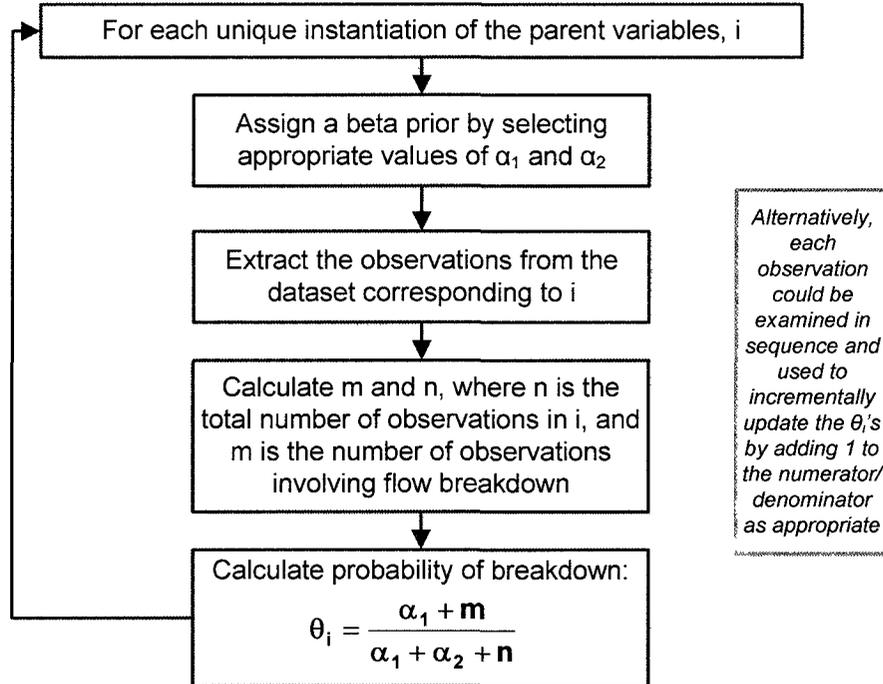
Given the above formulation, estimation of the parameter value simplifies to a straightforward counting exercise. For a given dataset, it is merely necessary to calculate the number of breakdown occurrences  $\mathbf{m}$  within the total set of  $\mathbf{n}$  observations. Values for  $\alpha_1$  and  $\alpha_2$  are selected to represent our initial beliefs concerning  $\theta$ .

In representing prior beliefs, it is sometimes helpful to think of  $\alpha_1$  and  $\alpha_2$  as ‘pseudo-counts’ which are added to the observed counts to reflect prior knowledge (Needham et al. 2007). Given this interpretation,  $\alpha_1$  represents the number of successes (i.e. flow breakdowns) that we would expect to occur (based on our prior knowledge) if we were to make  $(\alpha_1 + \alpha_2)$  observations. For this reason,  $\alpha_1 + \alpha_2$  is known as the equivalent sample size. A large equivalent sample size implies strong prior beliefs, since a greater number of observations is needed to significantly alter the prior value. If all values in  $[0,1]$  are considered to be equally likely (i.e. we have no knowledge of what the probability might be, or do not wish to impose our beliefs on the learning algorithm in the interest of objectivity), then  $\alpha_1 = \alpha_2 = 1$ .

---

<sup>9</sup> A family of distributions which remains within the same family after Bayesian updating (albeit with different hyperparameters) is referred to as a conjugate family of distributions.

To develop the conditional probability tables, the counting algorithm must be applied for each unique instantiation of the parent variables (i.e. for each combination of ramp geometry, and ramp/mainline flow). Figure 7-7 illustrates the process for a typical dataset comprised of various values of the parent variables. This is essentially the algorithm implemented in the software Netica referenced below, although a slightly different terminology is used.



**Figure 7-7 Counting Algorithm for Probability Updating**

The counting solution presented above can be expanded to multinomial problems by assuming a Dirichlet distribution for the parameter values. The resulting algorithm is widely used and has been implemented in many Bayesian software packages for learning probabilities (Korb and Nicholson 2004). Various techniques for dealing with incomplete data (noisy observations, hidden variables) have also been developed such as Gibbs sampling, and expectation-maximization.

The conditional probability table for flow breakdown was initially estimated using the software Netica. Figure 7-8 presents a snapshot of the Bayesian network developed in Netica based on the observed simulation data. Note that only one ramp geometry was considered during the initial data exploration / model development phase – the ramp with

the 175m speed change lane. Unless otherwise noted, all results which follow correspond to this scenario.

Since  $\Theta$  is represented by a probability density function, it is possible to compute not only the expected value of  $\Theta$  (which represents the probability of flow breakdown  $\theta$ ), but also the 95% probability interval for  $\Theta$ , which provides an indication of how likely it is that the true probability of breakdown is close to the expected value. To compute the 95% probability interval, the calculations were implemented in MATLAB and illustrated graphically in Excel. The resulting probability of breakdown parameters were identical to those estimated in Netica, but with the addition of 95% probability intervals for each parameter value. As shown in Figure 7-9, the probability of breakdown tends to increase with increasing ramp and mainline flow. Moreover, as the probability of breakdown increases, so does the corresponding 95% probability interval, implying lower confidence in the estimated value.

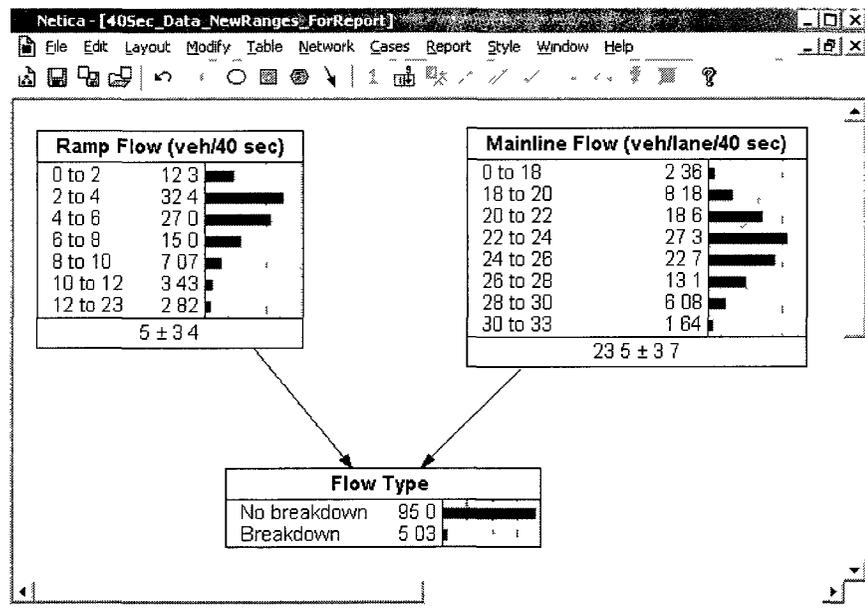
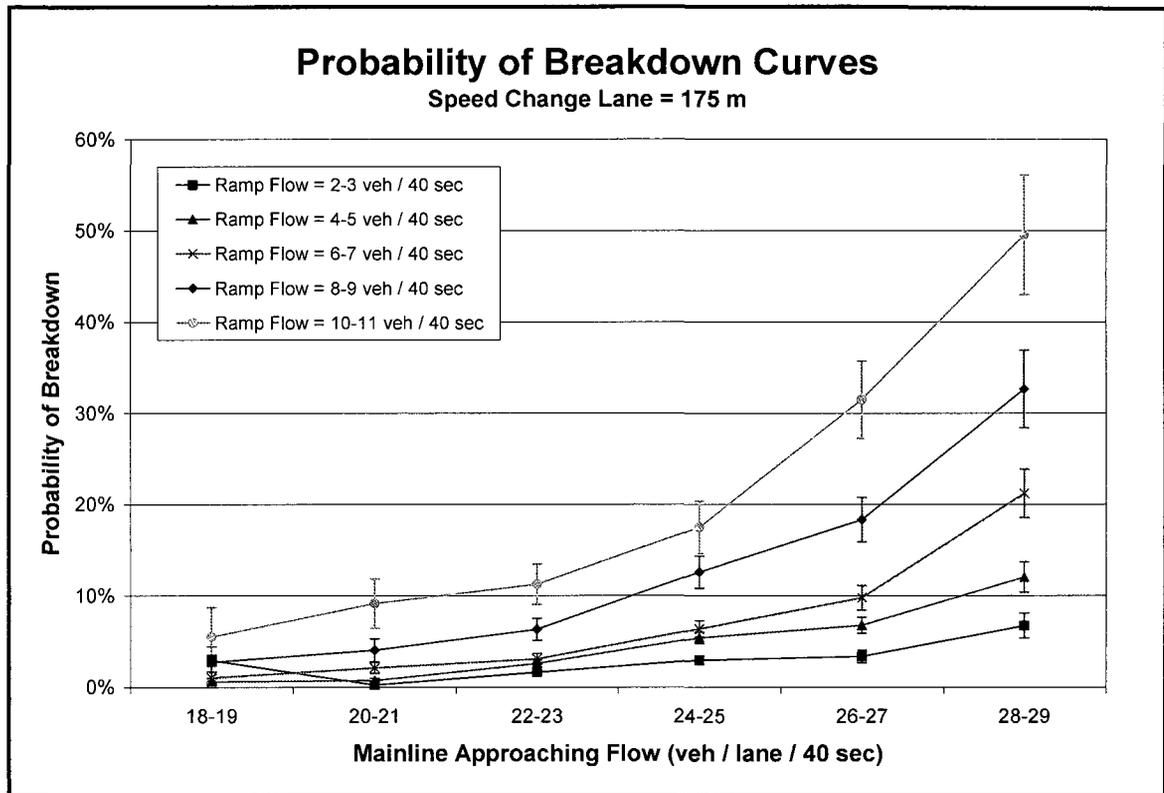


Figure 7-8 Flow Breakdown As Modelled in Netica



**Figure 7-9 Probability of Breakdown Assuming Independent Parameters**

While Figure 7-9 is based on 40 second flow observations, parameter values were also developed based on 20 second observations. In general, the 40 second data was found to provide a better basis for estimating flow breakdown, and as a result, the 20 second data was not considered further.

In developing the parameter values for Figure 7-9, different bin sizes were explored for discretizing the ramp and mainline flows. Initially, bins were developed based on 1 vehicle increments, however, the larger bin sizes shown in Figure 7-9 were found to produce smoother parameter curves. With larger bins, the sample size for any given parameter value increases, smoothing out anomalies. The reverse occurs as the bin size decreases. Thus, with smaller bins, an increase in ramp or mainline flow was sometimes found to result in a lower probability of flow breakdown or vice versa – unrealistic behaviour that could adversely impact the ramp metering algorithm.

While increasing the bin size may produce more stable parameter trends, the bin size cannot be increased indefinitely. With larger bins, it becomes more difficult to distinguish

the impact of different flow rates; if the flow rates fall within the same bin, the same probability of breakdown will be assigned. From a ramp metering perspective, this implies that two distinctly different metering rates could result in the same probability of breakdown, making it difficult for the ramp metering algorithm to determine which metering rate is preferred.

From the above discussion, a balance is clearly needed: if the bin size is too small, the algorithm may attempt to reduce on-ramp flow and end up with a greater probability of flow breakdown (or vice versa) due to inconsistent parameter trends. Conversely, if the bin size is too large, the ability of the model to differentiate between similar scenarios is compromised. Fortunately, the above issues can be mitigated to a certain extent by modelling local structure.

### Learning a Binomial Model with Local Structure

Local structure exists when there are dependencies between the parameters that make up the conditional probability table relating the parent and child variables. In the case of flow breakdown, we would expect the parameter values (i.e. the probability of breakdown) to increase with increasing levels of ramp and mainline flow. At the same time, an increase in the length of the speed change lane should decrease the probability of breakdown since there is a greater distance for drivers to complete the lane change manoeuvre. Such relationships can be used to develop models which predict the probability of breakdown as a function of the parent variables.

As before, the number of breakdown occurrences for a given instantiation of the parent variables can be represented by a binomial distribution:

$$\mathbf{y}_i \mid \boldsymbol{\theta}_i \sim \text{Bin}(n_i, \theta_i)$$

where  $\mathbf{i}$  refers to a particular instantiation of the parent variables (i.e. a particular level of ramp and mainline flow),  $\mathbf{y}$  represents the number of times breakdown occurs in a sample size of  $\mathbf{n}$  observations, and  $\boldsymbol{\theta}$  represents the probability of breakdown for each observation given the value of the parent variables. This formulation implies that the outcomes for a given group  $\mathbf{i}$  are exchangeable and independent with equal probabilities.

It is further assumed that the outcomes from different groups are independent of each other given the parameter values.

The simplest model of the relationship between  $\theta$  and the parent variables is linear:

$$\theta_i = \mathbf{a} + \mathbf{b}_1 \mathbf{x}_{1,i} + \mathbf{b}_2 \mathbf{x}_{2,i}$$

where  $\mathbf{a}$  and  $\mathbf{b}$  are the model coefficients to be estimated, and  $\mathbf{x}$  refers to the parent variables (i.e.  $\mathbf{x}_1$  = ramp flow and  $\mathbf{x}_2$  = mainline flow). To ensure that  $\theta$ , as a probability, lies between 0 and 1, a logistic transformation can be applied, giving:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \text{logit}(\theta_i) = \mathbf{a} + \mathbf{b}_1 \mathbf{x}_{1,i} + \mathbf{b}_2 \mathbf{x}_{2,i}$$

which has an inverse transformation of:

$$\theta_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \quad (\text{where } \eta_i = \mathbf{a} + \mathbf{b}_1 \mathbf{x}_{1,i} + \mathbf{b}_2 \mathbf{x}_{2,i})$$

The above equation is referred to as a logistic regression model. It falls within the category of Generalized Linear Models and can be fitted using maximum likelihood estimation (MLE) or Bayesian techniques. In the MLE approach, the parameters are assumed to be fixed, and are chosen to maximize the likelihood of the data without regard for any prior information that may be available. In the Bayesian approach, the parameters are treated as random variables; a prior distribution is specified for each parameter, which is then updated using information contained in the data through the process of Bayesian inference.

According to Bayes' Theorem, the posterior distribution of the model parameters can be estimated as:

$$\mathbf{P}(\lambda | \mathbf{y}) = \frac{\mathbf{P}(\mathbf{y} | \lambda) \mathbf{P}(\lambda)}{\mathbf{P}(\mathbf{y})}$$

where  $\mathbf{P}(\mathbf{y}|\lambda)$  is the likelihood of the observed data  $\mathbf{y}$  given a model with a set of parameters  $\lambda$  and  $\mathbf{P}(\lambda)$  is the prior probability of the parameter set before  $\mathbf{y}$  is observed. Thus, by Bayes' theorem, the joint posterior distribution of the model parameters is

proportional to the product of the likelihood and the priors. For the logit model developed above, the joint posterior distribution can therefore be expressed as:

$$p(\lambda | \mathbf{y}, \mathbf{n}, \mathbf{x}_1, \mathbf{x}_2) \propto p(\lambda) \prod_{i=1}^k p(\mathbf{y}_i | \lambda, \mathbf{n}_i, \mathbf{x}_{1,i}, \mathbf{x}_{2,i})$$

where  $\mathbf{x}$  represents the parent variables (i.e.  $\mathbf{x}_1$  = ramp flow,  $\mathbf{x}_2$  = mainline flow),  $k$  represents the number of unique instantiations of the parent variables (i.e. the number of unique combinations of ramp and mainline flow, also referred to as cases), and  $\lambda$  represents the model parameter set ( $\mathbf{a}$ ,  $\mathbf{b}_1$ ,  $\mathbf{b}_2$ ).

Furthermore, the likelihood of each case  $i$  in terms of the model parameters is given by:

$$p(\mathbf{y}_i | \lambda, \mathbf{n}_i, \mathbf{x}_{1,i}, \mathbf{x}_{2,i}) \propto [\text{logit}^{-1}(\eta_i)]^{y_i} [1 - \text{logit}^{-1}(\eta_i)]^{n_i - y_i}$$

$$\propto \left( \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{y_i} \left( \frac{1}{1 + e^{\eta_i}} \right)^{n_i - y_i} \quad (\text{where } \eta_i = \mathbf{a} + \mathbf{b}_1 \mathbf{x}_{1,i} + \mathbf{b}_2 \mathbf{x}_{2,i} \text{ as above})$$

To estimate the joint posterior distribution of the model parameters ( $\lambda | \mathbf{y}$ ), Markov Chain Monte Carlo (MCMC) methods are often used. Such methods are designed to generate a random sample from the posterior distribution, which can then be used to approximate the distribution or derive statistics such as the posterior mean or variance.

In general, MCMC techniques involve an iterative sampling process – at each step, the approximate distribution from which the sample is drawn becomes closer and closer to the target distribution ( $\lambda | \mathbf{y}$ ). The samples are drawn sequentially, and since the values sampled in one step depend only on the results of the previous step, a Markov chain is formed. The key is to structure the process so that the Markov chain converges to a unique stationary distribution which is the same as the posterior distribution of interest. Once the ensuing “chain” of parameter values has converged, it can be used to sample from the desired target distribution.

In the case of Gibbs Sampling, a popular MCMC technique, each unknown parameter is sampled sequentially from its (full) conditional distribution given each of the other model parameters (as most recently estimated) as well as the observed data. It can be shown that after many iterations, the samples thus generated correspond to the joint posterior

distribution of the unknown parameters. Any samples produced after convergence can then be used to derive summary statistics for the unknown parameters and make inferences about their true values. Additional information on Gibbs Sampling can be found in Appendix E (in the more general context of Bayesian inference).

To estimate the unknown parameters for the flow breakdown model, the software WinBUGS was used. WinBUGS provides a flexible, user-friendly tool for Bayesian analysis of complex models using MCMC methods (BUGS 2008; Lunn et al. 2000). The BUGS (Bayesian inference Using Gibbs Sampling) project was initiated in 1989 by the Medical Research Council's Biostatistics Unit in Cambridge, and has since evolved into a suite of products used by researchers around the world.

The WinBUGS model specification can found in Appendix H. In carrying out the WinBUGS analysis, independent 'noninformative' prior distributions were assumed for the unknown parameters values. Specifically, the priors were assumed to follow a normal distribution with a mean of 0 and a precision of  $1.0 \times 10^{-6}$  (where precision equals  $1/\text{variance}$ ). The analysis was conducted using 3 chains, each with different initial values for the stochastic variables. For each model developed, convergence was assessed after 15,000 simulation updates and deemed to be acceptable. In checking for convergence, several criteria were considered:

- **Chain stability:** In a practical sense, convergence requires that the parameter estimates stabilize around a set value.
- **Overlapping chains:** According to the WinBUGS User's Guide (Spiegelhalter et al. 2003), it is reasonable to conclude that convergence has been achieved if all the chains appear to overlap
- **Gelman-Rubin statistic:** The WinBUGS User's Guide (Spiegelhalter et al. 2003) describes the Gelman-Rubin statistic and its implementation within WinBUGS. In general, convergence is assumed to occur once the 'R' statistic has converged to 1, and the 'B' and 'W' statistics have converged to stability (represented by the red, green and blue curves in the WinBUGS bgr diagram, respectively).

After confirming that the convergence criteria were reasonably satisfied, an additional 25,000 updates were carried out to obtain a sufficient sample for posterior inference of the parameter values. In determining the appropriate number of iterations required after

convergence, the WinBUGS User's Guide (Spiegelhalter et al. 2003) offers the following advice:

*One way to assess the accuracy of the posterior estimates is by calculating the Monte Carlo error for each parameter. This is an estimate of the difference between the mean of the sampled values (which we are using as our estimate of the posterior mean for each parameter) and the true posterior mean. As a rule of thumb, the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample standard deviation.*

A sample consisting of 25,000 updates per chain was generally found to satisfy the above rule-of-thumb requirement.

Figure 7-10 illustrates the logit model that was developed from the WinBUGS analysis for the 175 m speed change lane scenario (i.e. the scenario used in the initial data exploration and model development phase).<sup>10</sup> As anticipated, the probability of flow breakdown increases with increasing ramp and mainline flow. The corresponding WinBUGS outputs for this scenario can be found in Appendix H. Appendix H also contains a chart which illustrates the model sensitivity to the posterior parameter distributions.

The logit model presented in Figure 7-10 is based on 40 second traffic observations, and does not include a term to capture interaction between the ramp and mainline flow. While this model formulation was selected as preferred, several other model forms were also explored in the initial model development phase.

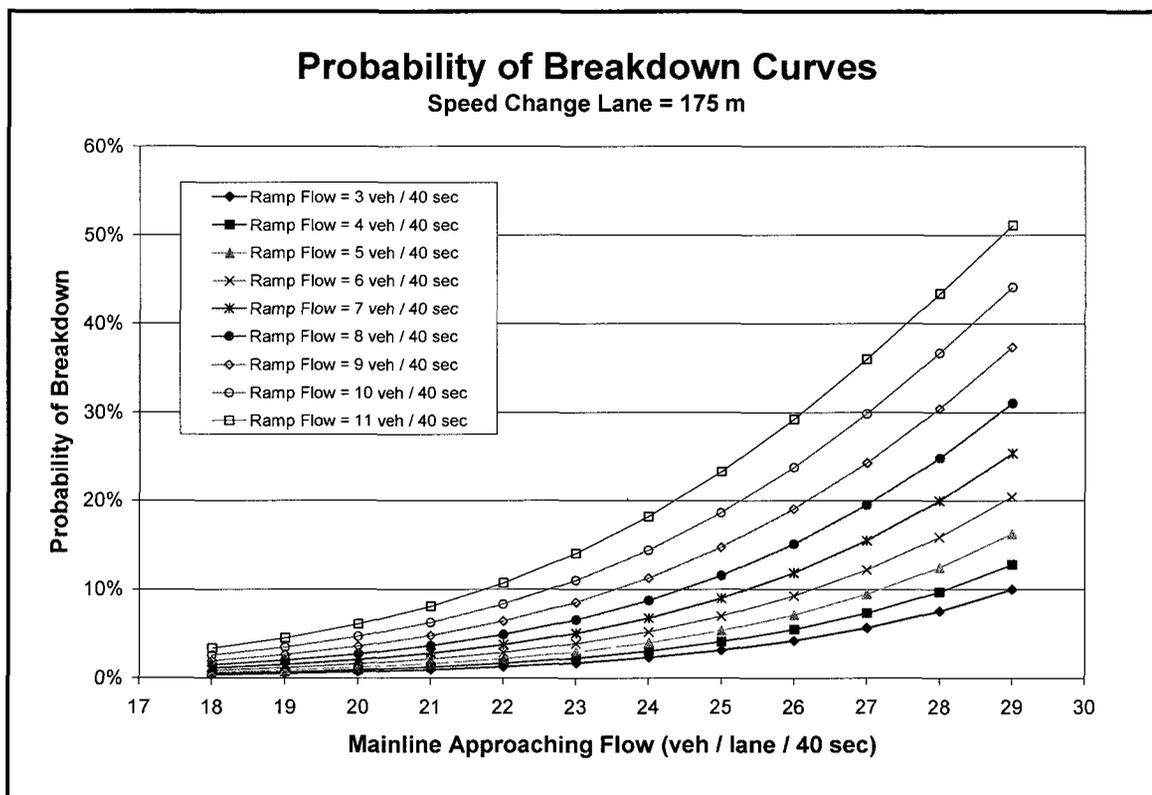
- Consideration was given to using a complimentary log-log model formulation, which, unlike the logit model, is asymmetrical. However, the logit model formulation was generally found to give superior results, as evidenced by the Deviance Information Criterion (DIC). [For any two models with the same observed data, the model with the smaller DIC is typically preferred – refer to Spiegelhalter et al. (2003) for more information].
- A version of the logit model was developed with an interaction term for ramp and mainline flow. While this model produced marginally better results (as evidenced

---

<sup>10</sup> For the 175 m speed change lane scenario, a sample of 83,283 observations was used for developing the breakdown model. Overall, the training dataset included 371 unique combinations of ramp and mainline flow (instantiations of the parent variables) as developed based on 40 second flow measurements expressed to the nearest whole vehicle.

by the DIC), the difference was considered negligible, and was not felt to justify the added model complexity.

- The impact of using 20 second flow observations was also explored. Accordingly, two logit models were developed: one based on 40 second data, and one based on 20 second data. While the DIC associated with the two models could not be compared directly (due to different observation datasets), the 40 second data was felt to be a better predictor of flow breakdown. This finding was confirmed using outputs from the SAS analysis described below.



**Figure 7-10 Probability of Breakdown Based on Logit Formulation**

To assess the reliability of the logit model illustrated in Figure 7-10, the observed number of flow breakdowns was compared against the model predictions. The comparison was conducted using a separate set of simulation observations developed specifically and exclusively for model validation purposes, comprised of 325 unique combinations of ramp and mainline flow for which a probability of breakdown could be computed (53,730 observations in total). In general, a reasonably good correlation was found between the observed and predicted data (correlation coefficient of 0.83 with the removal of 2

outliers), suggesting that the logit model is of sufficient accuracy for inclusion in the ramp metering algorithm.

Similar conclusions were drawn when comparing flow breakdown occurrence expressed as a probability, although in this situation, the comparison was restricted to ramp/mainline flow combinations having at least 10 observations in order to ensure a reasonable sample size for calculating probability values. For this restricted dataset, the correlation coefficient between the observed and predicted data was found to be in the order of 0.95 (excluding outliers). Diagrams illustrating the model validation results can be found in Appendix H, along with similar diagrams for the ‘training’ data used to develop the logit model.

For comparison purposes, the logit model parameters for the 175 m speed change lane scenario were also estimated using Maximum Likelihood Estimation. The analysis was carried out using SAS, a popular statistics program. Based on the SAS analysis, all of the model parameters are statistically significant at the 0.0001 significance level according to the Wald Chi-Squared test. Moreover, the global null hypothesis that all slope parameters are equal to zero can be rejected on the basis of the likelihood ratio and efficient score tests which assess the joint significance of the explanatory variables. In terms of model preference, the model fit statistics estimated by SAS (including the Akaike Information Criterion and the Schwarz Criterion) support the conclusions drawn from the WinBUGS analysis using the Deviance Information Criterion. The Generalized Coefficient of Determination can also be used to compare competing models, and in fact, provided strong support for basing the explanatory variables on 40 second flow observations as opposed to 20 second observations (for the logit model based on 40 second data, Nagelkerke’s R-Square value was found to be 0.24 compared to 0.08 for the 20-second model).

Table 7-3 provides a summary of the SAS parameter estimates, and the corresponding WinBUGS results, for the 175 m speed change lane scenario. The associated SAS output report can be found in Appendix H.

**Table 7-3 Comparison of SAS and WinBUGS Parameter Estimates**

<b>Parameter (175 m speed change lane model)</b>	<b>SAS – Maximum Likelihood Estimate</b>	<b>WinBUGS – Full Bayesian Approach</b>
a – intercept	-12.0511	-12.03
b <sub>1</sub> – ramp flow	0.2809	0.2809
b <sub>2</sub> – mainline flow	0.3105	0.3098

As Table 7-3 shows, the model parameters estimated using SAS are nearly identical to those developed in WinBUGS, demonstrating the validity of the two approaches (and confirming that both were applied correctly).

While Maximum Likelihood Estimation was found to produce very similar results to the full Bayesian approach, there are good reasons for adopting the Bayesian approach in this research. Not only does it fit with the Bayesian framework used in the ramp metering algorithm, but more importantly, it provides a robust methodology for combining real-world observations with simulation data. This latter capability is expected to be important in the development of real-world models, where it may be difficult or even impossible to obtain a sufficient sample of traffic sensor data for every possible situation that could conceivably arise.

From the results of the WinBUGS analysis, the advantages of modelling local structure are clear. By explicitly capturing parameter dependencies across different instantiations of the parent variables, there is no risk of instabilities or erratic trends; the resulting parameter curves are logical and smooth. Moreover, there is no need to group observations into larger bins, making it easier for the ramp metering algorithm to differentiate the impact of similar but distinct traffic flows. A comparison of the two approaches to developing parameter estimates can be found in Appendix H. As shown, both approaches yield similar results, but the parameters based on the logit model formulation form smoother curves, and are generally easier to work with.

*Development of Breakdown Model – Recommended Model*

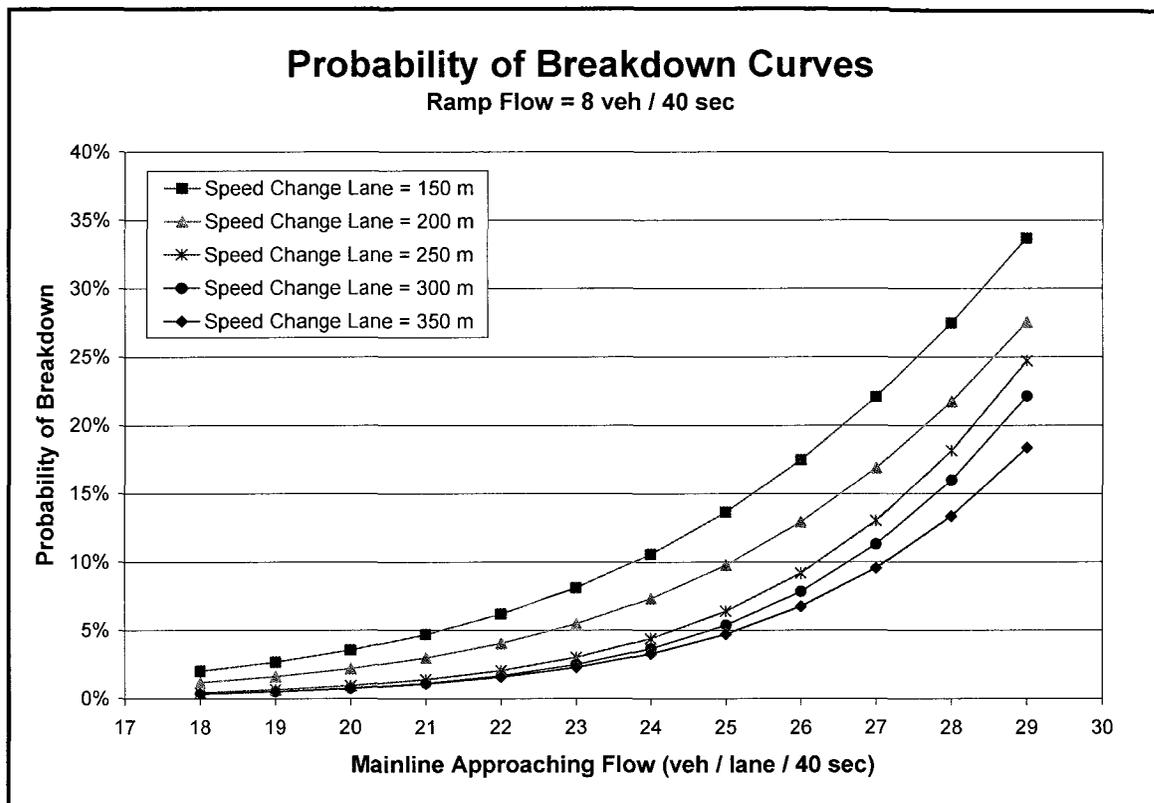
The recommended flow breakdown model is summarized in Table 7-4 and illustrated graphically in Figure 7-11. The model is based on a logit formulation as described above, where:

$$\Pr(\text{BD}) = \frac{e^{\eta}}{1 + e^{\eta}}, \text{ with } \eta = \mathbf{a} + \mathbf{b}_1(\text{RampVol}) + \mathbf{b}_2(\text{MainVol})$$

In the model, **a**, **b<sub>1</sub>**, and **b<sub>2</sub>** are calibration parameters. **RampVol** represents the ramp flow approaching the merge in vehicles per 40 seconds, while **MainVol** represents the mainline flow approaching the merge in vehicles per lane per 40 seconds. **Pr(BD)** is the resulting probability of breakdown to be incorporated in the Bayesian network of freeway traffic flow. Separate models were developed for speed change lanes varying in length between 100 m and 350 m. The model parameters were estimated using WinBUGS; more detailed outputs from the WinBUGS analysis can be found in Appendix H.

**Table 7-4 Flow Breakdown Model Parameters**

Geometric Configuration	Parameter Estimate		
	a	b <sub>1</sub>	b <sub>2</sub>
100 m Speed Change Lane	-8.742	0.3936	0.2258
150 m Speed Change Lane	-11.27	0.2648	0.2922
175 m Speed Change Lane	-12.03	0.2809	0.3098
200 m Speed Change Lane	-12.47	0.2986	0.3143
250 m Speed Change Lane	-14.93	0.3002	0.3936
300 m Speed Change Lane	-15.29	0.2953	0.4024
350 m Speed Change Lane	-14.6	0.2649	0.3789



**Figure 7-11 Graphical Depiction of Breakdown Model for various Merging Lengths**

The flow breakdown curves presented in Figure 7-11 correspond to a three-lane freeway section with a single lane on-ramp. Other geometric configurations may have different breakdown relationships. As Figure 7-11 illustrates, the probability of breakdown tends to increase with increasing mainline flow. However, as the length of the speed change lane increases, the probability of breakdown declines; the greater the distance for changing lanes, the easier it is to do so without disrupting mainline flow. While Figure 7-11 is based on a single value for the on-ramp flow, the relationship between ramp flow and flow breakdown can be seen in Figure 7-10 presented previously for the 175 m speed change lane scenario. As discussed, the probability of breakdown tends to increase as the ramp flow increases.

From the results presented above, it can be seen that, even at high flow rates, the probability of breakdown is relatively low. For example, for the 3-lane freeway depicted in Figure 7-10, if the approaching flow rate is 27 vehicles per lane per 40 seconds, and the ramp flow is 7 vehicles per 40 seconds, then total flow through the merge area over

the 40 second period is in the order of 88 vehicles (or 2640 vphpl). Despite this comparatively high traffic flow rate, the probability of breakdown is estimated to be only 15%. In comparison, the *Highway Capacity Manual* (HCM) defines freeway capacity to be in the order of 2250 to 2400 passenger cars/hr/lane, depending on the free-flow speed (TRB 2000).<sup>11</sup> While freeway sections may be able to sustain substantially higher flow rates over a very short time (as evidenced by the probability of breakdown in the example above), such flow rates cannot be supported over a longer period, and flow eventually breaks down.

In working with probabilities, there is a tendency to conclude that if the probability of flow breakdown is 10% at any one time step, than breakdown is likely to occur within 10 time steps assuming flows remain relatively constant. However, this interpretation is incorrect. What happens at one time step is independent of any previous time steps. Using the popular coin toss example, if the probability of tossing heads is 50%, there is a 50% chance that the next coin toss will be heads, even if the previous 2 or 5 or 100 coin tosses were tails. While the outcome of any one event is not impacted by previous events, it is possible to estimate the probability of a certain sequence of outcomes over time. With many coin tosses, roughly 50% will be heads. Therefore, it is unlikely (but not impossible) that a series of 100 coin tosses will result in all tails. The probability of such a sequence occurring can be calculated as the product of the probability of each individual coin toss:  $0.5^{100} = 8 \times 10^{-31}$ .

In the case of traffic congestion, the probability of flow not breaking down over a given time interval is calculated in a similar way. For example, if there is a 10% probability of flow breakdown at a particular time step, there is a 90% probability that flow will not break down. If this probability remains constant over 10 time periods, the probability that flow will not break down during this interval is  $0.9^{10} = 35\%$ . Thus, the probability that flow will break down is  $100\% - 35\% = 65\%$ . Once breakdown occurs, the probability relationship no longer holds; with the formation of mainline queues, the probability of flow breakdown in the next time step is affected.

---

<sup>11</sup> The flow breakdown relationship was developed assuming 2% heavy vehicles. If the HCM capacity values were adjusted to account for heavy vehicles, the capacities would be slightly lower.

Within the Bayesian network model, the probability of breakdown function is applied at every time step for every particle representing the freeway state. Once the probability of flow breakdown has been estimated for a given particle, a sample is drawn from the uniform distribution to determine whether or not flow actually breaks down. Thus, if the traffic flows are such that the probability of breakdown is in the order of 5%, then roughly 5% of the samples will predict the onset of traffic congestion. Where evidence is available, the probability distribution is refined (i.e. by selectively choosing the samples which support the evidence).

Where no evidence is available, the model must be capable of predicting flow breakdown in the cases where it is likely to occur (i.e. when demand levels consistently remain above the uncongested capacity). As a result, the model must be able to capture the increasing probability of flow breakdown as demand levels remain high over successive time steps, while still recognizing that the probability of breakdown at any one time step is independent of what happened previously. Using the particle filter inference technique described in Section 7.5, this occurs naturally; at each time step, samples which have not yet experienced flow breakdown are re-sampled based on the appropriate probability distribution to determine whether breakdown occurs. For those samples where breakdown has already been initiated, breakdown is assumed to persist into the next time step if the freeway density is above a pre-specified level, otherwise the probability of breakdown is re-assessed. Without this latter step, all samples would be re-tested at each iteration, and the percentage of samples predicting freeway congestion would roughly equal the probability of breakdown (i.e. even with high demand, the probability would remain low over successive time steps). However, by only re-sampling those particles which have not yet broken down, over several time steps, more and more samples will predict freeway congestion, even if the probability of breakdown is small. In general, the process is analogous to sampling with partial replacement.

### ***Limitations & Future Work***

The advantages of working in a simulation environment are considerable: the modeller has full control of the network geometry and travel demand for testing any scenario of interest, and full knowledge of every aspect of network performance. However, for real-

world implementation, real-world data must be used to develop the breakdown relationship used in the Freeway Traffic Model. Doing so poses a number of challenges:

- In real-world networks, sensor positions may not be optimally located
- Data may be incomplete or subject to error (or impacted by other unknown factors such as weather, construction, special events, etc.)
- Significant effort may be required to determine the flow type at a particular location. In particular, there is a need to distinguish between queue spillback and flow breakdown which requires knowledge of traffic conditions downstream
- Insufficient data may be available to establish statistically valid trends

This latter issue is of particular concern. With ramp metering, the potential for flow breakdown must be evaluated across the entire corridor, even if a section has never experienced breakdown in the past. Elimination of mainline queuing at one location may simply transfer the problem downstream. As a result, all sections must be treated as a potential bottleneck. Collecting the data required to model each freeway section, however, may prove problematic in real-world networks. Flow breakdown is a relatively rare event. On any given freeway system, breakdown may occur only once or twice a day at a limited number of ramps. As a result, it may be difficult to obtain sufficient data to fully capture the range of traffic and network conditions which may lead to breakdown occurrence.

Notwithstanding the above, the development of a probability of breakdown function for real-world networks is considered to be viable given existing data collection systems. Through advances in data storage, it is now possible to obtain flow observations covering multiple years. The introduction of data standards has made it possible to combine data from different networks to support model development, as long the underlying operational and geometric conditions are reasonably similar. Where data gaps continue to exist, it may be possible to combine real-world data with simulated data using the Bayesian approach to model development described above, for example, using the simulated data to develop parameter estimates, which then become the “prior distributions” to be updated when working with real-world data.

Moving forward, the following future work is recommended:

- **Develop breakdown models for other freeway elements.** The probability of breakdown function developed for the ramp control algorithm only applies to traffic congestion originating at on-ramps.<sup>12</sup> As a result, additional models are needed to predict flow breakdown at off-ramps and lane drops if breakdown is expected to occur at these locations. In the case of off-ramps with no lane drop, breakdown is most likely to result from mainline lane changes as drivers position themselves in the correct lane to exit the freeway. Such lane changes may occur over an extended distance, leading to difficulties in predicting where and when breakdown will occur. Since breakdown tends to occur less frequently at off-ramps (unless a lane drop is involved), obtaining the data to develop the prediction model may prove challenging.
- **Determine the optimal position of the cell boundaries in the Freeway Traffic Model from a flow breakdown perspective.** As noted in Section 7.4, the potential for flow breakdown is evaluated at cell boundaries, which should therefore be located at the start of potential bottlenecks. The challenge lies in determining where exactly queuing begins. At on-ramps, does breakdown occur at the start of the merge, near the end of the merge, or somewhere in between?<sup>13</sup> At lane drops, how far upstream do drivers begin to merge out of the lane, and where are such lane changes likely to trigger breakdown? While slight discrepancies between the actual and assumed location of flow breakdown are unlikely to have a significant impact on the model results, major variations in the location of breakdown could impact the accuracy of the model. This is of particular concern if incidents occur some distance from the cell boundary, or if breakdown occurs randomly over an extended freeway segment. Options to address this situation include reducing the cell size, or enhancing the model to account for the uncertainty in the breakdown location.
- **Expand the breakdown model to cover a range of freeway cross-sections.** The current breakdown model corresponds to a 3-lane corridor. However, it is anticipated that different freeway cross-sections may result in different lane utilization which could influence the potential for congestion at on-ramp merges.
- **Explore the impact of weather and other network characteristics influencing flow breakdown** (such as lane width, roadside environment, heavy vehicle proportion, etc.).
- **Explore other types of models that could be used to predict flow breakdown.** While the flow breakdown model presented above was found to work well in the Freeway Traffic Model, there may be opportunity to improve upon the model, or develop an entirely new model which results in better performance or requires less data for model development.

---

<sup>12</sup> For the test network under investigation, only congestion originating at on-ramps is considered. There are no lane drops, and off-ramp flows are relatively minor.

<sup>13</sup> In the test network described in Section 8.2, the cell boundaries were positioned at the end of the speed change lane, consistent with the sensor location used to develop the flow breakdown model.

- **Determine how the flow breakdown model could be expanded to predict breakdown in the event of a traffic incident (such a collision or vehicle breakdown).** It is currently envisioned that this could be accomplished by modelling the incident as a lane drop, with an appropriate assumption for the number of mainline lanes remaining open.

#### 7.4.2 Demand Estimation

The performance of the Freeway Traffic Model is highly dependent on the accuracy of the model inputs. One of the key inputs is the traffic demand entering and exiting the freeway system on both the freeway mainline (at the limits of the study area), as well the on-and off-ramps.

In **tracking mode**, the demand is estimated directly from traffic sensors. In using this sensor data, adjustments are required to account for sensor measurement error, as well as the duration of the data collection interval in relation to the model update interval.

- As discussed in Section 7.4, sensor error can be expressed in different ways. In the case of the VISSIM test network described in Section 8.2, the sensor error was assumed to be Gaussian with zero mean. The standard deviation of the sensor error was varied to reflect typical activity levels on different freeway system elements: 75 vphpl for the mainline and 50 vphpl for the ramps. For off-ramps, the exiting flow is expressed as a percentage of the mainline flow; in this case, the Gaussian error was assumed to have a standard deviation of 1%.
- In addition to accounting for sensor error, it is also necessary to allocate the observed traffic flows between model time steps. In developing the VISSIM test network, it was assumed that sensor data would be collected at 20 second intervals. However, the model update interval was assumed to be 10 seconds, based on the criterion that no vehicle should be able to cross more than one cell boundary during a given time step. Accordingly, the measured 20-second demand must be divided between the two 10 second time steps that fall within the measurement interval.
  - For mainline flow, the proportion of traffic arriving during the first 10 second time step was estimated by sampling from a uniform distribution ranging between 0.4 and 0.6. (i.e. it was assumed that between 40% and 60% of the observed demand would arrive during the first time step, with the remaining demand arriving during the second time step).
  - For ramp flow, the distribution of the arriving vehicles was found to be more sporadic. On the freeway mainline, traffic volumes are substantially higher, and are spread over multiple lanes, resulting in a reasonably stable distribution

of arriving flows.<sup>14</sup> However, on on-ramps, traffic volumes are typically lower. As a result, the headway between arriving vehicles can vary considerably; sometimes vehicle arrivals will be relatively uniform, at other times, vehicles will arrive in groups. Arrival headways will also be affected if the traffic entering the ramp is controlled by an upstream signal.

To account for the variability in ramp arrivals, the proportion of traffic arriving during the first 10 second time step was estimated by sampling from a uniform distribution ranging between 0 and 1, subject to the constraint that the maximum demand during any one time-step cannot be greater than the theoretical maximum ramp demand (assumed to be roughly 8 vehicles per 10 seconds). This latter constraint was imposed to minimize the potential for unrealistically high ramp flows in one 10-second time step while the second 10-second time step has much lower demand. If this constraint cannot be satisfied, a 50-50 distribution is assumed.

Figure 7-12 describes the above process in mathematical terms as implemented in the ramp control algorithm. The relationships in Figure 7-12 assume a total of  $n$  time steps occurring within the data collection interval (with  $n = 2$  for the VISSIM test network).

- Since off-ramp flow is expressed as a percentage of mainline flow, it was assumed that this percentage would remain reasonably constant over the entire data collection interval (i.e. the same percentage was applied to all time steps).

---

<sup>14</sup> With higher flows, the average vehicle headway is reduced. Since the minimum headway is fixed, this implies a reduction in at least some of the headways that are greater than minimum, which in turn implies a reduction in the headway variability.

Let **dc** be the duration of the data collection interval, and **dt** be the duration of the model time step, such that  $n = dc / dt$  (i.e. the number of time steps occurring within the data collection interval).

If **ObsFlow** is the flow observed during the data collection interval **dc** (in vph) and **MaxRampFlow** is the theoretical maximum ramp flow (in vph), then the proportion of the **ObsFlow** that occurs in time steps **1** to **n-1** can be estimated by sampling from the uniform distribution **(a,b)**, where **a** and **b** are calculated as follows:

$$a_1 = \frac{1}{n} - \frac{1/n}{n-1}$$

$$b_1 = \frac{1}{n} + \frac{1/n}{n-1}$$

$$a_2 = \frac{1-b_2}{n-1}$$

$$b_2 = \left[ \text{MaxRampFlow} \frac{dt}{3600} \right] / \left[ \text{ObsFlow} \frac{dc}{3600} \right]$$

$$a = \max(a_1, a_2)$$

$$b = \min(b_1, b_2)$$

$$\text{if } b < \frac{1}{n}, a = b = \frac{1}{n}$$

The proportion of ramp flow in time step **n** is simply 1 minus the summation of the values calculated above for each of the other time steps within the data collection interval.

**Figure 7-12 Allocation of Ramp Observations to Time Steps**

In **prediction mode**, traffic sensor data is again used to estimate the traffic demand entering the freeway system. However, in this situation, the demand inputs will be subject to greater uncertainty since historical data is being used to forecast future conditions over a prediction horizon of several minutes, rather than simply estimating demand over the most recent data collection interval.

It is assumed that the average historical demand from some previous number of observation intervals **N** will be roughly equal to the future average demand over the prediction horizon, subject to some level of uncertainty. In general, **N** must be large enough that any random fluctuations in the demand are smoothed out, but small enough that emerging trends in the data are not obscured. For the VISSIM test network described in Section 8.2, a value of **N** = 6 was assumed to be appropriate. Thus, the future demand

was estimated based on the average demand from the previous six 20-second observation intervals (i.e. two minutes).

To reflect the uncertainty in the use of historical data to estimate future conditions, the demand was assumed to be normally distributed with a standard deviation of 100 vphpl for mainline flow and 50 vphpl for the on-ramps. For off-ramps, where the demand is expressed as a proportion of the mainline flow, a standard deviation of 1% was applied.

By sampling from the distributions described above, the average demand over the prediction horizon can be estimated. However, it is also necessary to account for fluctuation in demand from one time step to the next. To do so, a sampling process was carried out for each time step to estimate the variation (noise) to be added to the average demand. The noise terms were again assumed to be normally distributed with zero mean and a standard deviation as above. Figure 7-13 illustrates the full process for estimating future demand over the prediction horizon.

Let **ObsFlow** (vph) be the flow observed during observation interval **z**, **AvgDemand** be the average demand over the prediction horizon (vph), and **Demand** be the demand observed during a particular future time step **j**. Then:

$$\mathbf{AvgDemand} = \frac{\sum_{z=1}^{\mathbf{N}} \mathbf{ObsFlow}_z}{\mathbf{N}} + \eta_1$$

$$\mathbf{Demand}_j = \mathbf{AvgDemand} + \eta_2$$

where **N** is as defined previously and  $\eta_1$  and  $\eta_2$  are noise terms reflecting the uncertainty in average demand over the prediction horizon and fluctuation in demand from one time-step to the next, respectively. In the case of the VISSIM test network, the distributions for  $\eta_1$  and  $\eta_2$  are assumed to be the same.

**Figure 7-13 Estimation of Future Demand**

Note that while the use of higher demand estimation error was explored (and felt to be justified based on the VISSIM simulation data that was examined), a decision was made to reduce the demand variation in order to reduce the number of particles needed for

prediction, which in turn impacts the algorithm speed (refer to Section 7.7.3). As a result of this decision, it was sometimes found to be necessary to include a safety margin in the ramp storage length to reduce the potential for queue spillback in the event that the actual ramp demand is higher than predicted. Ideally, a more realistic error distribution would be applied; nonetheless, the above compromise was found to yield acceptable results, and did not appear to have a major impact on the performance of the ramp control algorithm as currently implemented.

### *Capturing Traffic Diversion*

Within the Freeway Traffic Model, traffic diversion is not modelled explicitly, but is captured in the traffic sensor data used for tracking freeway conditions over time. As a result, the effects of any diversion activity triggered by freeway controls will be reflected in the traffic estimates.

While real-time sensor data provides a means to **track** how the system has evolved in response to a particular control action, such information may not be appropriate for **predicting** future conditions under different control strategies – an essential element of the ramp metering algorithm. Vehicle routing may change significantly under different control scenarios; if such behaviour is not accurately captured in the predictive process, the recommended ramp metering rates may not be optimal.

In some cases, previously observed data may in fact provide a suitable basis for predicting future outcomes. For example, at off-ramps, the only control measures to influence diversion are variable message signs (VMS). As long as the message remains similar (or there is no VMS), past observations are likely to provide the best estimate of future behaviour. Of course, if the message is changed as part of the control action, the diversion rate may also change. A model could certainly be developed to predict this phenomenon, however, such a model would require real-world data on diversion behaviour which is not readily available.

As a result, for the purposes of this thesis, it was assumed that variable message signs are not included as part of the control action, and that the process described previously for predicting future flows is suitable for capturing any off-ramp diversion that may occur.

Accordingly, within the ramp control algorithm, exiting rates at off-ramps are based strictly on past observations subject to a random walk (i.e. the exiting ratio over the prediction horizon is assumed to equal the average ratio from the previous **N** observation intervals plus or minus some random amount). This approach is considered to be appropriate even for freeways having variable message signs (controlled separately), since the impact of any message changes will be reflected in the real-time data. The only risk occurs if the message is changed during the prediction horizon, however, this is likely to occur only rarely. Moreover, since the algorithm continuously re-evaluates the most appropriate action at each control interval, any changes in diversion that do occur as a result of external changes to the VMS message will be quickly captured in the observed data and taken into account by the control algorithm.

Diversion at on-ramps is slightly different. While the introduction of ramp metering may cause some drivers to divert to the arterial network to avoid ramp delays (or divert to the freeway to take advantage of improved operating conditions), over time, traffic patterns are expected to stabilize as a new equilibrium is reached. However, drivers may still respond to real-time conditions, particularly if those conditions deviate significantly from normal, and information is available to advise motorists accordingly.

Within the ramp control algorithm, it is assumed that the ramp demand during the prediction horizon can be approximated from the demand observed during the previous **N** data collection intervals, subject to a random walk. While this approach ignores the relationship between ramp control and ramp usage, it is considered to be appropriate for the initial phase of algorithm development. The limitations of the proposed approach will be mitigated to a certain extent if it is assumed that the length of the ramp queue remains relatively stable over the peak period. Under such conditions, the overall change in diversion patterns over the prediction horizon is expected to be relatively small; as changes do occur, they will be reflected in the sensor data and incorporated into the control algorithm as ramp metering rates are continuously adjusted based on new data. The major risk with the approach lies in:

- Over-predicting the ramp demand when the ramp queue is increasing by failing to account for any associated traffic diversion away from the freeway system. This

will tend to cause the algorithm to over-estimate the ramp delay, which may in turn cause it to implement less aggressive metering rates than optimal.

- Under-predicting the ramp demand when the ramp queue is decreasing by failing to account for associated diversion to the freeway system. This will tend to cause the algorithm to under-estimate the ramp delay, which may in turn cause it to implement more aggressive metering rates than optimal.

The appropriateness of basing future traffic flow patterns on past behaviour was confirmed by Zhang and Levinson (2004a). In their investigation of off-ramp exit percentages, it was found that information from the current control interval could be used to predict exit percentages in the next control interval with a relative error of generally less than 10%. This stability in exiting behaviour can be attributed to the “complex dependencies of off-ramp exit percentages on metering rates at upstream ramps, together with slow-changing demand patterns and metering rates” (Zhang and Levinson 2004a, pg. 878). Obviously, the longer the prediction horizon, the less accurate such an approach is likely to be. It is also less clear whether such an approach is appropriate for on-ramps, where diversion patterns may be more susceptible to changes in the ramp control. Nevertheless, the approach is considered appropriate for an initial version of the new ramp control algorithm, recognizing that future enhancements may lead to improved performance.

### ***Estimating Ramp Demand in the Event of Queue Spill-Back***

A key challenge in working with sensor data is estimating the traffic demand when the queue spills back beyond the traffic sensor. This is a particular issue for on-ramps, since the sensor at the ramp entrance will not provide an accurate indication of the entering demand if the ramp queue exceeds the sensor position and spills back onto the arterial network. Without an accurate picture of the ramp demand, it is impossible to estimate the corresponding ramp queue, impacting the performance of the ramp control algorithm.

Certainly, additional traffic sensors could be implemented to track queue spillback on the arterial network (from which the corresponding ramp demand could be estimated).

However, in the interest of minimizing sensor costs, a simple methodology was

developed for estimating the ramp demand in the event of queue spillback. This methodology was briefly described in Table 7-2 in Section 7.4.

In essence, the model stores the sensor readings from the most recent observation intervals, and uses these historical values to compute a moving average. If queue spillback is detected, this moving average is used to estimate the ramp demand instead of relying on the observed data. To reflect the uncertainty in the estimated value (which is based on historical observations), an error term is applied. For the VISSIM test network described in Section 8.2, the error was assumed to be Gaussian with zero mean and a standard deviation of 100 vphpl. It was further assumed that traffic volumes over a 2-minute interval would provide a suitable basis for estimating the average ramp demand. Thus, with 20-second observation intervals, a total of six observations were used for computing the moving average. Once spillback has been detected, the average value is not updated again until spillback has dissipated, unless a higher average value would result based on the most recent sensor readings.

Since ramp demand is likely to change over time, the above methodology is not appropriate if queue spillback is considered acceptable, and allowed to continue over an extended duration. However, if the control algorithm is set to eliminate any queue spillback as soon as it occurs, the duration of queue spillback should be relatively short (as long as the ramp demand is less than the maximum metering rate – i.e. 900 vph for a single lane ramp). Under such conditions, the ramp demand is less likely to change significantly, and the above methodology may work reasonably well, particularly if the Freeway Traffic Model employs evidence on the existence of queue spillback to update the estimated queue length as described in Section 7.4.

In the current version of the algorithm, speed observations at the ramp entrance are used to detect the presence of queue spillback, although an occupancy-based criterion could certainly be used instead. While a speed threshold of 15 km/hr was found to provide an adequate basis for identifying when ramp storage has been exceeded, in practice, this value is likely to depend on the position of the ramp sensor relative to the start of the ramp. As a caution, an entering speed of zero does not necessarily mean that queue spillback has occurred, only that no vehicles were detected during the observation

interval. In such cases, observations from the previous interval are also considered in assessing the occurrence of spillback conditions.

### ***Potential Future Enhancements***

While the approach to demand estimation described above was found to produce reasonable results, more sophisticated demand estimation techniques exist which could potentially improve the performance of the algorithm significantly.

Vlahogianni et al. (2004) provide a summary of the numerous methods that have been developed for short-term traffic forecasting, ranging from time series models to neural networks. In just one example, Stephanedes and Kwon (1993) combine a logit-based behavioural model with an extended Kalman filter to predict on-ramp demand in real-time as a function of the traffic control. It is envisioned that something similar could be readily incorporated in the Bayesian framework of the control algorithm.

In addition to improving the approach to demand estimation, there is also opportunity to make use of sensor measurements on the arterial network; with better knowledge of upstream traffic flows, more accurate predictions of ramp demand can be achieved.

## **7.5 Inference Using Particle Filters**

Inference in dynamic Bayesian networks refers to the process of updating beliefs given new information. For a Bayesian network representing freeway traffic flow, inference involves using real-time sensor data to update probabilities about the freeway state.

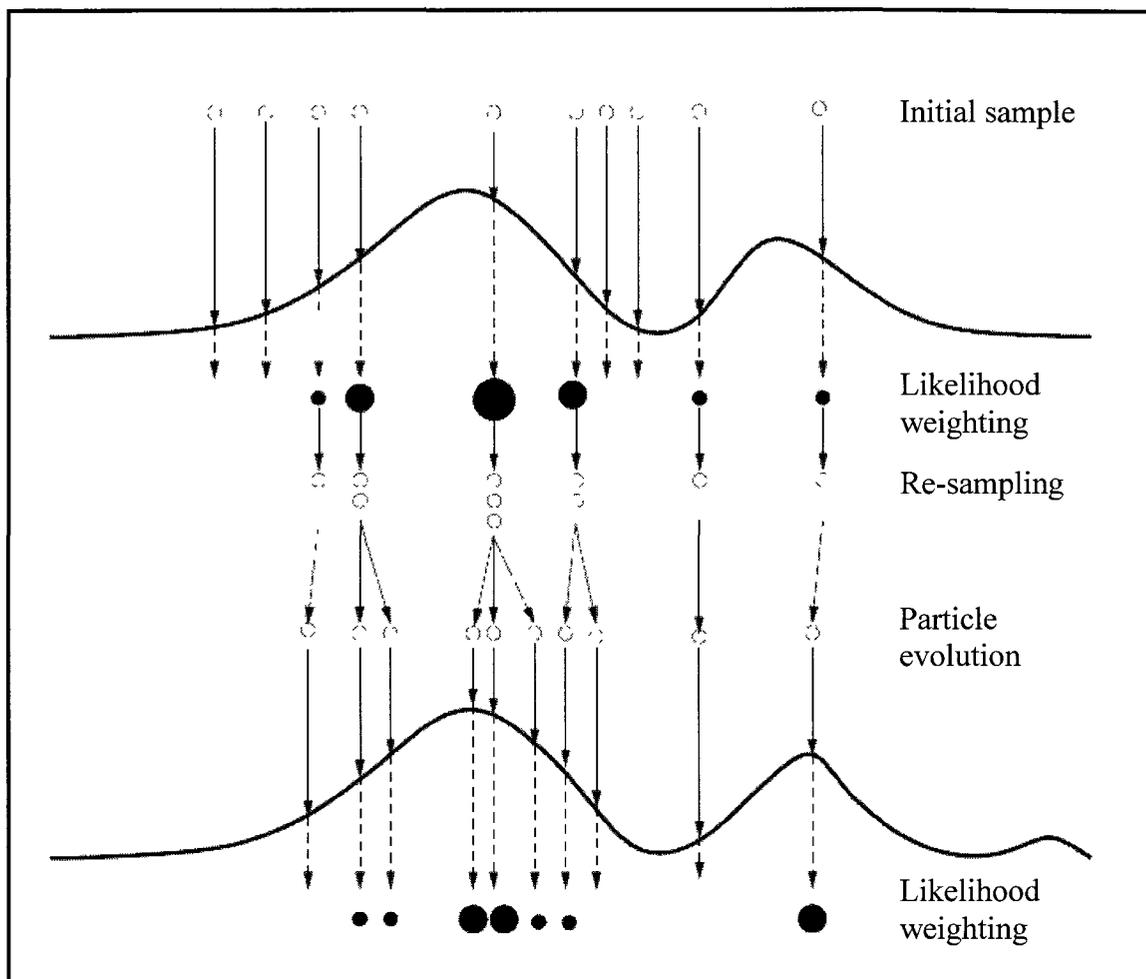
While several exact inference techniques have been developed, approximate inference methods are generally needed for dynamic Bayesian networks, particularly if the network is large or complex. A number of approximate inference techniques were reviewed to assess their appropriateness for the control algorithm. From this review, particle filters were identified as the most promising.

Particle filters approximate the belief state using a set of weighted samples (or “particles”):

$$P(\mathbf{X}_t | \mathbf{y}_{1:t}) \approx \sum_{i=1}^N \mathbf{w}_t^i \delta(\mathbf{X}_t, \mathbf{X}_t^i)$$

where  $\mathbf{X}_t^i$  is the  $i^{\text{th}}$  sample of state  $\mathbf{X}$  at time  $t$ ,  $\mathbf{y}$  is the evidence,  $\mathbf{w}$  is the sample weight and  $\delta$  is the Dirac delta function.

A graphical representation of a standard particle filter is provided in Figure 7-14. In the prediction step, the particles are subjected to a system model, and evolve to a new state. In the update step, the weight of each particle is computed based on the likelihood of the evidence. Over several successive iterations, many of the sample weights become negligible. To prevent such degeneracy, a re-sampling step is introduced. Particles with low weight are discarded, while particles with high weight are selected multiple times, forming a new sample which approximates the original weighted sample. Since the particle weights are represented by the selection frequency, the weights of the re-sampled particles are re-initialized. At each time step, the weighted particles represent the posterior probability distribution for the variables of interest, from which various statistics can be computed (such as the sample mean). Once re-sampling is carried out, this posterior distribution becomes the prior distribution for the next time step, and the process is repeated.



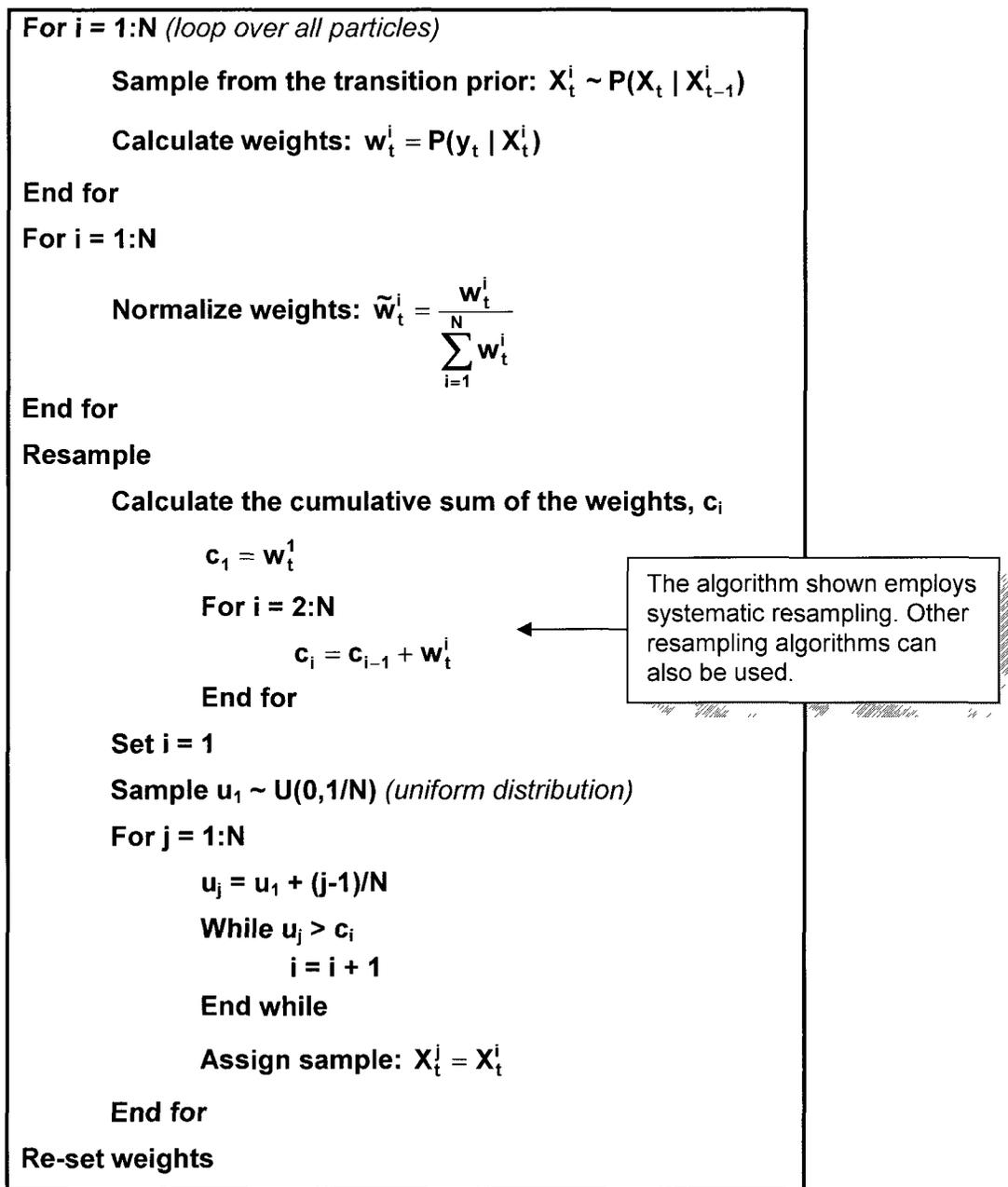
Source: Adapted from Van der Merwe et al. (2000, pg. 21)

**Figure 7-14 Graphical Illustration of a Particle Filter**

Particle filtering is essentially sequential importance sampling with re-sampling. With a prior distribution represented by a set of particles, the posterior distribution can be estimated by sampling from a proposal distribution, and weighting the samples appropriately. The most common proposal distribution is the transition prior:

$P(\mathbf{X}_t | \mathbf{X}_{t-1}^i)$ . In this case, the weights simplify to  $w_t^i = P(\mathbf{y}_t | \mathbf{X}_t^i)$ , where  $\mathbf{X}$  represents the state, and  $\mathbf{y}$  represents the evidence. An algorithm for this type of particle filter is presented in Figure 7-15. Figure 7-16 contains an extension of the algorithm for working with dynamic Bayesian networks where the transition and observation models are represented by a set of nodes, rather than a single equation.

A sequential Monte Carlo technique, particle filtering is also known as bootstrap filtering, the condensation algorithm, and survival of the fittest. A more detailed discussion of the theoretical basis for particle filters can be found in Murphy 2002, Gordon et al. 1993, Kanazawa et al. 1995, Ristic et al. 2004, and Doucet et al. 2001, from which much of the above discussion was derived.



Source: Adapted from Ristic et al. (2004)

**Figure 7-15 Particle Filter Algorithm**

```

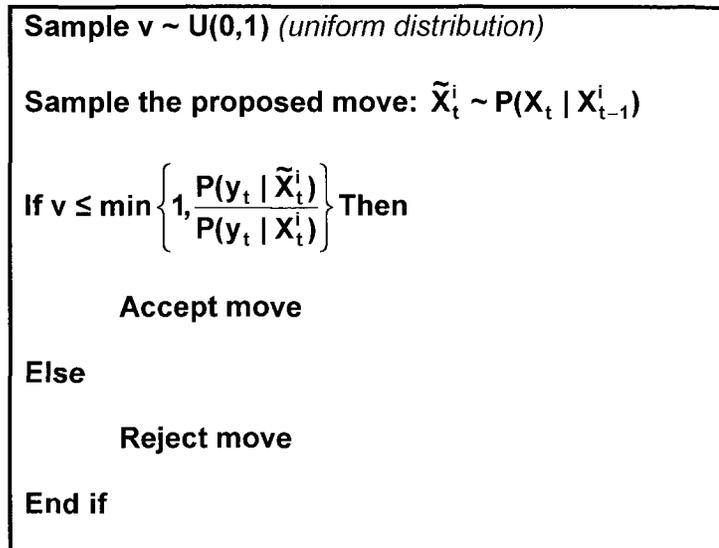
For  $i = 1:N$  (loop over all particles)
  Set  $w_t^i = 1$ 
  For each node  $X_j$  in topological order:
    Let  $u$  be the value of the Parents( $X_j$ )
    If  $X_j$  is a non-evidence node:
      Sample  $x_{j,t}^i \sim P(X_j \mid \text{Parents}(X_j) = u)$ 
    If  $X_j$  is an evidence node:
      Set the value of the node equal to the evidence,  $x_{j,t}^i = y_{j,t}$ 
      Update the sample weight:  $w_t^i = w_t^i \times P(X_j = x_j \mid \text{Parents}(X_j) = u)$ 
    End for
  End for
For  $i = 1:N$ 
  Normalize weights:  $\tilde{w}_t^i = \frac{w_t^i}{\sum_{i=1}^N w_t^i}$ 
End for
Compute expected value:  $E(X_j) = \sum_i \tilde{w}_t^i \times x_{j,t}^i$ 
Resample

```

Source: Adapted from Murphy (2002)

**Figure 7-16 Particle Filter Extension for Dynamic Bayesian Networks**

A simple particle filter with a Markov Chain Monte Carlo (MCMC) step has been implemented in the ramp control algorithm with acceptable results. The MCMC step was introduced to address particle impoverishment. Since particles with high weights are selected many times, only a few particles are carried forward at each time step, resulting in a loss of diversity. The MCMC step introduces sample variation while still approximating the posterior distribution. In the case of the ramp control algorithm, an MCMC step based on the Metropolis-Hastings algorithm was adopted, in which candidate moves are generated by sampling from the transition prior. The moves are either accepted or rejected based on the acceptance criterion, which reflects the likelihood of the observed evidence (refer to Figure 7-17).



Source: Adapted from Van der Merwe et al. (2000)

**Figure 7-17 MCMC Step**

A number of particle filter variants are described in the literature which offer the potential for improved performance. Of these, the unscented particle filter proposed by Van der Merwe et al. (2000) is considered to have particular merit.

## 7.6 Formulation of the Utility Function

In the ramp control problem, the basic objective is to select the “best” ramp metering strategy for the freeway corridor under investigation. However, the “best” strategy will vary depending on the objectives of the freeway operator, and the relative importance assigned to each objective. If the goal is simply to minimize system-wide travel time (or some similar performance measure), the ramp metering rates can be determined using standard optimization techniques. The task becomes more difficult with multiple objectives, particularly if some of the objectives conflict, so that improving one measure of performance causes another to deteriorate. In such situations, it is difficult to examine the results of different metering strategies and determine which one is preferred. Is it better to have no freeway congestion by imposing substantial delay at ramp meters upstream of the bottleneck, or is it preferable to allow some level of freeway congestion if the ramp delays can be reduced? The answer depends on how drivers value ramp delay versus freeway delay, and also on how attitudes may change as the level of delay increases. Due to the random nature of traffic flow, the outcomes of any ramp metering

strategy are subject to uncertainty, and as a result, attitudes towards risk may also influence the decision problem.

Given the uncertainty inherent in freeway performance under different ramp metering scenarios, and the inclusion of both equity and efficiency objectives in the decision problem, the selection of an optimal ramp metering strategy is not necessarily a simple task.

Utility theory has emerged as a promising technique for solving multi-attribute decision problems under uncertainty. Where outcomes are subject to a random process, it is assumed that the best course of action is to select the alternative with the highest expected utility. Expected utility (**EU**) reflects the probability **P** of each possible outcome **x** associated with alternative **X**, as well as each outcome's corresponding utility **U**:

$$EU(X) = \sum_{x \in X} P(x)U(x)$$

Thus, the expected utility for a given alternative can be calculated by simply multiplying the probability of each potential outcome by the utility of that outcome and summing the results. Appendix I provides an introduction to utility theory, while the following section describes its application to the ramp control problem.

### ***The Ramp Control Problem***

In the Bayesian decision network for the ramp control problem, decision nodes represent the ramp metering rates to be applied at different on-ramps, while utility nodes represent the multi-attribute utility corresponding to the freeway state under the imposed ramp metering control. Since the predicted freeway state is subject to uncertainty, it is expressed as a probability distribution over potential outcomes. By multiplying the probability of each outcome by its corresponding multi-attribute utility and summing the results, the expected utility can be computed. This process is repeated for each set of ramp metering rates under consideration; the scenario with the highest expected utility is selected as preferred.

Since the expected utility provides the basis for decision-making, it is important that the underlying multi-attribute utility function captures the objectives of the decision problem. In the case of the ramp metering algorithm, control objectives were defined as follows:

- Minimize freeway congestion
- Minimize ramp delay
- Minimize inequity
- Minimize impacts to the arterial road network

The challenge lies in selecting appropriate performance measures for assessing how well each of these objectives have been achieved. To address this challenge, a comprehensive evaluation of alternative performance measures was carried out to determine which measures to carry forward for inclusion in the multi-attribute utility function. The final recommendations from this evaluation are presented in Table 7-5, while the full evaluation can be found in Appendix J.

**Table 7-5 Recommended Freeway Performance Measures**

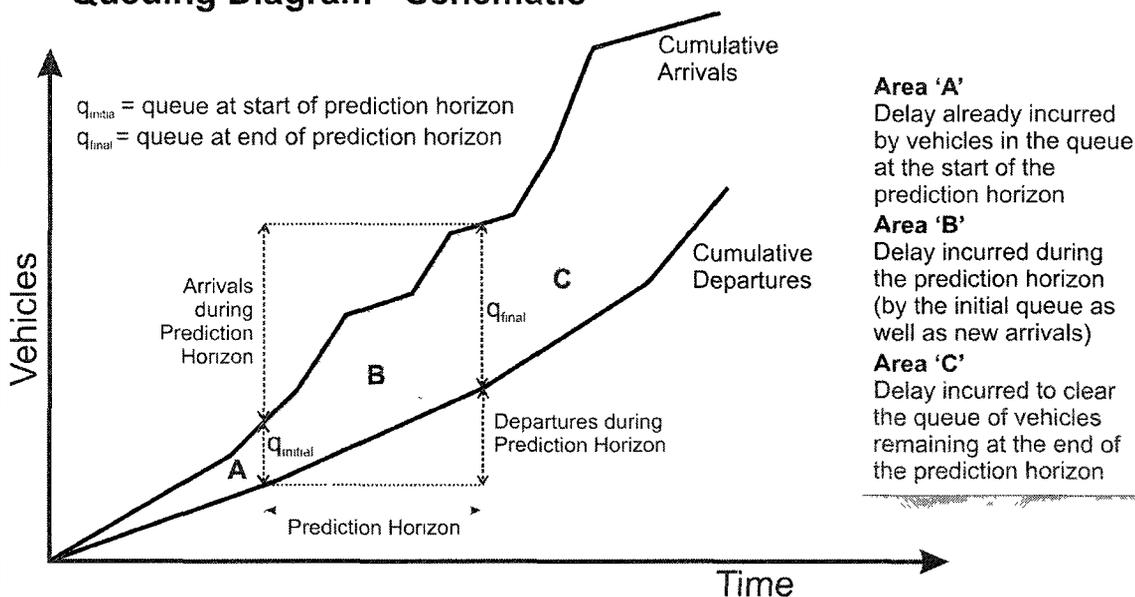
<b>Objective: Minimize freeway congestion</b>
<p><b>Recommended Performance Measure:</b></p> <p>Average freeway speed</p>
<p><b>Method of Aggregation:</b></p> <p>The average freeway speed is calculated as the weighted average of the speed in each freeway cell, where the weights are based on the number of vehicles per cell.</p>
<p><b>Issues / Implications:</b></p> <p>The average speed as calculated above does not represent the average speed to travel the corridor. Rather, it is the average speed being experienced by all drivers on the freeway at a given point in time. As a result, the speed provides an indication of system performance, and may not be representative of what individual drivers are experiencing.</p> <p>Rather than calculating the utility associated with the average freeway speed, it is also possible to compute the utility for each individual freeway cell based on the cell speed, and then weight the utilities together based on the cell volume. The benefit of the latter approach is that the calculated utilities correspond to the conditions that drivers are actually experiencing. However, the former approach provides a single measure of effectiveness for the corridor (i.e. the weighted speed) which is easy to communicate and compare.</p> <p>Regardless of which approach is adopted, the utility will be sensitive to the corridor length and number of cells. A 2 km corridor with 1 km of congestion will have a much lower average speed (utility) than a 10 km corridor with 1 km of congestion.</p>

## Objective: Minimize ramp delay

### Recommended Performance Measure:

Average ramp delay per vehicle, for those vehicles impacted by ramp metering during the prediction horizon (refer to the figure below for an explanation of how this value is calculated)

### Queuing Diagram - Schematic



$$\frac{(A + B + C)}{q_{initial} + \text{Arrivals}} = \frac{(A + B + C)}{q_{final} + \text{Departures}} = \text{Average delay per vehicle impacted by ramp metering during the prediction horizon}$$

'A' will be the same for all scenarios. Objective is to minimize 'B' & 'C'

### Method of Aggregation:

The average delay per vehicle is calculated by dividing the total delay incurred at all ramps along the corridor by the total on-ramp demand. Thus, aggregation is not required.

### Issues / Implications:

The proposed approach looks at all the vehicles delayed at some point during the prediction horizon due to ramp metering, and calculates the total ramp delay experienced by these vehicles, including any delay which may occur outside the prediction horizon. This is necessary in order to compute a meaningful value for the average delay per vehicle.

The impact of any delay incurred prior to the prediction horizon will be the same for all ramp metering scenarios examined. However, the delay incurred during the prediction horizon will be different, as will the delay required to clear the queue remaining at the end of the prediction horizon.

From the algorithm tests, it was determined that in some cases, the algorithm was acting to spread out the ramp queue among multiple ramps in order to reduce the delay incurred while dissipating the final queue at the end of the prediction horizon (with more 'servers', the queue can be dissipated more quickly). In the situation examined, the delays were acting to limit the ramp queues, making the queue length constraints virtually unnecessary.

The approach assumes that the ramp metering rates do not change after the prediction horizon has finished, which will impact the calculation of delay incurred while clearing the final queue.

### **Objective: Minimize inequity**

#### **Recommended Performance Measure:**

Standard deviation of the ramp delays along the corridor at the end of the prediction horizon, where the ramp delay is measured as the time required to clear the last vehicle in the queue assuming similar ramp metering rates persist into the future

#### **Method of Aggregation:**

The performance measure is already based on an aggregation of the results for different ramps, and no further aggregation is required.

#### **Issues / Implications:**

Section 3.2.1 provided an overview of the different approaches that have been used in the literature to measure the equity of ramp metering systems. Although the recommended measure is relatively simplistic, it is easy to calculate and provides a good indication of the variation in ramp delays along the corridor. Moreover, test results suggest that it works reasonably well within the new algorithm.

A simple ratio of the minimum to maximum ramp delay was also considered as a potential performance measure, based on the approach recommended by Meng and Khoo (2010). However, this approach focuses exclusively on extreme values, and tends to obscure any equity gains made at ramps with intermediate levels of delay.

A potential concern relates to which ramps are included in the equity assessment. For large systems, it may be more appropriate to calculate an equity measure for a group of ramps, rather than the entire corridor (i.e. ramps providing access to the freeway from the suburbs, within the downtown, etc.).

### **Objective: Minimize arterial impacts**

#### **Recommended Performance Measure:**

Occurrence of queue spillback. If spillback is predicted, the solution is deemed to be unacceptable. In the event that spillback cannot be avoided, the algorithm will attempt to minimize the extent of spillback in the network, while continuing to optimize the metering rates at all other ramps in accordance with the utility function.

#### **Method of Aggregation:**

To allow the algorithm to search for a solution which minimizes the extent of spillback in the network, the total number of queued vehicles exceeding the ramp storage is subtracted from the estimated utility. Since the algorithm is attempting to maximize the utility, it will search for the ramp metering rate at each ramp that will minimize the amount of spillback to the greatest extent possible, given the anticipated ramp demand. *At those ramps where spillback can be avoided, the algorithm will continue to adjust the metering rate to maximize the overall utility.*

#### **Issues / Implications:**

The option of including queue spillback directly in the utility function was also explored. Initially, a utility function was developed based on the remaining ramp storage. However, it was felt that this utility had little practical meaning since the remaining storage is either acceptable (i.e. no spillback is occurring) or it's not. The number of spaces remaining has little bearing; the objective is to use the available ramp storage as efficiently as possible. A preferable approach would be to develop a utility function based on the extent of queue spillback (with no spillback having a utility of 1, and any spillback greater than a certain critical value having a utility of 0).

In the end, the current approach was adopted, since it was felt to be critical to avoid impacts to the arterial network. Under the current arrangement in Ontario, the province oversees operation of the freeway network, while the local municipality has responsibility for most arterial and collector roads. It is therefore difficult to foresee a situation where the municipality would allow freeway traffic to purposefully spill back onto the arterial network, unless movement was made towards a more integrated system.

It should be emphasized that while the recommended performance measures in Table 7-5 were selected for testing the current version of the ramp metering algorithm, in practice, any number of different performance measures could be used instead, depending on the objectives of the transportation authority and the preferred means of measuring such objectives. In particular, while the definition of equity used in the utility function is primarily concerned with providing equality of service to all drivers accessing the freeway system, other definitions of equity could certainly be used instead. **This flexibility is one of the key strengths of the new ramp metering algorithm: the ability to tailor the algorithm to trade-off competing objectives in accordance with local attitudes and preferences.**

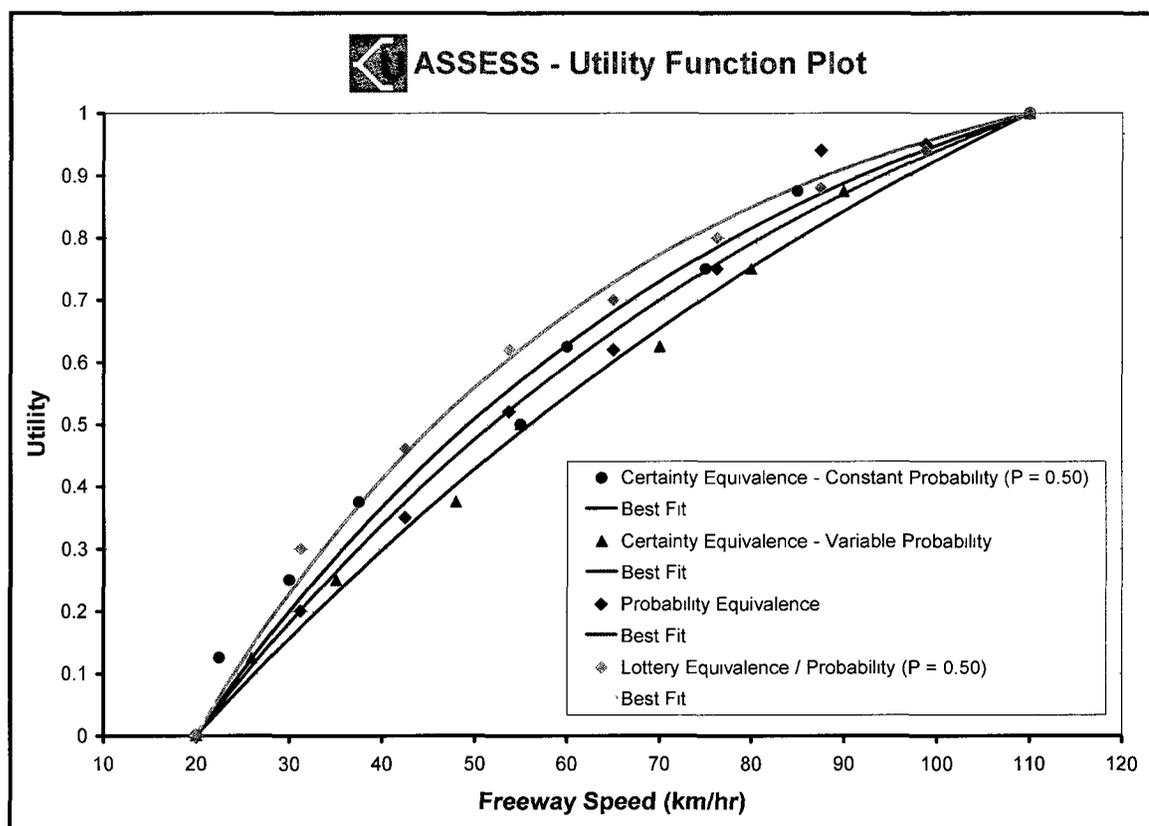
Once suitable attributes have been selected for evaluating freeway performance in relation to the control objectives, the next step is to develop a utility function for each attribute which quantifies preferences for different outcomes in the face of uncertainty and risk. These single-attribute utilities can then be combined together, forming a multi-attribute utility function which trades-off competing objectives. The resultant multi-attribute utility reflects both the relative importance of the individual attributes and their degree of interaction, providing an indication of the overall utility of the predicted freeway state for inclusion in the ramp control algorithm.

To develop the utility functions, the software ASSESS was applied (Delquié 2008). ASSESS is an interactive computer program which uses the concepts of indifference probability and certainty equivalence to derive the utility function for a particular attribute. The program poses a series of hypothetical questions designed to elicit the user's preferences towards various risky alternatives. From the response to each question, the corresponding utility can be computed. These utilities are then plotted and an equation is fitted to the data, providing an estimate of the user's utility function and associated risk tolerance.

In assessing utilities, one of four methodologies can be applied: certainty equivalence with constant probability, certainty equivalence with variable probability, probability equivalence, or lottery equivalence. Since these methodologies impact the type of question posed by the program, certain users may feel more comfortable using one

methodology over another. To examine the consistency of an individual's responses, the analysis can be repeated using different assessment methods.

Figure 7-18 compares the utility estimates for freeway speed calculated using the four approaches in ASSESS. As shown, similar results were obtained regardless of the approach applied. In general, the 'certainty equivalence with constant probability' technique was found to be the most intuitive to work with, and was therefore used to develop the utility curves included in the ramp control algorithm.



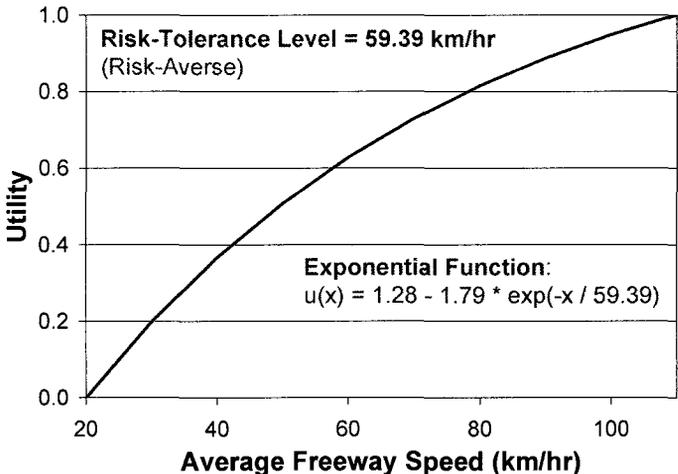
**Figure 7-18 Impact of Utility Assessment Techniques for Freeway Speed**

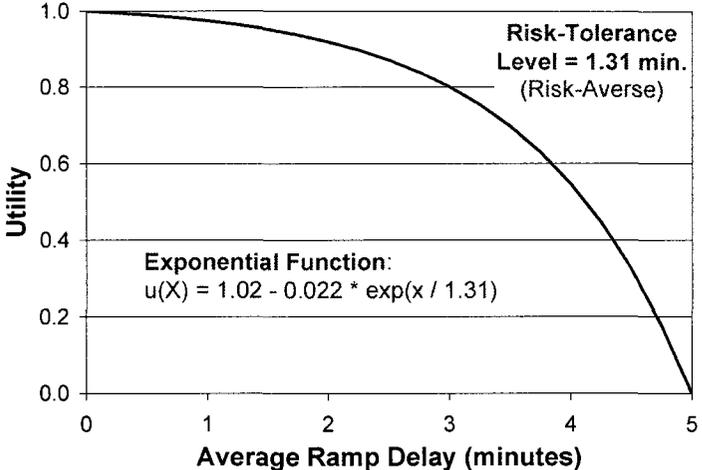
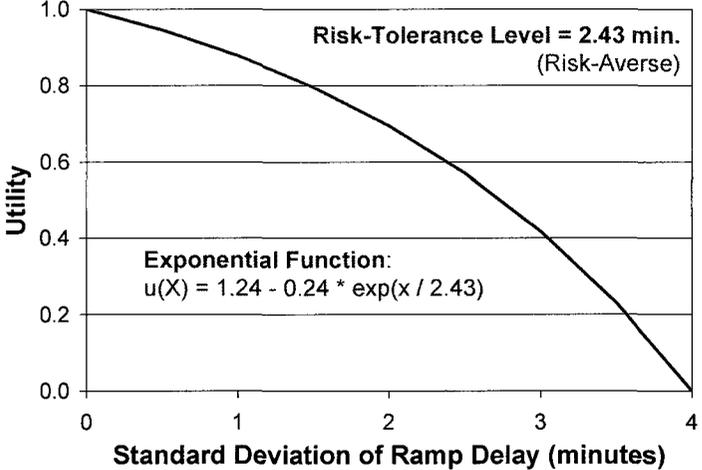
For the purposes of this research, the utility functions were developed from the perspective of a typical transportation engineer tasked with managing the freeway system, and thus reflect the preferences and priorities of a hypothetical system operator. Such an approach was considered appropriate for testing the control algorithm, since in most cases, it will be the transportation agency (and its engineering staff) who specify how the system should operate. For the hypothetical case considered in this research, it

was assumed that the main priority was reducing congestion and delay, but that some loss in efficiency could be tolerated if it means a more equitable system which is more acceptable to the public.

Table 7-6 presents the utility functions that were developed for each of the freeway system attributes described in Table 7-5. The functions were generated using the ASSESS software described above. In applying the software, responses were given which were felt to reasonably imitate how a typical transportation engineer might respond. As Table 7-6 shows, all of the utility functions are concave, indicating risk aversion. Given the consequences of traffic congestion and people's reaction to delay, such conservative behaviour is considered realistic.

**Table 7-6 Utility Functions Used in Current Algorithm**

Objective	Attribute	Utility Function
Minimize freeway congestion	Weighted average freeway speed (km/hr)	 <p data-bbox="774 930 1173 991">Risk-Tolerance Level = 59.39 km/hr (Risk-Averse)</p> <p data-bbox="981 1165 1340 1226">Exponential Function: <math>u(x) = 1.28 - 1.79 * \exp(-x / 59.39)</math></p>

Objective	Attribute	Utility Function
Minimize ramp delay	Average ramp delay per vehicle (minutes)	 <p>The graph shows a concave utility function for average ramp delay. The y-axis is labeled 'Utility' and ranges from 0.0 to 1.0. The x-axis is labeled 'Average Ramp Delay (minutes)' and ranges from 0 to 5. The curve starts at (0, 1.0) and ends at (5, 0.0). Text in the graph includes: 'Risk-Tolerance Level = 1.31 min. (Risk-Averse)' and 'Exponential Function: <math>u(X) = 1.02 - 0.022 * \exp(x / 1.31)</math>'.</p>
Minimize inequity	Standard deviation of the maximum ramp delay at each ramp at the end of the prediction horizon	 <p>The graph shows a concave utility function for the standard deviation of ramp delay. The y-axis is labeled 'Utility' and ranges from 0.0 to 1.0. The x-axis is labeled 'Standard Deviation of Ramp Delay (minutes)' and ranges from 0 to 4. The curve starts at (0, 1.0) and ends at (4, 0.0). Text in the graph includes: 'Risk-Tolerance Level = 2.43 min. (Risk-Averse)' and 'Exponential Function: <math>u(X) = 1.24 - 0.24 * \exp(x / 2.43)</math>'.</p>

In combining the individual utilities to form a multi-attribute utility function, a linear (additive) relationship was assumed. For additive utility to apply, not only must the attributes exhibit mutual utility independence,<sup>15</sup> but the attributes must also be additive independent, that is, preferences over lotteries must depend only on the marginal probability distribution of the attributes.

<sup>15</sup> An attribute **j** is considered to be utility independent of an attribute **k** if preferences between lotteries on **j** are independent of the value of **k** (refer to Appendix I). Note that both utility independence and additive independence deal with the independence of preferences, not independence of the attributes themselves. For example, for freeway speed to be utility independent of ramp delay, the decision-maker's preferences for uncertain outcomes involving freeway speed should be the same regardless of the ramp delay, which is generally expected to be the case.

According to Keeney and von Winterfeldt, if the objectives used to develop the utility model meet the criteria for fundamental objectives (i.e. objectives which define the basic reasons for being interested in the decision – the ends that are trying to be achieved rather than the means), then a strong case can be made that an additive model is appropriate.<sup>16</sup> Since the objectives used in the ramp control algorithm generally meet the criteria for fundamental objectives, use of an additive utility formulation was considered to be appropriate. Moreover, in applying the ASSESS software to develop preliminary estimates of the function parameters, the resulting values generally supported a linear model, at least for the typical transportation engineer being emulated.

With additive utility, the coefficients of the individual utility terms add to one; interaction is not considered. As a result, the utility coefficients can be thought of as weights which capture the relative importance of the various objectives. In developing appropriate values for these weights, a two-step process was applied. First, the weights were estimated using the ASSESS software. Then, a series of simulations was carried out to refine the utility weights, ensuring that the algorithm was performing as intended. In theory, such a process should not be required if the decision-maker's preferences are elicited correctly. In practice, the process of "tuning" the weights is likely unavoidable given the difficulty in rationalizing complex trade-offs, particularly in cases where the attributes are not necessarily intuitive.

Table 7-7 presents the utility weights that were applied in the ramp control algorithm as part of this research. Two scenarios were considered:

1. Optimization for efficiency only (to allow comparison with other traditional ramp metering algorithms)
2. Optimization for both efficiency and equity (to demonstrate the strength of the utility approach)

As with the individual utility functions, the utility weights were developed from the perspective of a typical transportation engineer; in practice the weights would be tailored to reflect agency priorities and objectives, local practice, or even feedback from the public. Indeed, the power of the utility approach lies in the ability to develop custom

---

<sup>16</sup> See Chapter 13, "Practical Value Models", in Edwards et al. (2007).

utility functions to suit specific needs. Other formulations of the utility function (incorporating different utility weights or even different attributes or objectives) are not only feasible, but desirable to ensure the ramp control system successfully optimizes freeway operations in accordance with local preferences.

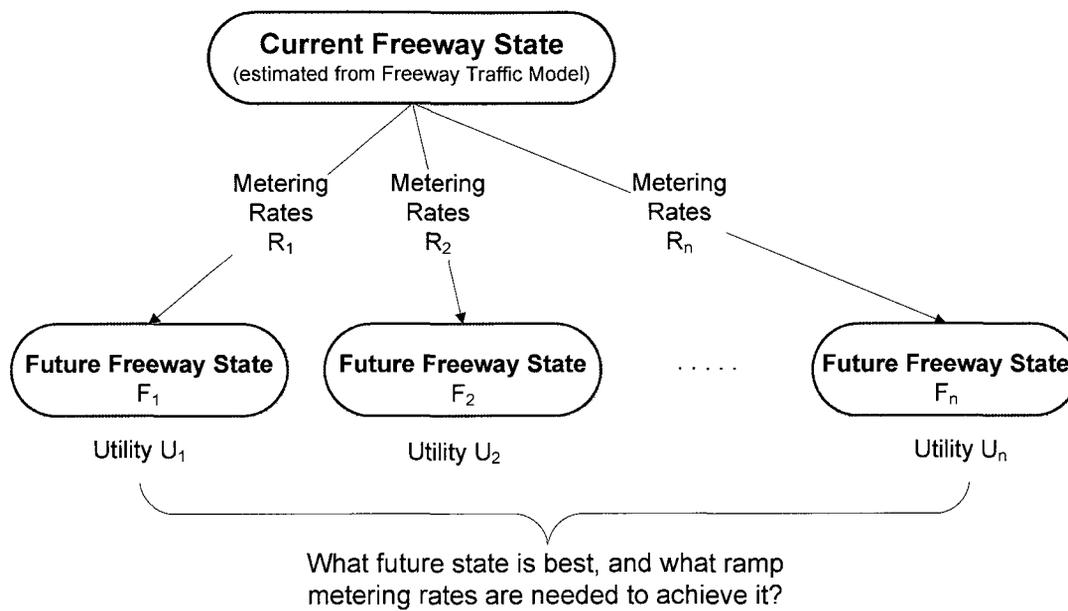
**Table 7-7 Multi-Attribute Utility Function Weights**

Objective	Attribute	Utility Weights	
		Efficiency Only	Efficiency + Equity
Minimize freeway congestion	Average freeway speed	0.6	0.55
Minimize ramp delay	Average ramp delay per vehicle	0.4	0.25
Minimize inequity	Standard deviation of the ramp delay	0.0	0.20

## 7.7 Solving the Control Problem

### 7.7.1 Overview

Given the current freeway state, what control solution will result in the “best” future state? This question, depicted graphically in Figure 7-19, forms the basis of the control problem. As discussed in Section 7.6 above, the “best” solution is defined in terms of a utility function which captures the importance placed on the various performance objectives. The challenge then is to find the solution with the maximum utility. The following sections provide an overview of the methodology for addressing this challenge. In section 7.7.2, the notion of predictive control is formally introduced, while Section 7.7.3 explores the use of pattern search techniques for determining the optimal control strategy.



**Figure 7-19 The Control Problem**

### 7.7.2 Predictive Control

To find the global optimal solution to a control problem, all input variables must be known in advance. Given the complex, dynamic nature of freeway systems, the ability to predict such information a priori is highly questionable. Real-time algorithms take advantage of data as it becomes available during the control process, and are thus more responsive to actual freeway conditions.

Real-time algorithms can operate in one of two ways. In the simplest approach, the control action in the next time step is determined based on current conditions, without regard for the future. This is the approach adopted by Zhang and Levinson (2004a). By ignoring dependencies between time slices, the problem is simplified. However, the optimal ramp metering rates calculated under such conditions may not be optimal if a longer timeframe were considered, resulting in less efficient usage of system resources as the peak period progresses.

The second approach involves predictive control – actions in the next time step are determined based on how the system is expected to respond in the future (Karimi et al. 2004; Bellemans et al. 2006a/b). Since the potential for traffic congestion may only

become apparent by looking several time steps ahead, this approach allows appropriate action to be taken before problems arise.

An excellent description of predictive control is provided by Karimi et al. (2004): Model predictive control “is a control algorithm that searches for the optimal future control sequence that minimizes a pre-defined objective function over a near-future time horizon” (pg. 491). Predictive control involves three distinct time intervals:

- The control interval defines when the ramp meter controllers are updated. Typically, control parameters are updated every 30 to 60 seconds.
- The prediction horizon defines the interval over which system outputs are to be estimated. Optimization over this interval is carried out to determine the sequence of control actions which best meets the control objectives.
- The control horizon is shorter than the prediction horizon, and defines the period over which ramp metering rates are to be optimized. Once the control horizon has passed, ramp metering rates are assumed to be constant for the remainder of the prediction horizon. The use of a control horizon reduces the computational complexity of the optimization problem.

In carrying out predictive control, a receding horizon is used. The algorithm gets updated data from traffic sensors, computes the optimal control strategy for the remainder of the prediction horizon (by estimating how the network will perform under various scenarios), implements the strategy for the next control interval, and then repeats the process. In general, the optimal strategy consists of a series of ramp metering rates to be implemented at each ramp meter over the prediction horizon, one rate for each control interval. However, only the first metering rate is actually applied. All future rates are re-calculated during the next iteration as the prediction horizon is shifted forward.

When establishing the prediction horizon, it is important to consider the length of time required for vehicles to travel from one end of the corridor to the other. If any of the vehicles entering the corridor contribute to downstream congestion, it may be necessary to restrict freeway access. At the same time, the prediction horizon should not be so long that computational issues arise. The algorithm must be able to arrive at a solution in the allocated time. Prediction accuracy is also a consideration – the further one projects into the future, the more uncertain the results.

### 7.7.3 Calculation of Control Parameters

Solution of the control problem in real-time is not a trivial task given the stochastic, non-linear relationships involved. From a review of the available options, a pattern search technique was selected for solving the decision problem. MATLAB's Direct Search toolbox contains a collection of functions that can be directly integrated into the control algorithm.

The Direct Search functions in MATLAB are intended to extend the capabilities of standard optimization techniques, solving problems for which the objective function is not differentiable, not continuous, or stochastic. According to the MATLAB user's manual, direct search methods do not require information on the gradient of the objective function or higher derivatives (MathWorks 2010). Instead, the algorithm searches a set of points around the current point in the search for an optimal solution. If a new point is found which reduces the objective function, it becomes the current point, and the process is repeated. If a better solution is not found, the mesh of points is refined and the search begins again. The mesh is formed by adding the current point to a scalar multiple of a set of vectors which define a pattern over which to search. A successful "poll" results in the mesh being expanded over the solution space, whereas an unsuccessful "poll" causes the mesh to be contracted. The process continues until the stopping criteria are satisfied. Over time, the algorithm produces a sequence of points which get successively closer to the optimal value.

While the pattern search algorithm will eventually arrive at a solution, the challenge lies in doing so within a limited amount of time. Indeed, it was recognized from the outset that the speed of the solution algorithm was likely to be an issue given the real-time constraints of the ramp metering problem. It was generally assumed that, as long as the extent of the deficiency was not substantial, any speed-related issues could be resolved in subsequent phases of work using a variety of hardware and software techniques for improving code performance (such as conversion to a faster programming language or use of parallel processing). As a result, the primary focus of the current research was assessing the feasibility of the proposed approach; no attempt was made to enforce the real-time restrictions required for field implementation. Nonetheless, a number of

measures were explored to improve the speed of the pattern search routine and ensure viability for real-world application. A brief overview is provided below.

- **Elimination of infeasible solutions** – The speed of the solution algorithm can be improved by reducing the size of the solution space. One way of doing so is to impose constraints on the solution space prior to beginning the optimization problem. Accordingly, the ramp metering algorithm was modified to include upper and lower limits on the control solution. Within the algorithm, initial limits are defined based on the minimum and maximum allowable ramp metering rates. The upper bound is then modified to reflect the anticipated level of ramp demand, while the lower bound is adjusted by determining which ramp metering rates are likely to be infeasible due to queue length constraints. By imposing such constraints up front, the number of feasible solutions is reduced, facilitating the search process.
- **Solution precision** – The level of precision required for the control solution was also explored. In the end, a decision was made to round the optimal ramp metering rates (expressed in vehicles per minute) to the nearest whole number, reducing the size of the solution space considerably.<sup>17</sup> This was accomplished in the pattern search algorithm by adjusting the mesh tolerance. Overall, the precision of the ramp metering solution is expected to have only a minor impact on the algorithm performance from a practical implementation perspective. Clearly, there is no point in applying greater precision than can be implemented in the ramp controller. There is also no point in applying greater precision if the difference amounts to only a few vehicles over the entire peak period. At the same time, the solution should not be so coarse as to limit responsiveness. By rounding one minute flow rates to the nearest whole number, the maximum error per minute is 0.5 vehicles, or 5 vehicles over 10 minutes. However, since the flow rate is adjusted every minute, the rates can be selected such that the errors tend to cancel out over time.
- **Optimization of MATLAB code** – The Profiler Utility in MATLAB was used to identify inefficient sections of code which were then targeted for improvement. While substantial gains in performance were achieved, it is anticipated that a more thorough code review would result in even further improvements.
- **Selection of pattern search options** – There are a number of options which can be specified when working with the pattern search tool. After some experimentation, a decision was made to use the parallel processing option to make the most efficient use of the available computer hardware. In doing so, cache and vectorization must be set to “off”, while complete poll must be set “on”. Other changes to the default settings include: poll method = GSSPositiveBasis2N; initial mesh size = 2; mesh scaling = “off”; and mesh

---

<sup>17</sup> It is also possible to control the precision of the cycle length rather than the metering rate. However, doing so tends to create uneven jumps in the metering flows.

rotation = “off”. While these settings were found to yield reasonable performance, further testing is needed to confirm that the selected values are truly optimal.

- **Development of custom patterns (heuristics)** – The pattern search tool allows the user to specify an optional search routine which is carried out each iteration prior to polling. To take advantage of this option, a series of custom patterns was developed to hone in on solutions faster using knowledge of the most effective metering techniques. For example, the most efficient metering schemes tend to restrict metering to those ramps closest to the bottleneck. For the most part, the custom patterns were tailored towards efficiency objectives, however custom patterns that address equity could certainly be developed.
- **Assessment of other solver algorithms** – In addition to the pattern search tool, other options for solving the control problem were also examined, including the genetic algorithm and various linear solvers in MATLAB, with limited success. Although these options were dropped from consideration for the current research, further assessment of these and other alternatives may be warranted in subsequent phases of work if field trials are pursued.

The speed of the solution algorithm can be further increased by modifying certain key parameters. Doing so, however, requires careful judgement to ensure that the performance of the algorithm is not adversely affected. Relevant parameters include:

- **Duration of the control interval** – While the duration of the control interval does not impact the speed of the solution algorithm, it does impact the time available for finding a solution; the longer the control interval, the longer the algorithm has to determine the optimal ramp metering rates. Accordingly, control intervals of 40, 60, and 80 seconds were explored. For the majority of the simulation tests, a 60 second control interval was adopted. In general, the longer 80 second interval did not provide sufficient sensitivity to changing freeway conditions, the shorter 40 second interval did not offer sufficient improvement to justify the loss in calculation time.
- **Number of unique control intervals** – The control horizon dictates the number of control intervals for which a unique set of ramp metering rates must be derived. If three sets of ramp metering rates are to be implemented over the prediction horizon, then the solution algorithm must solve for 27 different variables, assuming 9 on-ramps within the freeway system (as in the test network described in Section 8.2). With only two sets of ramp metering rates implemented over the prediction horizon (i.e. one set of rates for the first control interval, and another set of rates for all remaining control intervals), then the number of variables to solve for drops to 18. If the ramp metering rates are assumed to be constant over the entire prediction horizon, then there are only 9 variables to estimate, significantly reducing the computational effort.

In general, allowing the ramp metering rates to vary over the prediction horizon allows for more flexible control, enhancing efficiency. For example, the algorithm can meter hard if necessary, and then back off on the metering rate to prevent queue spill-over or respond to declining mainline demand. Conversely, the algorithm can delay imposing a more restrictive metering rate until sometime later in the prediction horizon when flow levels become more critical. Indeed, by allowing the metering rates to vary, the algorithm can more precisely time when different levels of metering are required at different points along the corridor such that bottleneck capacity is not exceeded, recognizing that upstream actions may involve a time lag before the effects are felt downstream. Thus, any move to restrict the number control intervals with different ramp metering rates may impact how the algorithm performs.

In the end, the case with three unique sets of ramp metering rates was found to be difficult to solve for the test network under investigation. All algorithm testing was therefore carried out assuming either 1 or 2 unique control intervals over the prediction horizon.

- **Length of the prediction horizon** – The shorter the prediction horizon, the fewer the calculations needed to predict the future freeway state under a particular ramp metering scenario. At the same time, the prediction horizon must be long enough for actions taken at the far end of the corridor to be felt at the bottleneck location; if the time to reach the bottleneck from a particular on-ramp exceeds the length of the prediction horizon, the algorithm will see no benefit in imposing ramp control.
- **Number of particles used for predicting the freeway state** – The number of particles used to represent the freeway state has a direct impact on the algorithm speed; a 50% reduction in the number of particles will cut the run-time by approximately 50% as well. However, with fewer particles, the variability of the resulting traffic estimates is increased, and it becomes more difficult to distinguish the optimal solution, especially when the utilities are similar. Not only does this lead to poorer freeway performance, but the run-time may actually increase since random variation in the utility results may cause a scenario to be discarded in one iteration only to be selected as optimal sometime later, causing the algorithm to fluctuate between solutions.
- **Amount of demand uncertainty** – The ramp demand over the prediction horizon is subject to uncertainty. While this uncertainty should theoretically be captured in the freeway analysis, it was found that the algorithm performed equally well with less demand uncertainty as long as a suitable ‘safety margin’ was built into the ramp storage capacity. With less uncertainty, the variability of the utility estimates is reduced, requiring fewer particles to predict future conditions.
- **Maximum allowable change in the metering rate** – The option of limiting the maximum change in the ramp metering rate from one control interval to the next was also explored. This has the benefit of reducing the size of the solution space, and also prevents wild swings in the metering rate from one interval to another.

However, since ramp demand can vary quite significantly, corresponding changes in the ramp metering rate may sometimes be justified.

Each of the above parameters were modified in isolation and in combination to determine the optimal settings in terms of algorithm performance (including speed). The parameter values adopted in the final algorithm test runs are described in Section 9.2, while Section 9.3 provides an indication of the average solution run-times that were achieved.

In general, the results of the algorithm tests suggest that the proposed approach is viable for real-time control. While the time required to find the control solution currently exceeds the duration of the control interval, the extent of the deficiency is relatively minor depending on the assumptions adopted (refer to Section 9.3). Such results imply that it should be possible to meet the real-time constraints of the problem by improving the MATLAB coding, converting the code to a more efficient language, upgrading the computer hardware, or using more parallel processing to carry out different computing tasks simultaneously. While such options were not explored as part of the current study, it seems reasonable that solutions can be found.

## **8 VALIDATION OF THE FREEWAY TRAFFIC MODEL**

### **8.1 Test Philosophy**

As part of the “proof of concept” phase of the research, an off-line version of the Bayesian decision network was developed in MATLAB for testing the validity of the proposed approach. The objectives of this initial phase of testing were twofold:

1. To confirm the accuracy of the Freeway Traffic Model which forms an integral part of the Bayesian network; and
2. To test the use of particle filters for probabilistic inference in dynamic Bayesian networks from a speed and reliability perspective.

### **8.2 The Freeway Test Network**

The test network was created using the micro-simulation software, VISSIM. The network is comprised of an eastbound section of freeway with 6 off-ramps and 9 on-ramps. In terms of cross-section, the freeway has three through lanes, with auxiliary speed changes lanes provided at all merge and diverge locations.

Since the Freeway Traffic Model used in the Bayesian network simulates the flow of vehicles between successive freeway segments (or cells), the freeway corridor in VISSIM was divided accordingly. Cell boundaries were generally located near on-ramps to model the potential for flow breakdown at these locations. Cell boundaries were also positioned near off-ramps so that sensor data used to determine off-ramp exit percentages could also be included as mainline flow evidence in the process of belief updating. To prevent vehicles from crossing more than one cell boundary in one 10 second time step, the minimum cell length was set at 350 m.

Although initial testing of the Freeway Traffic Model was carried out under the assumption of no ramp metering, the test network developed in this phase of work was later used to assess the performance of the new ramp metering algorithm. For this reason, the network was designed from the outset to accommodate ramp metering operations.

A diagram illustrating the network layout can be found in Figure 8-1, along with a snapshot of a small portion of the network as implemented in VISSIM. Within Figure 8-1, the cell boundaries are noted, as well as the location of traffic sensors for providing speed and flow data to the control algorithm.

To ensure a realistic network configuration, key freeway elements such as the interchange spacing, length of auxiliary speed change lanes, and ramp length were loosely based on the characteristics of the Highway 417 freeway corridor in Ottawa. The positioning of the ramp meter stop bar on each ramp was determined by reviewing ramp meter design guides (Arizona DOT 2003; Caltrans 2000; Mn DOT 2001; Mn DOT 2009; Nevada DOT 2006). In general, ramp meters were positioned roughly 150 m from the start of the speed change lane to allow sufficient distance for acceleration up to merging speeds after leaving the stop position.

It was further assumed that all ramps would have two lanes of vehicle storage, the normal travel lane, and the ramp shoulder, with vehicles released onto the freeway in an alternating pattern, one vehicle per green. Since only one ramp lane joins the freeway, the shoulder storage lane is tapered out after the ramp meter. This configuration is applied in practice to increase the ramp storage – a key determinant of ramp metering success. Without sufficient storage, vehicles either spill back onto the arterial network, or are released onto the freeway at inopportune times, creating congestion.

To test the algorithm under a range of traffic conditions, a 2.5 hour simulation was performed. Travel demand for the test network was generated so as to create realistic traffic patterns, with congestion forming and dissipating along the freeway corridor over the simulation period. In developing the demand matrix, a trial and error process was used, with the objective of creating sufficient freeway congestion to demonstrate the benefits of ramp metering. Ramp demand was selected so as not to exceed the capacity of the ramp meter (i.e. 900 vph assuming one vehicle per green). It was assumed that 2% of the travel demand was comprised of heavy vehicles. Although vehicle routing was determined internally by VISSIM using dynamic assignment, convergence of traffic volumes was not an issue since there is at most one path between each origin-destination pair (i.e. each entry/exit point along the corridor, typically located at on- and off-ramps).

Additional details on the VISSIM test network can be found in Appendix K.

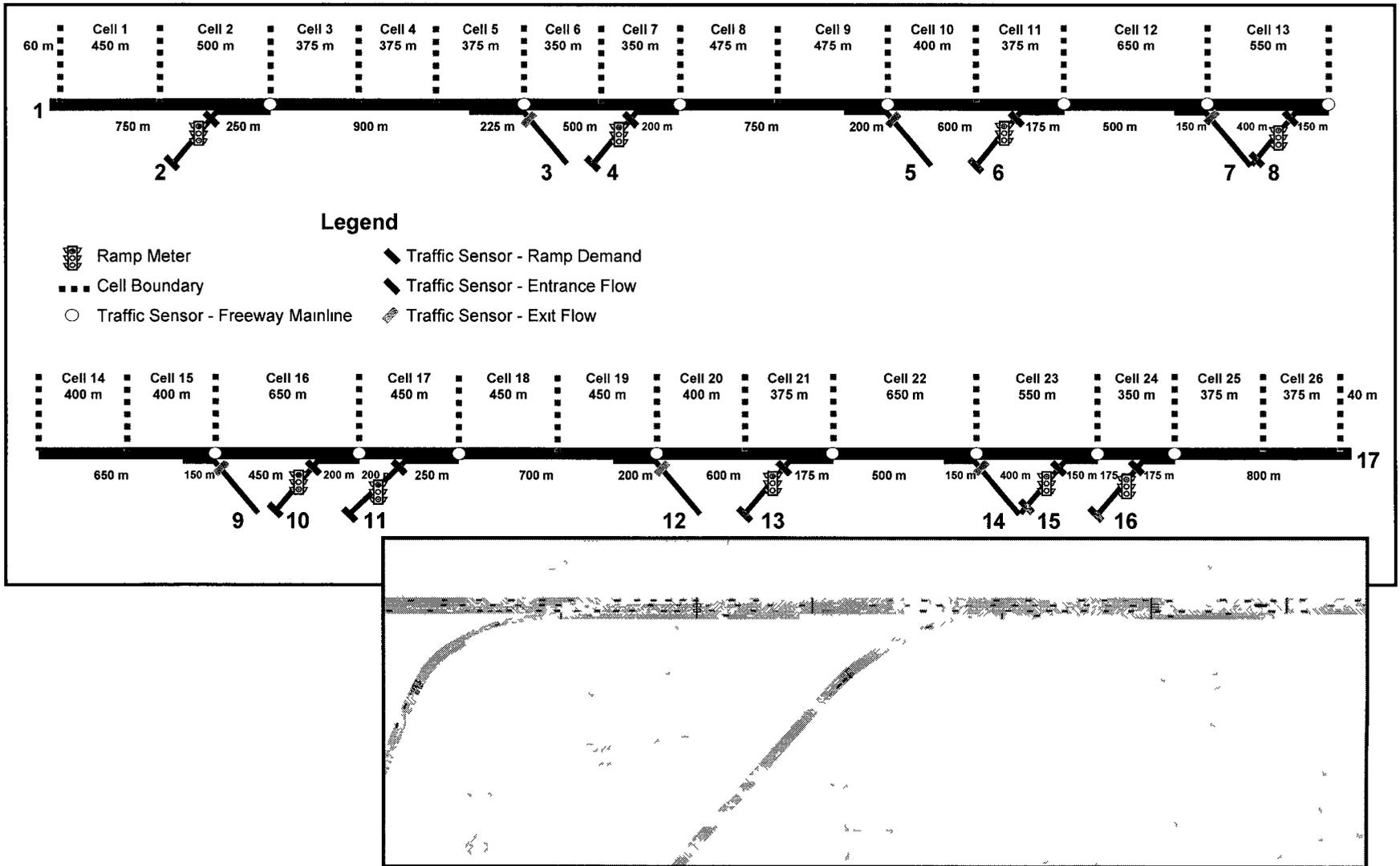


Figure 8-1 The VISSIM Test Network

### 8.2.1 Modelling of On-Ramp Merging

As a hypothetical test network, no effort was made to calibrate the VISSIM model. However, care was taken to ensure that the model parameters were reasonable, and that the driver behaviour being simulated was realistic. In assessing the performance of the test network, it was observed that the default driver behaviour parameters in VISSIM were resulting in abnormal operations at freeway merges. In particular, it was noted that drivers attempting to enter the freeway under peak conditions were often unable to find a gap in mainline traffic, causing a queue to form on the ramp. This queue would gradually increase over time, becoming unrealistically long as more and more drivers became trapped on the ramp unable to complete the merge manoeuvre. At the same time, with no merging conflicts, drivers on the freeway were able to maintain free-flow speeds despite levels of ramp demand which would typically result in breakdown conditions. In reality, merging vehicles are generally able to ‘force’ their way onto the freeway even under conditions of high demand due to the cooperative behaviour of approaching vehicles. As mainline traffic approaches the merge area, drivers will either change lanes or slow down to allow ramp traffic to merge – actions which tend to trigger congestion when traffic volumes are high. In VISSIM, the former behaviour is not modelled, while the latter behaviour can be controlled by adjusting the driver behaviour parameters for lane changes.

Another issue which was identified was the tendency for close-calls (or even collisions) to occur during the merge manoeuvre, particularly when the vehicle changing lanes was accelerating from a stopped (or near stopped) position in the speed change lane after waiting for a gap in mainline flow. Such collisions/close-calls may occur in reality, however, at a much lower frequency than observed in the simulations.

Discussions with PTV America (the developers of VISSIM) provided suggestions for addressing the above issues. Unfortunately, however, many of the options to facilitate merging were also observed to increase the number of unrealistic collisions and vice versa. In the end, the major change adopted was to increase the default value for the Maximum Deceleration for Cooperative Braking from  $-3.0 \text{ m/s}^2$  to  $-7.5 \text{ m/s}^2$ . This change was found to substantially improve the ability of ramp vehicles to merge with mainline

traffic by causing vehicles in the shoulder lane to brake more frequently when confronted with a ramp vehicle. Appendix L provides additional information on the various options that were considered to improve the operation of the VISSIM model in merge areas.

Although the number of unrealistic collisions/close-calls observed in the simulations remains higher than desirable, the impact on the overall operational performance of the freeway is expected to be minor. While the simulation shows the vehicles colliding, they simply continue on with their journey as if nothing had occurred – similar to the case in reality where no collision does occur – producing minimal impact on the model results. Of greater concern, but more difficult to address, is the fact that drivers in reality would act more cooperatively to avoid these collisions/close-calls, either by changing lanes or slowing down. Thus, it is not so much the fact that vehicles collide that is important, but the fact that the model may be under-representing the cooperative behaviour of drivers, and the resultant impact on capacity. This issue is highlighted in the following correspondence received from PTV America (2010):

*At this time, a limitation of all simulation models is the ability to explicitly replicate tactical driving behavior. In a merging scenario, this is the behavior where the driver on the entrance ramp and the driver on the freeway in the right most lane begin to “cooperate” to form a gap to facilitate the merge before both vehicles arrive in the merge area. This tactical behavior is not available in simulation models, yet. Therefore, to achieve the capacities that result from this cooperation requires adjusting factors like the desired safety distance reduction factor, acceptable headway settings, etc. until the desired capacity is achieved. It also involves checking to see if the visualization (frequency of collisions) is reasonable given this tactical driving behavior limitation.*

After increasing the Maximum Deceleration for Cooperative Braking, the driver behaviour being modelled in VISSIM was deemed acceptable for assessing the impacts of ramp metering. However, moving forward, there is a need to address the limitations of current software to more accurately depict driver behaviour in merge areas, in particular, the tendency of drivers to move out of the shoulder lane to avoid conflicts with merging vehicles.

### **8.2.2 The Ramp Signal Controller in VISSIM**

To model ramp control within the VISSIM test network, the control logic must be specified. In VISSIM, each signal head is controlled by a signal controller. Different

types of controllers are available, designed to emulate the various controller types used in real-world networks. Users also have the option of defining their own signal control logic using the VAP (Vehicle Actuated Programming) module. To use VAP, the control logic is specified in a text file using the VAP programming language. During the simulation run, VISSIM retrieves information on the current detector status, interprets the control logic in light of this information, and implements the desired signal changes in VISSIM. To aid in the creation of VAP files, a graphical programming interface is available called VisVAP which represents the control logic in the form of a flow chart.

Given the need to develop specialized control logic which would work with the ramp metering algorithm being implemented in MATLAB, VisVAP was used to develop a custom signal controller for the VISSIM test network using the sample files provided with VISSIM as a starting point. Essentially, the control logic takes the optimal cycle lengths calculated by the ramp metering algorithm and implements them in accordance with its internal logic. Key elements of the control logic include:

- **Action to take when the meter is first turned on** – When the ramp meter is first turned on, the signals for both storage lanes should transition to red.
- **Criteria for turning the meter off** – Even though the ramp metering algorithm may recommend turning the meter off, the meter should remain on as long as a ramp queue exists. The release of a large number of vehicles simultaneously is both unsafe and likely to trigger traffic congestion. Thus, if a ramp queue is detected, the cycle length is set equal to its minimum value (i.e. 4 seconds) until the queue has dissipated, at which point the meter is turned off.
- **Logic for switching the signal display from red to red-amber<sup>18</sup>** (or green if a red-amber phase is not used) – Vehicles are to be released in accordance with the specified cycle length. In practical terms, this means that the control logic should initiate the red-amber (or green) phase once an amount of time equal to the cycle length has passed, as long as vehicles are detected at the meter. However, if no vehicles are detected, the signal should remain red until the next vehicle arrives. Complicating the situation is the fact that vehicles are being released from two lanes. If a queue exists at the ramp, vehicles are released in an alternating pattern. If no queue exists, then the signal should respond to whichever vehicle arrives first (unless two vehicles arrive before the cycle length has ended, in which case the first lane to receive the green indication depends on which lane was previously served). Since the cycle length is continuously being updated, it is also

---

<sup>18</sup> The red-amber phase indicates to drivers that the green phase is about to begin.

important for the control logic to ensure a smooth transition from one cycle length to the next.

- **Duration of each signal phase** – Within the control logic, the red-amber, green, and amber phases each have a duration of 1 second, while the duration of the red phase is varied according to the cycle length. Since the minimum length of the red phase is 1 second, the above values imply a minimum cycle length of 4 seconds. Note that the use of a red-amber phase is not strictly required, and could be replaced with an additional second of green time. In a simulation environment, both alternatives produce similar results, however, in real-world networks, the use of a red-amber phase should be in accordance with local practice.
- **Logic for transitioning from red-amber to green to amber to red** – While the transition from one phase to another may seem straight-forward, in VAP, all signal changes must be explicitly defined.

The VisVAP flow chart illustrating the ramp control logic can be found in Appendix M. Appendix M also contains the corresponding VAP file in text format.

### **8.3 Methodology & Key Assumptions**

Validation of the Freeway Traffic Model was carried out for three different VISSIM runs with varying levels of congestion. For each VISSIM run, sensor readings and other simulation outputs were summarized for testing with the Freeway Traffic Model. Speed and flow measurements at cell boundaries and ramps were used as evidence in the model, while cell density measurements were used to assess the accuracy of the model compared to what was actually observed. Although all testing was done off-line, the inputs are based on real-time observations from the simulation runs, and are believed to provide a suitable basis for assessing model accuracy. Moreover, by employing an off-line approach, the Freeway Traffic Model can be applied multiple times to the same underlying VISSIM data without the need to repeat the simulation, reducing the run-time considerably, and providing a consistent dataset for model testing and refinement.

To assess the accuracy of the model predictions, results from the Freeway Traffic Model were compared with the simulation outputs. In addition to reviewing graphical trends, the estimated density for each freeway segment was compared to the actual density observed during the simulation. Since the Freeway Traffic Model is probabilistic, predictions of freeway performance will vary from one run to the next. To capture the impacts of such

stochastic variation, the model was applied multiple times for the same set of VISSIM sensor readings. The accuracy of each run was estimated by calculating the Root Mean Squared Error (RMSE), which takes into account the difference between the estimated and actual density of each freeway segment,  $i$ , at each time step,  $t$ , where:

$$\text{RMSE} = \sqrt{\frac{\sum_{i,t} (\text{Density}_{\text{Est}} - \text{Density}_{\text{Act}})^2}{\text{NumObs}}}$$

To assess the performance of the model with and without evidence, the Freeway Traffic Model was applied in two modes: tracking (where the model updates its belief about the current freeway state using evidence from traffic sensors) and prediction (where the model predicts the future freeway state over a given horizon without receiving evidence from traffic sensors).

At the outset of the analysis, it was assumed that data would be collected from traffic sensors at 20-second increments. This assumption is consistent with the data collection interval applied by the Ontario Ministry of Transportation for the COMPASS system in Toronto (Foo 2006). As noted in Section 8.2, the model update interval was set to 10 seconds while the minimum freeway cell length was set to 350 m, effectively preventing vehicles from entering and exiting a given cell within the same time step. When applied in tracking mode (with evidence), a total of 5000 particles was used to approximate the belief state; in prediction mode (without evidence), 800 particles were used with a prediction horizon of 6 minutes (refer to Sections 8.5 and 8.6 for the rationale for these assumptions).

Since the Freeway Traffic Model is applied at the freeway segment (or cell) level, it makes no difference whether the on-ramp flow is controlled by a ramp meter or not, as long as the magnitude of the flow is known (or can be estimated).<sup>19</sup> Accordingly, validation of the Freeway Traffic Model at the “proof of concept” phase was based on VISSIM runs with no ramp metering installed. However, the ability of the Freeway

---

<sup>19</sup> In later phases of the work, it was determined that this statement may not be strictly correct. There appears to be some evidence that the freeway capacity may increase slightly when the on-ramp flow (whether metered or not) is low. Refer to Section 7.4 for details.

Traffic Model to predict freeway performance with ramp metering was examined in subsequent phases of work as part of evaluating the new ramp control algorithm of which the Freeway Traffic Model forms an integral part.

#### **8.4 Calibration of Model Parameters**

Key parameters in the Freeway Traffic Model were described in Section 7.4. This section also outlined the parameter assumptions that were adopted for use with the VISSIM test network. The choice of parameter values was largely determined by evaluating the freeway traffic data available from the VISSIM model. For example, the relationship between the uncongested travel speed and cell density was determined by plotting the speed-density data obtained while the freeway was uncongested. Likewise, the queue discharge flow was estimated by examining the flow at the bottleneck under congested conditions, while the expected density within the queue was determined based on VISSIM density readings for freeway segments experiencing traffic congestion.

Since most of the parameters are treated as uncertain within the Bayesian framework, it was necessary to estimate a probability distribution for each parameter, rather than simply an average or typical value. This was accomplished by assuming a normal distribution and assessing the sample mean and variance. Once initial distributions were developed for each parameter, a process of testing and refinement was carried out to improve the accuracy of the model estimates (as determined by the RMSE for the segment density). For the most part, such adjustments were relatively minor. In certain cases, the Freeway Traffic Model was calibrated to model less variation in the parameter value than actually observed. It is hypothesized that with less variability, the model has greater difficulty predicting rare values, but is better able to predict the majority of values which lie close to the mean, resulting in better performance overall. Presumably, increasing the number of particles would also address this issue, but at the expense of model run-time.

A more formal calibration approach could be employed to ‘optimize’ the model parameters. For example, parameters for an earlier version of the Freeway Traffic Model were estimated using the genetic algorithm and pattern search tools in MATLAB with the

objective of minimizing the RMSE between the estimated and actual freeway performance. In addition, an automated process was developed for iterating through a series of test scenarios, each with different parameter assumptions. Although it was not feasible to test every possible combination of parameter values using this approach, by carefully reviewing the findings from one set of test scenarios and developing new scenarios accordingly, acceptable (although not necessarily optimal) parameter values could be determined.

For the final version of the Freeway Traffic Model, a less formal calibration approach was deemed to be sufficient as described above. A similar approach is recommended for calibrating the Freeway Traffic Model for real-world networks, although it is recognized that it may be slightly more challenging without the extensive range of network performance data that is available within a simulation environment. Nonetheless, it is anticipated that standard freeway sensor data should be sufficient for developing all of the necessary parameter distributions.

### **8.5 Performance of the Freeway Traffic Model with Evidence**

Figure 8-2 illustrates the performance of the Freeway Traffic Model assuming evidence is received from traffic sensors at 20 second intervals. The graphs in Figure 8-2 correspond to a freeway scenario with relatively high congestion (VISSIM Run 'A'). Due to space constraints, only selected segments are presented.

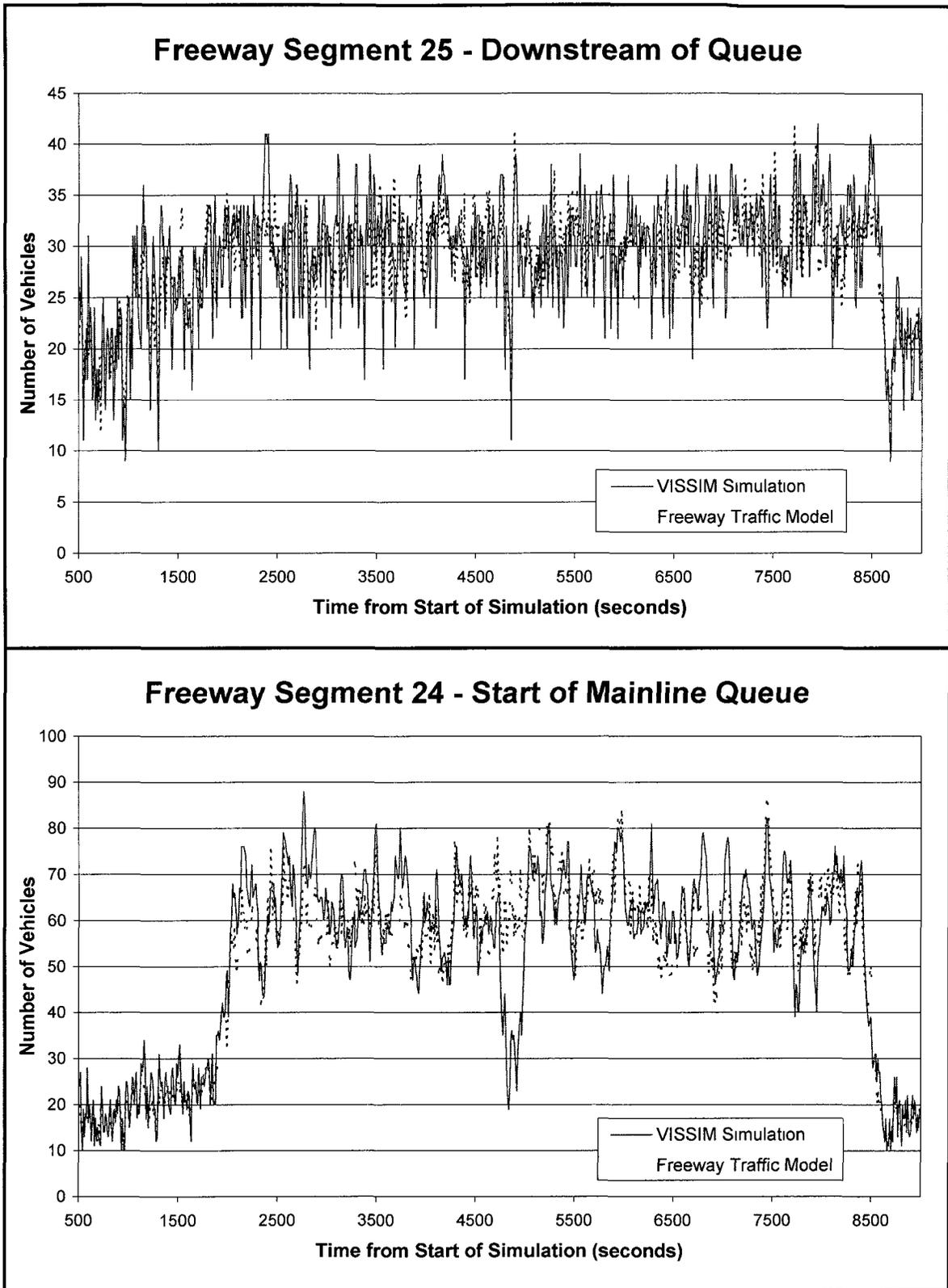


Figure 8-2a Model Performance With Evidence – Selected Segments

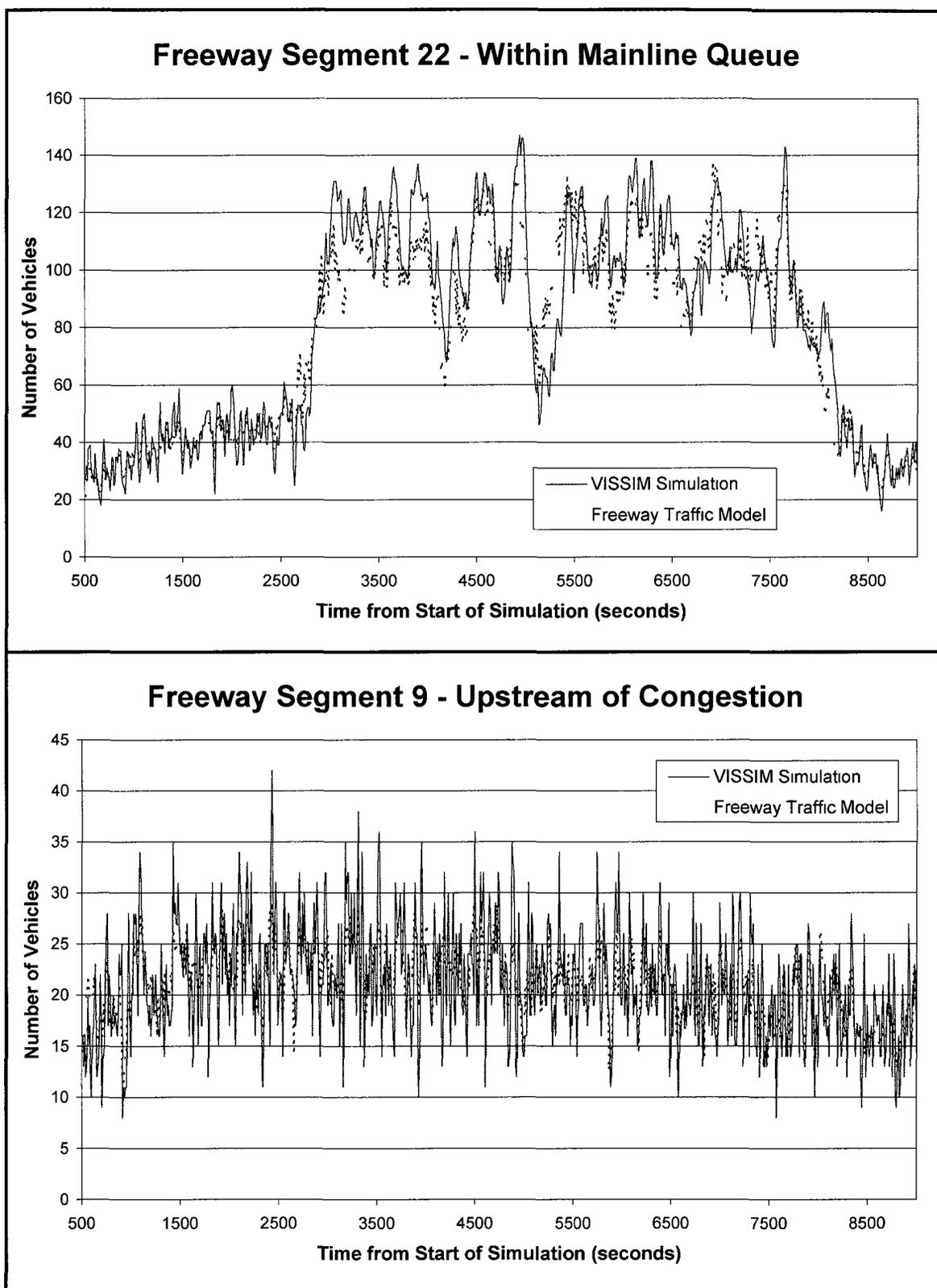


Figure 8-2b Model Performance With Evidence – Selected Segments

For the VISSIM scenario shown, the Freeway Traffic Model was applied 10 times to assess the variability in the model results. For the 10 runs, the RMSE varied from a low of 5.7 veh/km/lane to a high of 6.6 veh/km/lane, with an average error of 6.1. Overall, this level of accuracy is considered acceptable for freeway control applications. Indeed, as shown in Figure 8-2, **the number of vehicles estimated by the Freeway Traffic Model closely approximates the actual number of vehicles observed in the VISSIM simulation.** The model is able to replicate free-flow conditions, the onset of freeway congestion, queue spillback onto upstream links, and congestion dissipation as demand drops over time.

The results in Figure 8-2 are based on a Freeway Traffic Model run with a RMSE of 6.0 veh/km/lane. Results for the other 9 runs are generally comparable. In Figure 8-2, the scale of the y-axis varies from segment to segment to better illustrate model performance with and without mainline queuing.

To guard against the possibility of model over-fitting, the Freeway Traffic Model was tested using two other VISSIM simulations with different levels of congestion. In both cases, the performance of the Freeway Traffic Model was more or less equivalent to that presented above.

- For VISSIM Run ‘B’ with moderate congestion, the RMSE for the Freeway Traffic Model ranged from 5.3 to 5.8 veh/km/lane (5.6 veh/km/lane on average, based on 10 runs).
- For VISSIM Run ‘C’ with the lowest level of mainline congestion, the RMSE for the Freeway Traffic Model ranged from 4.5 to 6.1 veh/km/lane (5.0 veh/km/lane on average, based on 10 runs).

In absolute terms, the largest estimation errors tend to occur on higher-density freeway segments with mainline queues. As a result, the overall accuracy of the Freeway Traffic Model (as measured by the RMSE) tends to increase with lower levels of congestion.<sup>20</sup>

While the RMSE is a useful statistic, it can at times be misleading since the results for congested and uncongested freeway segments are lumped together. As noted above,

---

<sup>20</sup> This is particularly true since the RMSE weights larger errors more heavily (due to the process of squaring the differences).

congested freeway segments tend to have higher error; by combining all segments together, not only is it difficult to distinguish the model performance associated with different operating conditions, but the results will tend to be skewed depending on the level of congestion.

Any model runs which under- or over-estimate the duration/extent of mainline congestion will have a major impact on the RMSE. In such situations, some cells will exhibit extremely high error for a period of time, although the freeway state is otherwise predicted accurately. Given the intended role of the Freeway Traffic Model within the ramp control algorithm, the ability to accurately predict queue behaviour is less important than the ability to predict flow breakdown; as long as the model does a reasonable job at tracking mainline queues, it is expected to perform well in ramp control applications.

To carry out probabilistic inference within the Bayesian network model, the posterior probability distribution is represented by a set of particles. In total, 5000 particles were used to approximate the belief state for the model runs described above. Given the size of the freeway corridor and the extent of mainline congestion, it was anticipated that substantially more particles would be needed to obtain reasonable results, and in fact, there is some evidence that more particles would be beneficial. When mainline queues spread back over several cells, often only a few unique particles are carried forward to the next time step reflecting the low probability of the observed freeway state. From these results, it would appear that the model has difficulty predicting shockwave behaviour; the use of a relatively high standard deviation for the expected queue density means that there is a wide range of possible density values which each cell can assume when subject to mainline queuing. As the number of cells experiencing freeway congestion increases, the probability of predicting the right combination of density values (and corresponding flow rates) decreases, resulting in fewer and fewer particles that are consistent with the evidence.

Despite the above observation, it was found that increasing the number of particles beyond roughly 10,000 would actually lead to worse results, contrary to expectations. The reason for this behaviour lies in the approach used to estimate the capacity flow for the active bottleneck. Since the capacity value was found to have such a major impact on

the model results, and since the capacity was also found to vary significantly depending on the characteristics of the bottleneck, a method was sought to improve the capacity estimates. As described in Section 7.4, the Freeway Traffic Model has the ability to update the probability distribution for the congested capacity using evidence from traffic sensors. If all of the particles agree that the freeway segment has broken down, evidence reversal is employed, and the sensor data is taken as input to the model for estimating the capacity directly. Otherwise, the sensor data is used as standard evidence to determine which particles should be weighted more heavily based on the likelihood of the observations. Even with additional particles, this latter approach may lead to less accurate estimates for the capacity flow, since there is a risk of choosing “poor” particles to represent the posterior distribution; particles which most closely match the observed capacity may not be given the highest weighting when evidence for the entire corridor is considered.<sup>21</sup>

While it is preferable to estimate the segment capacity directly from the sensor observations, this is only an option if all of the particles agree that traffic flow has broken down at the location in question. With more particles, there is less chance that this will be the case, and there is therefore less opportunity to take advantage of evidence reversal for estimating the capacity flows, leading to less accurate performance overall. However, it is anticipated that as the number of particles increases beyond a certain threshold, this trend will be reversed, with accuracy increasing again as the number of particles increases.

By employing evidence reversal to estimate the segment capacity directly, the accuracy of the traffic predictions could be improved while substantially reducing the number of particles needed to represent the freeway state. Other factors that may influence the performance of the model include:

- **Number of Traffic Sensors** – In general, the more evidence available, the more accurate the results. However, at the same time, with more evidence to match, the likelihood of the observed dataset for any given particle tends to decline, resulting in fewer unique particles being carried forward. If too few particles are selected, the performance of the model could suffer. Increasing the number of particles

---

<sup>21</sup> This is particularly true under heavy congestion due to the difficulty in accurately modelling shockwave behaviour. In this situation, only a few particles may be consistent with the observed evidence in the queue, and these particles may not necessarily provide the best estimate of the bottleneck capacity.

would help to address this issue, but computer run-time will also increase, suggesting that a trade-off is needed between the number of traffic sensors used in the analysis and the number of particles representing the freeway state.

- **Corridor Length & Complexity** – Traffic operations within any given freeway segment are subject to uncertainty; the greater the number of segments, the greater the possible permutations of uncertain events, and the more particles needed to represent the belief state. Corridor length also impacts sensor requirements. With more sensors, fewer particles are likely to be consistent with the evidence. As a result, longer corridors will tend to require more particles to achieve a comparable level of accuracy.
- **Assumed Sensor Error** – The probability distribution used to represent the evidence can also affect the accuracy of the model and the number of particles carried forward. If the sensors are relatively accurate, the probability distribution for the evidence will be quite narrow. As a result, the likelihood of the evidence for any given particle will often be low. This results in fewer particles being carried forward, although the particles that are selected should more accurately reflect the actual freeway state. Again, it may be necessary to increase the number of particles to ensure an adequate number are carried forward from one time step to the next. In practice, it was found that using a slightly wider probability distribution for the evidence (suggesting greater uncertainty or error) worked reasonably well in conjunction with the other assumptions adopted.
- **Model Uncertainty** – During the model calibration process, it was found that better results could be achieved by reducing the uncertainty associated with some of the model parameters. It is hypothesized that in the majority of cases, the true parameter value lies close to the mean – with less uncertainty, more particles are focused in this area, resulting in better parameter estimates. While the model may have more difficulty dealing with rare events, overall, performance is improved. Again, such behaviour relates to the number of particles; with more particles, the sample corresponding to the central portion of the probability distribution increases, so that there is less need to limit the extent of uncertainty by excluding rare events.

The above considerations all play a key role in determining how many particles should be used to represent the belief state in order to obtain sufficiently accurate results without violating the real-time constraints of the problem. For the VISSIM test network, 5,000 particles were found to work well when combined with evidence reversal for estimating capacity flows. With 5000 particles, the Freeway Traffic Model took roughly 2 to 3 seconds to complete each tracking update. This run-time was achieved using an Intel Core i7 930 2.8 GHz computer (over-clocked to 3.6 GHz) with 6 GB of RAM.

Of the various factors described above, evidence reversal had the greatest impact on the model performance for the subject corridor. Adjustments to the sensor error and parameter uncertainty tended to trade off accuracy for fewer particles, however, since the conditions for using evidence reversal are more likely to be achieved with fewer particles, the overall performance of the model generally improved. Without the use of evidence reversal for estimating capacity flows, different assumptions would be needed to achieve comparable results.<sup>22</sup>

## **8.6 Performance of the Freeway Traffic Model without Evidence**

The Bayesian network that forms the basis for the Freeway Traffic Model is generally well-suited to modelling freeway operations when evidence is available, as demonstrated in the previous section. However, to solve the ramp control problem, the model must also perform well in prediction mode when evidence is not available.

Figure 8-3 illustrates the performance of the Freeway Traffic Model corresponding to the most congested freeway scenario (i.e. VISSIM Run 'A'), assuming that no evidence is available from traffic sensors to refine the probability distribution for the freeway state. Figure 8-3 was developed assuming that all ramp flows are known, as well as the mainline flow entering Cell 1. As a result, Figure 8-3 provides an indication of how well the traffic model performs in prediction mode, without the added error from estimating the demand.

When operating in full prediction mode, future demand inputs will not be known, but must instead be estimated from previous observations. Given the difficulty in accurately predicting future demand based strictly on past evidence, the prediction horizon must be limited in duration; not only do demand levels and patterns change throughout the peak period, but the inherent randomness of traffic flow makes it difficult to accurately predict future volumes entering the freeway network for more than a few minutes in the future. As a result, any attempt to use the model with a prediction horizon of more than 5 to 10

---

<sup>22</sup> Refer to Table 2 in Section 7.4 for a discussion of the evidence reversal process and its implications in terms of the various factors described above. Overall, without the use of evidence reversal for capacity estimation, it would not be possible to achieve the same level of accuracy unless the number of particles was radically increased. Doing so would most certainly violate the real-time constraints of the problem, and may also lead to computer memory issues.

minutes must either incorporate a more sophisticated demand estimation model, or assume the demand inputs are known in advance (as was done in preparing Figure 8-3).

As shown in Figure 8-3, the model without evidence does a reasonable job of predicting the number of vehicles per segment when the freeway is uncongested. The model is unable to replicate all of the peaks and valleys, but generally provides a good estimate of freeway usage which falls within the range of observed values.

In contrast, without evidence, the model is less accurate at predicting traffic congestion. While the model is able to predict the on-set of flow breakdown, at least at the main bottleneck location,<sup>23</sup> any shockwave behaviour within the traffic queue is ignored. Instead, the cell density in the queue remains relatively constant (approximately equal to the average queue density), and any fluctuations due to passing shockwaves are disregarded. In addition, the model also under-estimates both the duration of congestion, as well as the maximum queue length. This behaviour is a direct result of the initial assumption for the congested capacity, which, without evidence from traffic sensors, is not updated during the model run.

As noted in Section 7.4, it was found that the congested capacity is sensitive to conditions at the start of the bottleneck section. When turbulence from on-ramp merging is high (due to high on-ramp flows), the flow out of the queue tends to be somewhat lower; when turbulence is reduced (due to lower on-ramp flows, such as might be observed during ramp metering), the flow increases. Within the Freeway Traffic Model, the initial capacity distribution was calibrated to work well with conditions likely to be encountered during ramp metering. As a result, the model is less well-suited to predicting mainline

---

<sup>23</sup> The Freeway Traffic Model is able to predict flow breakdown at the main bottleneck in each of the three VISSIM scenarios that were investigated, even when evidence is not available from traffic sensors. However, without evidence, the model appears to have more difficulty predicting smaller episodes of congestion. While such behaviour is evident in all of the VISSIM scenarios to a limited extent, it is most noticeable in VISSIM Run 'B' where the model fails to predict a small traffic queue that forms upstream of the main congestion. Such results highlight the challenge of predicting unlikely events when evidence is not available. In some cases, the model may fail to predict the onset of congestion entirely, particularly if the probability of breakdown is low. In others, the model may predict the timing of flow breakdown incorrectly (i.e. where congestion occurs sooner than expected at a low probability of breakdown, or where congestion is delayed even though the probability of breakdown is relatively high). In cases of transient breakdown, the queue may dissipate before the effects can be reflected in the belief state.

queuing in heavily congested freeway corridors with no ramp metering, such as encountered in VISSIM Run ‘A’, unless evidence is available from traffic sensors.

Given the inability of the model to accurately predict the duration/extent of mainline queuing for VISSIM Run ‘A’, it is not surprising that the RMSE for this scenario declined significantly, from an average of 6.1 veh/km/lane with evidence, to an average of 13.1 veh/km/lane without evidence. For the less congested VISSIM scenarios, the drop in performance is less dramatic, since the assumed distribution for the congested capacity more closely matches the actual capacity observed in the simulation.<sup>24</sup>

- For VISSIM Run ‘B’ (moderate congestion), the average RMSE without evidence was 10.4 veh/km/lane, compared to an average of 5.6 veh/km/lane with evidence.
- For VISSIM Run ‘C’ (lowest level of mainline congestion), the average RMSE without evidence was 6.8 veh/km/lane, compared to an average of 5.0 veh/km/lane with evidence.

Again, 10 model runs were conducted for each VISSIM scenario. However, unlike the case with evidence, without evidence, the RMSE was virtually identical for all runs corresponding to the same scenario. As in the previous tests, 5000 particles were used to represent the belief state. However, similar results were obtained using only 1000 particles, based on trials involving VISSIM Run ‘A’.

While the results presented above suggest deficiencies in the ability of the Freeway Traffic Model to predict traffic operations in the absence of evidence, it is important to recognize that, in the ramp metering algorithm, the “no evidence” form of the Freeway Traffic Model will only be applied over a relatively short time horizon, limiting the opportunity for errors to be propagated far into the future. Indeed, the starting point for predicting future operating conditions will be based on the current traffic state as estimated from the Freeway Traffic Model with evidence. Since this latter model is much more accurate, the corresponding traffic predictions are also likely to be more accurate. Offsetting this improvement however, is a decrease in accuracy resulting from the need to estimate future demand inputs over the prediction horizon.

---

<sup>24</sup> For both scenarios, the performance of the Freeway Traffic Model in congested and uncongested conditions is similar to that described above for VISSIM Run ‘A’. The major difference lies in the ability to predict the duration/extent of traffic queues.

Figure 8-4 presents selected results from applying the Freeway Traffic Model without evidence over a six minute prediction horizon.<sup>25</sup> To obtain the results in Figure 8-4, the traffic model was first applied in tracking mode, using evidence from traffic sensors to update beliefs about the current freeway state. At pre-determined intervals, these updated estimates were then used as a starting point for predicting traffic operations over a six minute interval using the version of the model without evidence. It is these latter results that are provided in Figure 8-4. To distinguish between predictions, an alternating colour scheme has been applied.

The results in Figure 8-4 are based on two sets of model runs. In the first, the demand inputs over the prediction horizon were assumed to be known; in the second, the demand inputs were estimated based on past observations as described in Section 7.4.2. The second set of runs (shown in column 2) reflects the actual conditions under which the model will be applied in the ramp control algorithm. A comparison of the two scenarios provides an indication of the extent of error introduced by estimating the ramp and mainline demand.

As expected, the prediction model performs better when the demand inputs are known in advance. Under such conditions, the model is able to predict the number of vehicles per segment with reasonable accuracy for both congested and uncongested conditions, although it may have difficulty capturing unlikely events, such as flow breakdown which occurs at low probability.

When the demand inputs are not known in advance, but must be estimated from previous observations, the model predictions are somewhat less accurate. However, while the model is unable to capture the minute variations in cell density and speed which characterize freeway performance, the predictions are generally adequate for the purposes of establishing control parameters. From the results obtained, the model has the most difficulty predicting shockwave behaviour. On some segments (often located downstream of an on-ramp), the predicted queue density fluctuates wildly, with little resemblance to the actual density, particularly where demand projections have been utilized (see for

---

<sup>25</sup> The results in Figure 33 represent typical findings; since the model is stochastic, different results will be obtained with different model runs.

example Segment 22 in Figure 8-4). In others, the density within the queue seems to settle towards a relatively constant value, completely ignoring any variability, and resulting in similar predictions whether the demand is known in advance or not.

Since the ramp control algorithm is intended to reduce the extent of mainline queuing, the ability to predict queue growth is less important than the ability to predict the on-set of flow breakdown. Although not evident from Figure 8-4, it was observed that the model sometimes predicts congestion at the main bottleneck a few minutes earlier than actually occurs. As long as the time lag involved is relatively small, such behaviour is not considered a serious deficiency, since it allows the algorithm to take action to prevent breakdown before it occurs. However, while congestion may have been initiated later than predicted in the freeway scenario examined, the timing of flow breakdown is subject to uncertainty and it is entirely possible that a different outcome could be observed under nearly identical conditions. Of the alternatives, it is preferable for congestion to occur later than predicted, rather than sooner, otherwise the algorithm is left responding to congestion after the fact, rather than implementing pre-emptive counter-measures to prevent flow breakdown in the first place.

Given the uncertain nature of freeway flow, it is anticipated that the Freeway Traffic Model may sometimes miss important trends when working in prediction mode. For this reason, the control parameters are continually re-assessed as new evidence becomes available, reducing the impact of prediction errors. Nonetheless, it may be worthwhile to improve the demand prediction module as part of future work.

The six-minute model predictions presented in Figure 8-4 were developed using 800 particles. As discussed in Section 7.7.3, it was necessary to limit the number of particles used for prediction in the ramp control algorithm in order to reduce the time required to compute the control strategy. To determine the optimal number of particles for prediction purposes, a series of tests was carried out, each corresponding to a different initial freeway state, ranging from free flow to heavily congested. For each test, traffic predictions were developed using different numbers of particles to represent the freeway state. To assess the variability in the results, 10 model runs were carried out for each

particle scenario.<sup>26</sup> Variability was evaluated based on the standard deviation of the estimated freeway utility as computed within the ramp control algorithm. The results are presented in Appendix N. Based on a review of the findings, between 700 and 900 particles are considered a reasonable compromise between minimizing solution variability and maximizing calculation speed.

The results of the particle tests imply that the number of particles can be reduced significantly when working in prediction mode. Such findings are consistent with comments from the literature (see for example Kanazawa et al. 1995) which suggest that stochastic simulation is most effective when no evidence is available. When used strictly in predictive mode, fewer particles are needed to represent the system state since no particles are discarded due to inconsistency with the evidence.

Another factor influencing model performance is the prediction horizon. Based on the results obtained from the various model tests, a 6-minute prediction horizon is considered to be appropriate for use in the ramp control algorithm. With longer intervals, demand levels can change significantly, hindering prediction. Moreover, run-time begins to become an issue, since the solution algorithm applies the prediction model numerous times. Within the algorithm, the first minute of the prediction horizon is used to assess freeway performance for the current control interval. The impact of future control actions is thus evaluated over a 5-minute window. For the length of the corridor in question, this is generally considered adequate for the effects of control actions at one end of the corridor to be felt at the other within the time limits for prediction; unless the algorithm sees the benefit of a particular action within the prediction horizon, it will have no incentive to implement it.

---

<sup>26</sup> If the variability is too high, the algorithm will have difficulty determining which control solution is preferred (with high variability, there is greater risk that the true optimal solution will not have the maximum utility).

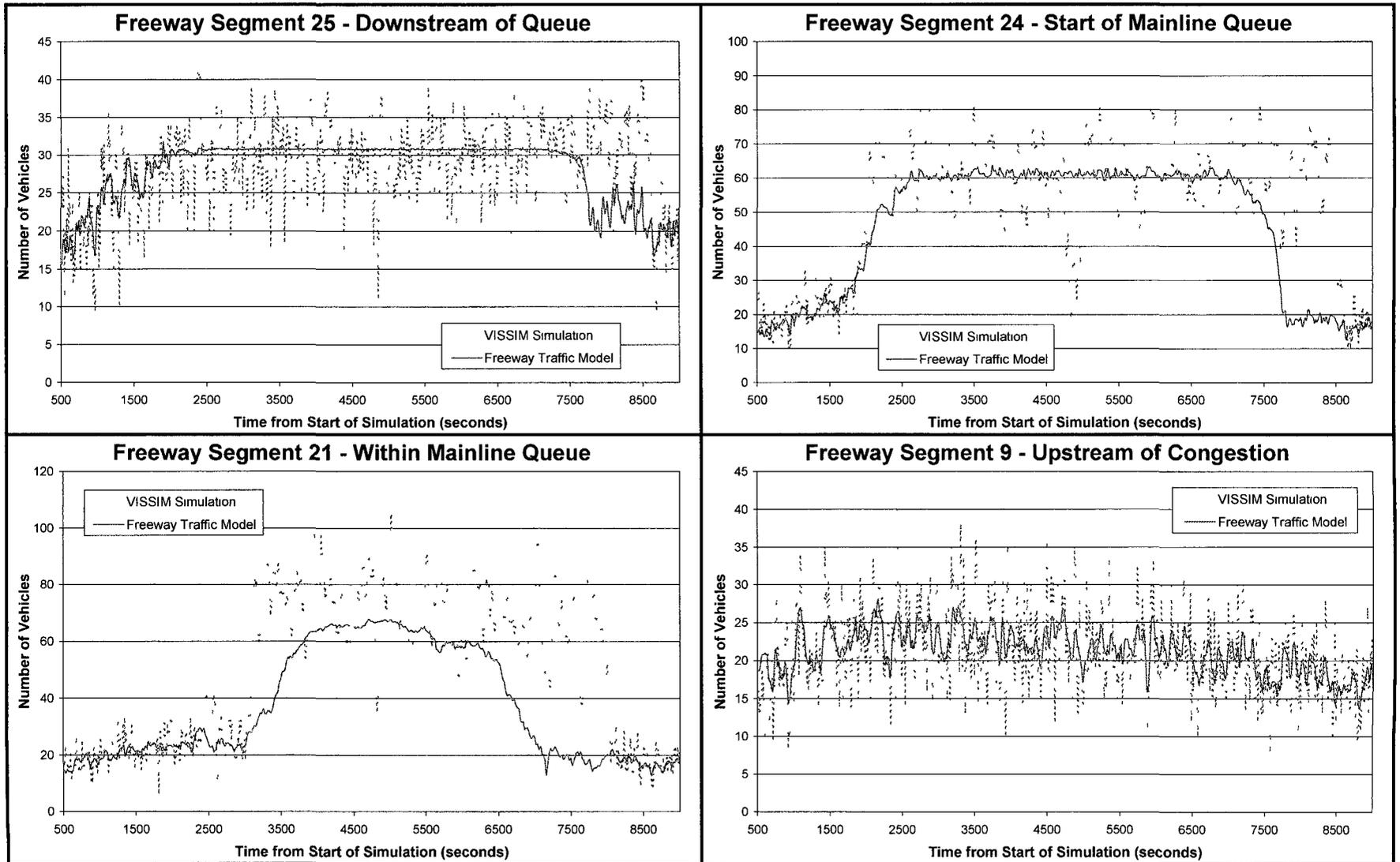
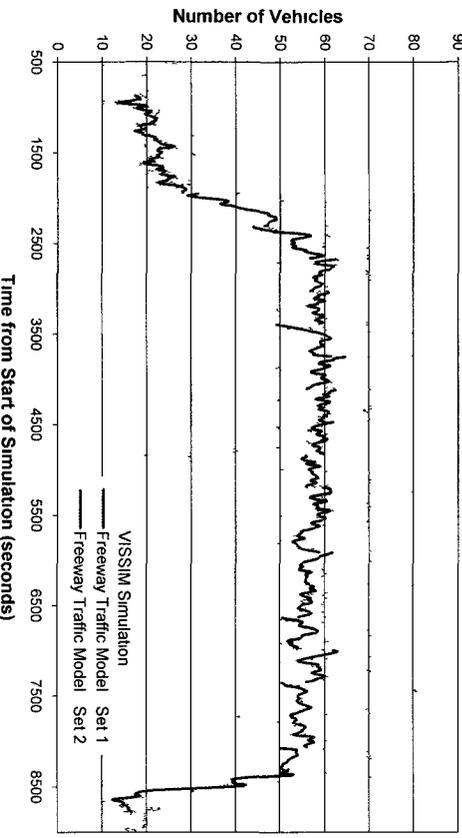


Figure 8-3 Model Performance Without Evidence – Selected Segments

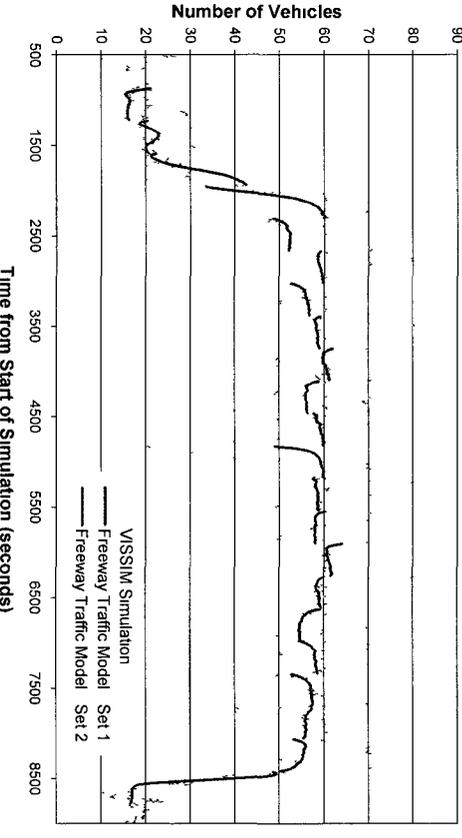
**DEMAND INPUTS ASSUMED TO BE KNOWN**

**Freeway Segment 24 - Start of Mainline Queue**

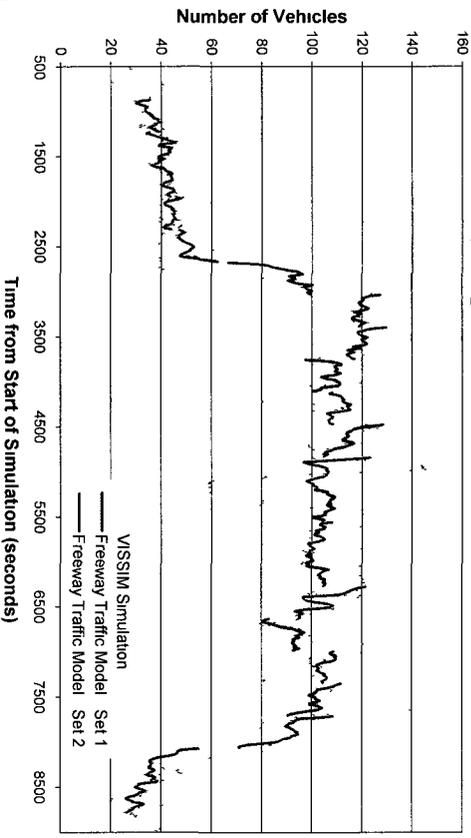


**DEMAND INPUTS ESTIMATED FROM PAST OBSERVATIONS**

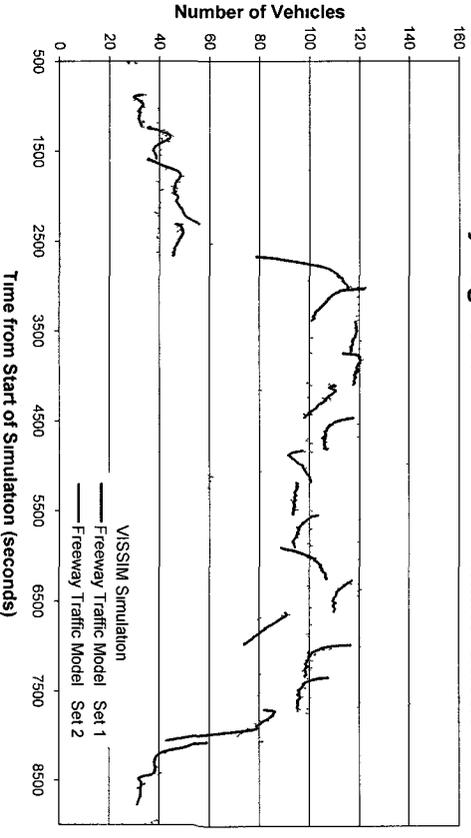
**Freeway Segment 24 - Start of Mainline Queue**



**Freeway Segment 22 - Within Mainline Queue**



**Freeway Segment 22 - Within Mainline Queue**



**Figure 8-4 Six-Minute Model Predictions Without Evidence – Selected Segments**

## 9 ALGORITHM PERFORMANCE UNDER TEST CONDITIONS

### 9.1 Overview of the Test Environment

The performance of the new algorithm was assessed using the VISSIM test network described in Section 8.2. Since the algorithm was developed in MATLAB, a methodology was required to integrate the algorithm with the VISSIM model. This was achieved by controlling the simulation in MATLAB via the VISSIM com interface.

Essentially, MATLAB allows the simulation to proceed at 20 second increments (i.e. the assumed data collection interval). At the end of each increment, MATLAB collects data from the VISSIM traffic sensors which are then used by the algorithm to update probabilities regarding the current freeway state. If the 20-second increment corresponds to the end of the control interval, MATLAB also passes new ramp metering rates to VISSIM to be implemented in the simulation. Thus, a two-way flow of information is achieved: VISSIM provides the traffic sensor data to the control algorithm in MATLAB, which in turn uses this data to estimate the control parameters. These parameters are then fed back into VISSIM for controlling freeway flow.

A more detailed description of the approach used to integrate VISSIM and MATLAB can be found in Figure 9-1. Figures 9-2 and 9-3 illustrate the key calculations carried out by the control algorithm and how each calculation fits with the flow of data to and from the freeway network. Figure 9-2 corresponds to the case with a real-world network, while Figure 9-3 applies to the VISSIM simulation environment. Although use of a simulation environment necessitates a sequential approach to software application, the flow of data between the algorithm and simulation environment is identical to what would occur if the algorithm was applied in real-time on an actual freeway.

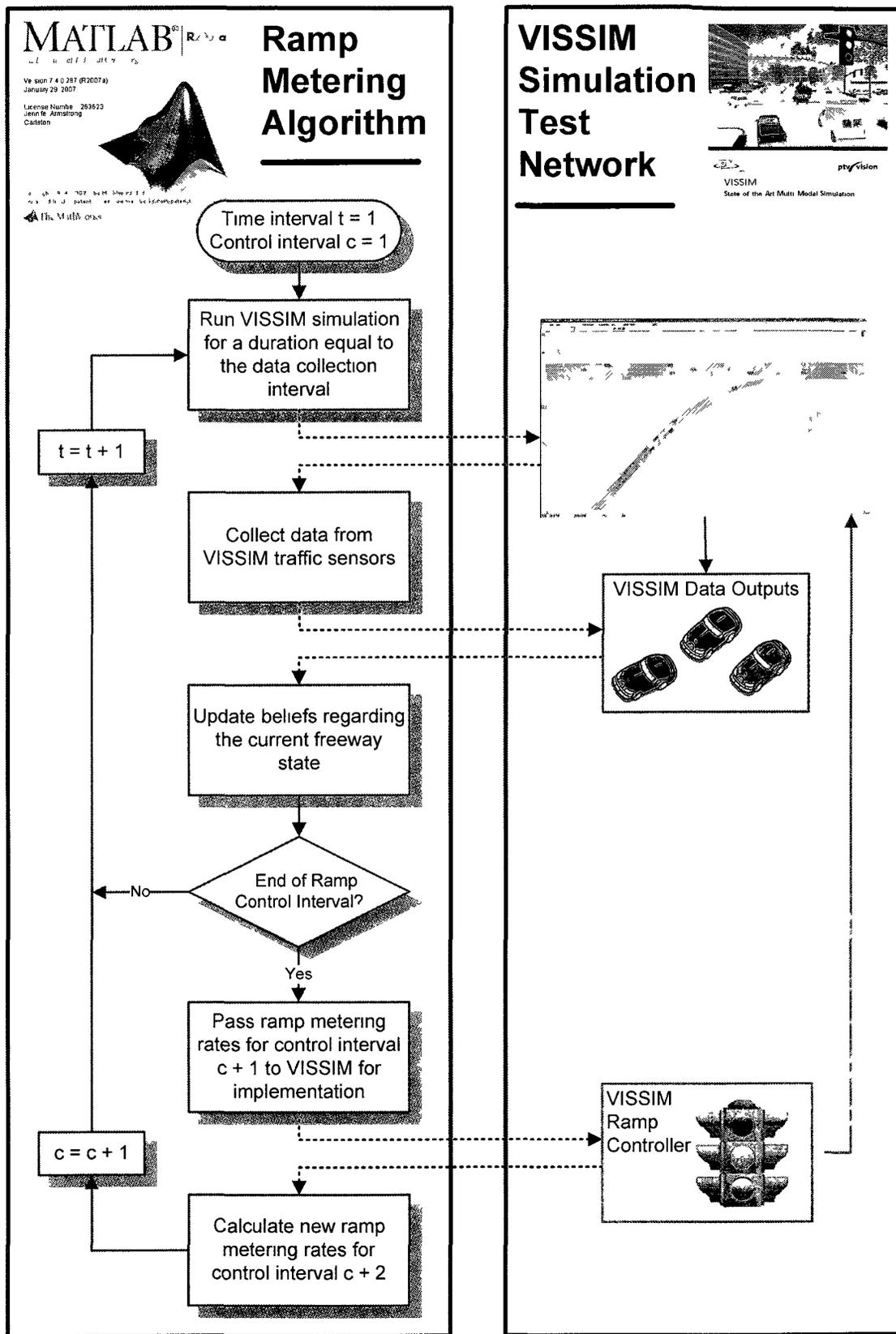


Figure 9-1 VISSIM & MATLAB Integration

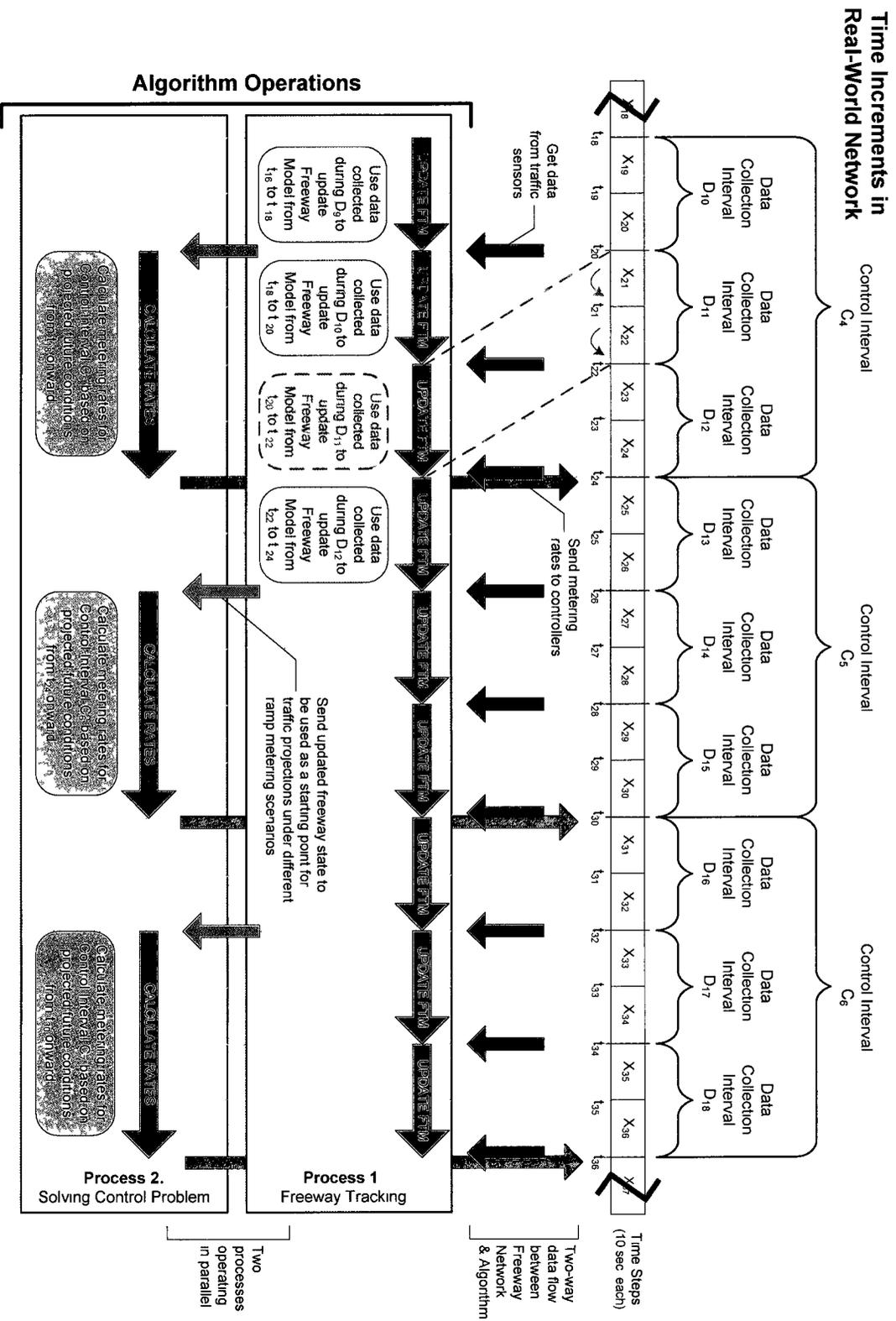


Figure 9-2 Algorithm Calculations – Real-World Network

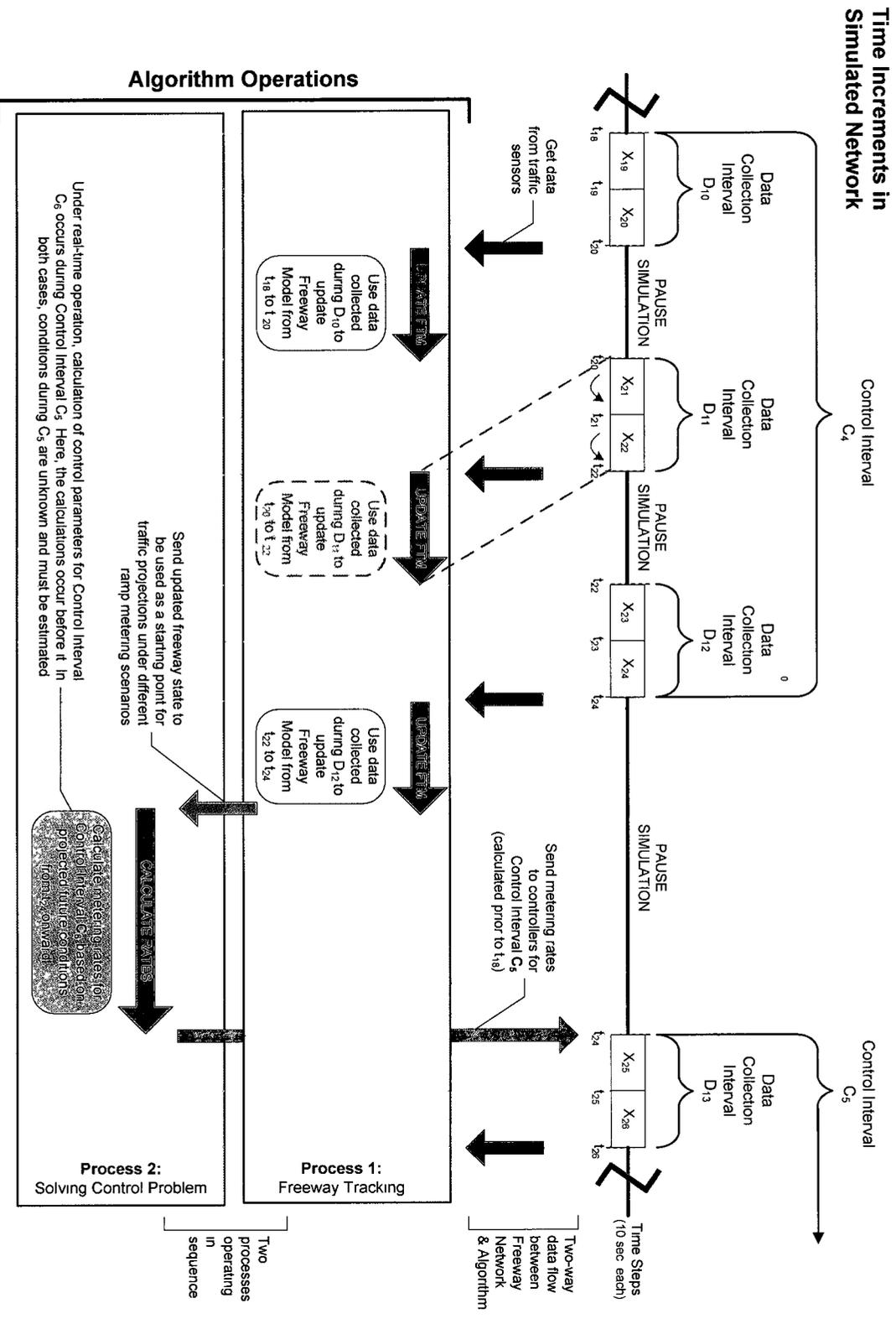


Figure 9-3 Algorithm Calculations – Simulated Network

## 9.2 Key Assumptions

In setting up the ramp control algorithm, there are a number of operational parameters which must be specified. To determine appropriate values for these parameters, a series of tests was carried out. Given the time required to carry out a single model run, and the random nature of the simulations (which means that several runs are needed to establish trends), tests were carried out as strategically as possible, honing in on areas of the algorithm likely to have the greatest impact on results.

Table 9-1 presents the final assumptions that were adopted for the algorithm tests described in the following sections.

**Table 9-1 Algorithm Assumptions**

<b>Parameter</b>	<b>Assumed Value</b>	<b>Comments</b>
Data collection interval	20 seconds	Established based on the characteristics of the COMPASS system in Toronto (Foo 2006)
Model update interval	10 seconds	Set so that no vehicle can cross more than one cell boundary in a single time step
Control interval	60 seconds	A shorter control interval would allow the algorithm to respond more quickly to prediction errors, but provides less time for solving the control problem (and increases model run-time considerably for testing purposes). A 60-second control interval was felt to be a reasonable compromise.
Prediction horizon	360 seconds	Note that this value includes 60 seconds for the current control interval. Since the ramp metering calculations for the next control interval take place during the current interval, and since operating conditions during the current interval are unknown, the current interval must be included in the prediction horizon. The selected prediction horizon is considered to be sufficiently long that the effects of metering can be felt throughout the corridor, but not so long as to unduly jeopardize run-time (and prediction accuracy).

Parameter	Assumed Value	Comments
Number of unique control intervals over the prediction horizon	1	While it is preferable to increase the number of control intervals with unique ramp metering rates, this increases the size of the solution space significantly, negatively impacting run-time. Since acceptable results could be achieved assuming constant ramp metering rates over the prediction horizon, this assumption was adopted. However, it is anticipated that better results could potentially be achieved if two unique control intervals were used, since this allows the algorithm to meter hard initially, and then back off as necessary to respect queue length constraints.
Maximum change in the metering rate from one control interval to the next	Unconstrained	With only 1 unique control interval, it is important that the algorithm have complete flexibility to tailor the metering rate to the anticipated demand. Moreover, it is less critical that a restriction be imposed since the size of the solution space has already been reduced considerably by limiting the number of unique control intervals to 1.
Number of particles used for tracking	5,000	Selected based on the off-line tests described in Section 8.
Number of particles used for prediction	800	Selected based on the off-line tests described in Section 8.

### 9.3 Implementation Issues

In implementing the new ramp control algorithm, a number of issues were encountered. Key issues and their resolution are described below.

- ***Dealing with queue spillback due to demand estimation errors***
  - In some cases, queue spillback was observed in the simulation due to errors in the demand estimation process. As a result, a mechanism was introduced to over-ride the calculated ramp metering rate if the ramp controller detects the presence of queue spillback at any point during the control interval.
  - The over-ride mechanism is triggered on the basis of the speed of the vehicles entering the ramp; if this speed drops below a pre-defined threshold (assumed to be 25 km/hr for the VISSIM test network), queue spillback is considered to be imminent, and action is taken to increase the metering rate to its maximum level.
  - It should be noted that the speed for triggering the over-ride mechanism is different from the speed used for defining queue spillback in the Freeway

Traffic Model. The latter is concerned with detecting when spillback is actually occurring, while the former is designed to prevent spillback before it occurs.

- ***Failure to meter ramps hard enough initially in order to avoid queue spillback later in the prediction horizon***
  - This issue arises due to the fact that the ramp metering rates are assumed to be constant over the prediction horizon (since there is only one unique control interval). With no opportunity to adjust the metering rate, there is a tendency for the algorithm not to meter hard enough initially in order to ensure that sufficient ramp storage is available at the end of the horizon.
  - In reality, although the optimal metering rate is calculated based on conditions over the full 6 minute horizon, it is only implemented over the next 60 second control interval before a new rate is introduced. Since the metering rate can be adjusted every minute, even if the algorithm meters hard in one 60-second interval, the rate can be reduced in subsequent intervals to prevent queues from spilling back.
  - To address this issue, a ramp storage adjustment factor was applied to allow the algorithm to think that more storage space is available than actually exists, thus encouraging the use of higher metering rates. Since the algorithm is able to adjust the ramp metering rate again in the next control interval, the risk of queue spillback is considered minimal. In the case of the VISSIM test network, a factor of 1.25 was found to work well.
- ***Turning the Meter Off***
  - It was found that when there is a low risk of mainline congestion, the algorithm sometimes has difficulty determining when it is appropriate to turn the meter off. For metering rates close to demand, there are essentially no ramp queues, however, a marginal improvement in freeway speed is sometimes observed due to lower freeway densities resulting from vehicles temporarily delayed at the ramp meters. As a result, the algorithm perceives a benefit in having the meter on.
  - To address this behaviour, a delay penalty was introduced whenever the ramp meter is in operation, reflecting the fact that, even when metering at demand (with minimal ramp queues), vehicles experience some delay simply by virtue of having to stop at the ramp meter.
- ***Identifying the Preferred Solution / Dealing with Random Variability when Solutions are Similar***
  - The stability of the results was also cause for concern. It was found that different scenarios often have very similar utilities, making it difficult to

determine which is best. Compounding the problem is the stochastic nature of the utility estimates; since the utility for a given scenario is likely to vary slightly from one run to the next, the optimal solution in one test may be different from the optimal solution in another. In some cases, it is suspected that this causes the algorithm to cycle between solutions during the iteration process.

- To address this issue, minor adjustments were made to the utility functions to help the algorithm better distinguish between scenarios. In addition, a decision was made to round the utility values to the nearest thousandth, since a greater level of precision is not meaningful given the variability of the utility results.
- Where the utilities for two scenarios are identical, the algorithm chooses the solution with the least restrictive metering. This is accomplished by adding the sum of the metering rates divided by 100,000,000 to the estimated utility value. Unfortunately, this has the by-product of sometimes causing the algorithm to continue searching for a solution for longer than would otherwise be the case (especially if random variation in the utility estimates causes cycling between solutions).
- It was found that the algorithm has particular difficulty distinguishing between utility values when metering close to demand, since only a very few particles may predict ramp queues. Accordingly, a maximum ramp metering rate was established to avoid scenarios that essentially behave the same. This maximum rate was assumed to be equal to the average ramp demand plus 100 vph, reflecting the characteristics of the underlying demand distribution.

- ***Solution Speed***

- In real world operation, the maximum time allowed for computing the optimal solution cannot exceed the duration of the control interval. In the test environment, this constraint was relaxed since the primary focus of this phase of algorithm development was to assess the feasibility of the proposed approach. Nonetheless, a number of options were explored to increase the algorithm speed, if only to reduce the time required for the VISSIM tests.
- To reduce the size of the solution space to be searched, the maximum ramp metering rate for a given ramp was limited to the maximum ramp demand expected over the prediction horizon, as described above.
- In addition, lower bounds for the metering rates were determined so as to avoid queue spillback. This had the effect of reducing the number of ‘infeasible’ solutions that were examined.
- Other actions included: optimizing the MATLAB code, adjusting the settings in the pattern search algorithm (and developing custom patterns), modifying the solution precision, and selecting appropriate control parameters for items

such as the control interval, prediction horizon, number of particles, etc. (refer to Sections 7.7.3 and 9.2).

- In the final version of the algorithm, the full 2.5 hour simulation generally required approximately 3.75 hours of computer run-time. Of this 3.75 hours, roughly 15 minutes was required for running the traffic simulation in VISSIM. The remaining 3.5 hours were required for the ramp control calculations. With the first 500 seconds of the simulation used for loading the network, there are 141 60-second control intervals for which ramp metering rates must be determined. Dividing 3.5 hours by 141 gives an average of roughly 90 seconds per control interval for calculating the control parameters.<sup>27</sup> Since there is only 60 seconds actually available for this calculation (i.e. the duration of the control interval), the calculation speed would need to increase by roughly 50% to meet the real-time constraints of the problem. However, it should be emphasized that the above results correspond to average conditions; in many cases (particularly when the network was uncongested), a control solution could be found in only 40 to 50 seconds, in other cases, several minutes of computation time were required.

## 9.4 Algorithm Performance

### 9.4.1 Base Case with No Ramp Metering

Using the VISSIM test network, a base case scenario was developed to model freeway performance with no ramp metering in place. This scenario provides the benchmark against which all other scenarios were compared. To account for stochastic variation in the simulation results, a series of 10 simulation runs was carried out using different random seeds.

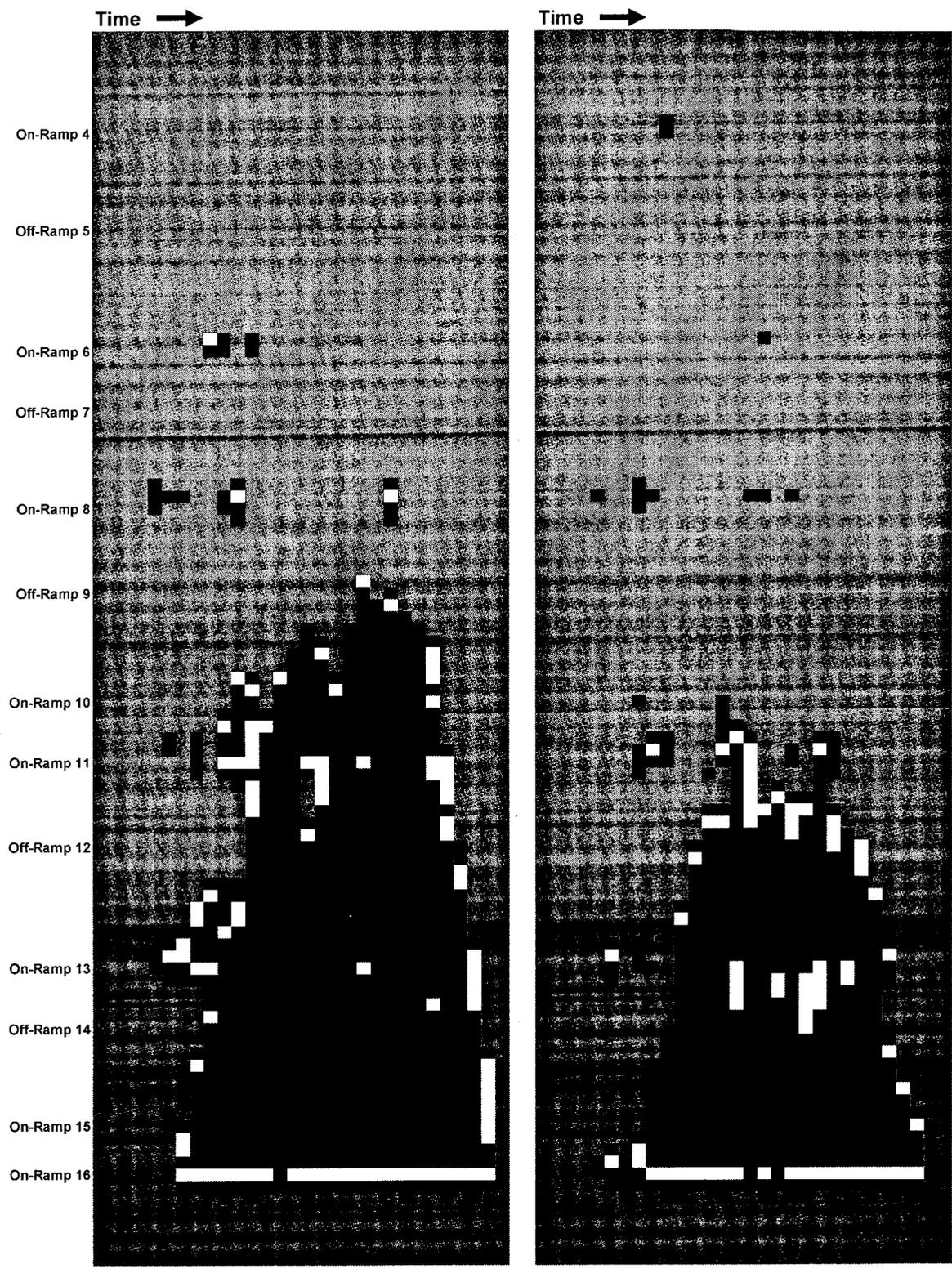
Table 9-2 presents a summary of the key performance statistics that were obtained for the network as a whole. This summary is based on vehicular activity over the entire 2.5 hour simulation period, excluding the first 500 seconds of the simulation while the network is being loaded.

---

<sup>27</sup> Note that the time required to compute the control solution should not be confused with the prediction horizon (which in this case was assumed to be 360 seconds). The prediction horizon indicates how far the traffic model looks into the future when predicting the effects of different ramp metering scenarios, whereas the control interval indicates how frequently the ramp metering rates get updated in the ramp controller (which in turn dictates the time available for computing the optimal solution).

While total travel time is often used as a measure of network performance, a decision was made to focus on the average network travel speed instead. The average travel speed is simply calculated as the total distance travelled by all vehicles in the network divided by the total travel time, and as such, accounts for any random variation in vehicular travel between different simulation runs. As Table 9-2 shows, the average network travel speed for the base case scenario was 72.8 km/hr. This value incorporates travel on both ramps and mainline segments, during periods of congestion as well as periods of free flow; at any given point in the network at any given time, the speed actually experienced by drivers may be substantially different. Generally speaking, a reduction in the duration/extent of mainline congestion will tend to increase the average speed – an increase that may be partially offset by any increase in ramp delay attributed to ramp metering.

Diagrams showing the magnitude and duration of mainline congestion for the base case scenario can be found in Figure 9-4. To provide an indication of the range of simulation results, the simulation runs with the highest and lowest congestion are presented. Within Figure 9-4, time is shown on the x-axis, while location is shown on the y-axis, providing an indication of both the temporal and spatial distribution of mainline queues. Each ‘cell’ in the diagram represents the average travel speed experienced in a 100 m freeway section over a 5 minute interval. Red cells indicate a freeway section with an average travel speed of 50 km/hr or less, while yellow corresponds to a speed of between 50 and 70 km/hr. Speeds between 70 and 90 km/hr are represented by blue, while speeds greater than 90 km/hr are shown as green. According to Figure 9-4, the maximum queue length on the freeway varies between roughly 3.4 km and 4.6 km with no ramp metering in place. Such results imply considerable room for improvement with the introduction of ramp controls.



Seed 1 - Highest Mainline Congestion

Seed 8 - Lowest Mainline Congestion

Figure 9-4 Congestion Maps – Base Case with No Ramp Metering

**Table 9-2 Network Performance Statistics – Base Case with No Ramp Metering**

Performance Measure	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average
# of vehicles that have left the network	21,337	21,330	21,395	21 336	21,332	21 330	21,324	21,343	21,338	21,368	21,343
# of vehicles in network at end of simulation	496	498	466	499	497	518	516	496	515	506	501
Total distance traveled [km]	148,956	149,027	149,284	149,167	149,016	149 023	148,956	148,874	148,896	149,093	149,029
Total travel time [h]	2193	2082	1969	2060	2036	1924	2073	1920	2076	2177	2051
Average speed [km/h]	67.9	71.6	75.8	72.4	73.2	77.5	71.8	77.6	71.7	68.5	72.8
Total delay time [h]	773	659	545	637	615	502	652	499	656	755	629
Average delay time per vehicle [s]	127	109	90	105	101	83	108	82	108	124	104
Number of stops	43,007	38,551	28,893	38,614	35,391	25,573	41,110	26,928	33,633	46 672	35,837
Average number of stops per vehicle	2.0	1.8	1.3	1.8	1.6	1.2	1.9	1.2	1.5	2.1	1.6
Total stopped delay [h]	14.9	12.9	9.0	14.0	11.4	7.9	15.9	8.9	10.8	16.4	12.2
Average stopped delay per vehicle [s]	2.5	2.1	1.5	2.3	1.9	1.3	2.6	1.5	1.8	2.7	2.0

\* Excludes the first 500 seconds of the simulation while the network is being loaded

#### 9.4.2 Utility based on Efficiency Only

While the new ramp control algorithm incorporates equity considerations directly, few other algorithms do. Since equity and efficiency objectives often conflict, it is unrealistic to expect the new algorithm to outperform other algorithms when assessed on an efficiency basis alone. Such comparison is only meaningful if the equity component of the new algorithm is weighted at zero. Accordingly, a utility weighting scheme was developed based on the sole objective of maximizing efficiency (while satisfying ramp queue length constraints). To test the impact of this scenario, a total of 10 VISSIM simulations were carried out.

The results of the model runs are presented in Table 9-3 in terms of overall network performance. From this table, it is clear that the new ramp metering algorithm had a positive impact on network performance. Comparing Table 9-3 with Table 9-2, the average network travel speed improved from 72.8 km/hr with no ramp metering to 82.3 km/hr with ramp metering, a gain of roughly 13%. This difference is statistically significant (probability < 0.001) based on the one-sided t-test.

Congestion maps for the model runs with the highest and lowest freeway congestion are presented in Figure 9-6 (with interpretation as described in Section 9.4.1). These maps show a significant reduction in the magnitude/duration of mainline queues compared to the “no metering” scenario. This improvement in freeway congestion comes at the expense of ramp delays. Figure 9-5 shows the average travel time at the various on-ramps along the corridor for different time intervals throughout the simulation. The results have been aggregated for all 10 simulation runs. In general, the results show the highest metering occurring at the four on-ramps closest to the bottleneck. At these ramps, the average travel time ranges from roughly 150 to 200 seconds per vehicle, with delays at Ramp 13 approaching 250 seconds per vehicle (most likely due to the lower on-ramp volume which allows more restrictive metering rates without the risk of violating ramp queue length constraints). From the results in Figure 9-5, it is clear that the efficiency-only utility function results in a distribution of ramp delay that is clearly inequitable – a distribution which could conceivably generate significant opposition from the travelling public.

In addition to ramp delay, it is also important to examine potential impacts to the arterial network due to queue spillback. In general, the results of the algorithm appear to be acceptable. Spill-back does occur on occasion, however, each spillback event is typically less than one minute in duration, with maximum queues exceeding the available storage by less than 100 m. Ramp 16 (at the start of the bottleneck) does show more significant spillback in two or three model runs, however, the worst episode lasts only 8 minutes (out of a total 2.5 hour simulation), and there are only 5 instances (out of all 10 simulation runs) where the spillback exceeds 100 m.

Although the ramp control algorithm was able to prevent queue spillback in all but a few minor cases, the algorithm would likely perform better if the ramp demand predictions were more accurate. Any errors in the estimated ramp demand impact the queue length calculations, which in turn influence the selection of the ramp metering rates. This is particularly true at on-ramps where the maximum queue length is being constrained by the ramp storage. If the demand is over-estimated, the algorithm may allow more vehicles onto the freeway than otherwise necessary in an effort to prevent spillback. If the demand is under-estimated, the resulting vehicle queue may exceed the available storage.

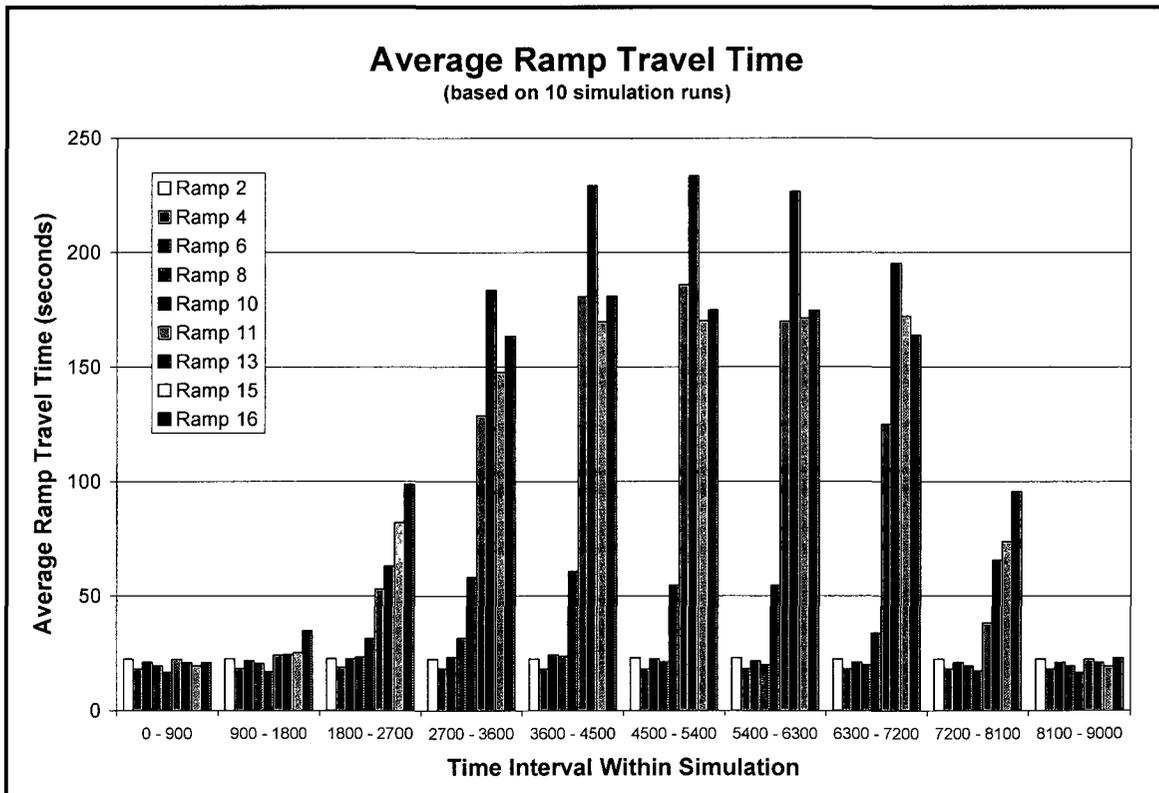
While the use of a ramp storage adjustment factor can mitigate the potential for spillback (by ensuring that a few extra storage spaces are held in reserve in case the demand is higher than expected), under such an approach, the ramp storage will only be used to its full capacity when the demand error is at its highest level, implying less efficient use of the ramp storage at other times. This inefficiency is reflected in the queue length results at the ramps being metered the hardest. Rather than being able to maintain the queue at a relatively stable level (just under the storage capacity), the queue length fluctuates significantly, with the algorithm letting too many vehicles onto the freeway at some times, and too few vehicles on at other times due to demand predictions that proved to be invalid. From these results, improvements to the demand prediction module would seem to be warranted; by reducing the demand prediction error, the algorithm should be able to maintain more stable ramp queues with fewer spillback episodes and more efficient utilization of the available storage.

Overall, the results of the algorithm tests suggest that the algorithm is generally working as intended. By metering traffic at the ramps closest to the bottleneck, an improvement in system-wide performance is achieved. Given these results, it would appear that the underlying Freeway Traffic Model is performing well in an on-line mode. This was confirmed by examining selected outputs from the Freeway Traffic Model collected during the algorithm tests and comparing them to the actual conditions observed in the VISSIM simulation. Sample results for a typical model run can be found in Appendix O.

**Table 9-3 Network Performance Statistics – New Algorithm (Efficiency Only)**

Performance Measure	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average
# of vehicles that have left the network	21,343	21,335	21,347	21,358	21,375	21,362	21,373	21,320	21,357	21,354	21,352
# of vehicles in network at end of simulation	503	525	494	478	493	484	507	524	470	494	497
Total distance traveled [km]	149,075	148,743	149,095	149,059	149,069	148,942	149,030	149,091	148,860	149,123	149,009
Total travel time [h]	1886	1839	1758	1748	1721	1938	1700	1832	1706	2023	1815
Average speed [km/h]	79.1	80.9	84.8	85.3	86.6	76.8	87.6	81.4	87.2	73.7	82.3
Total delay time [h]	463	419	336	326	299	517	277	409	286	599	393
Average delay time per vehicle [s]	76	69	55	54	49	85	46	67	47	99	65
Number of stops	43,836	39,258	33,764	33,507	30,182	47,298	27,103	40,023	28,392	53,908	37,727
Average number of stops per vehicle	2.0	1.8	1.5	1.5	1.4	2.2	1.2	1.8	1.3	2.5	1.7
Total stopped delay [h]	48.2	42.1	38.0	38.1	36.4	48.3	32.6	46.8	34.5	51.8	41.7
Average stopped delay per vehicle [s]	7.9	6.9	6.3	6.3	6.0	8.0	5.4	7.7	5.7	8.5	6.9

\* Excludes the first 500 seconds of the simulation while the network is being loaded



**Figure 9-5 Average Ramp Travel Time – New Algorithm (Efficiency Only)**

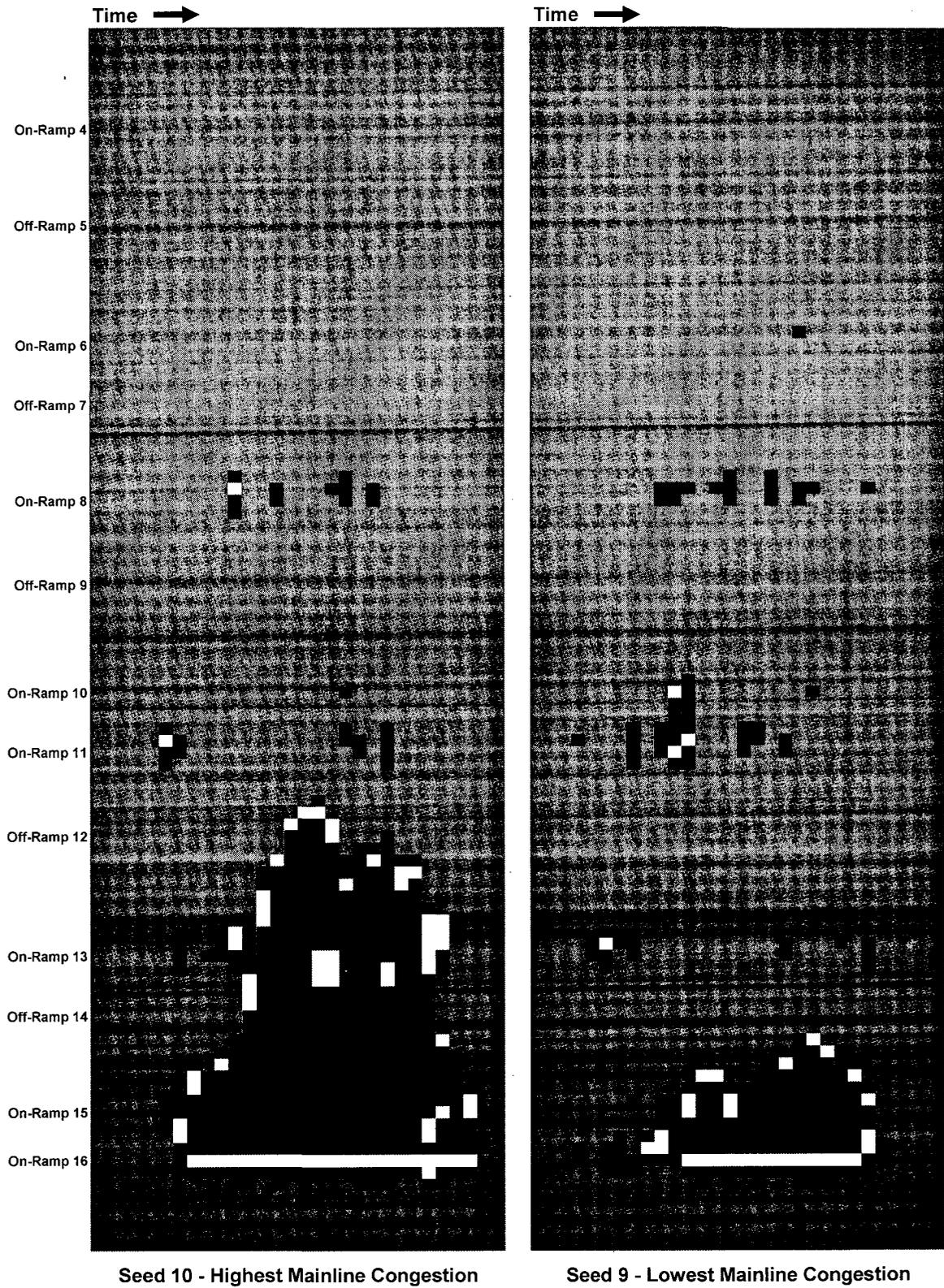
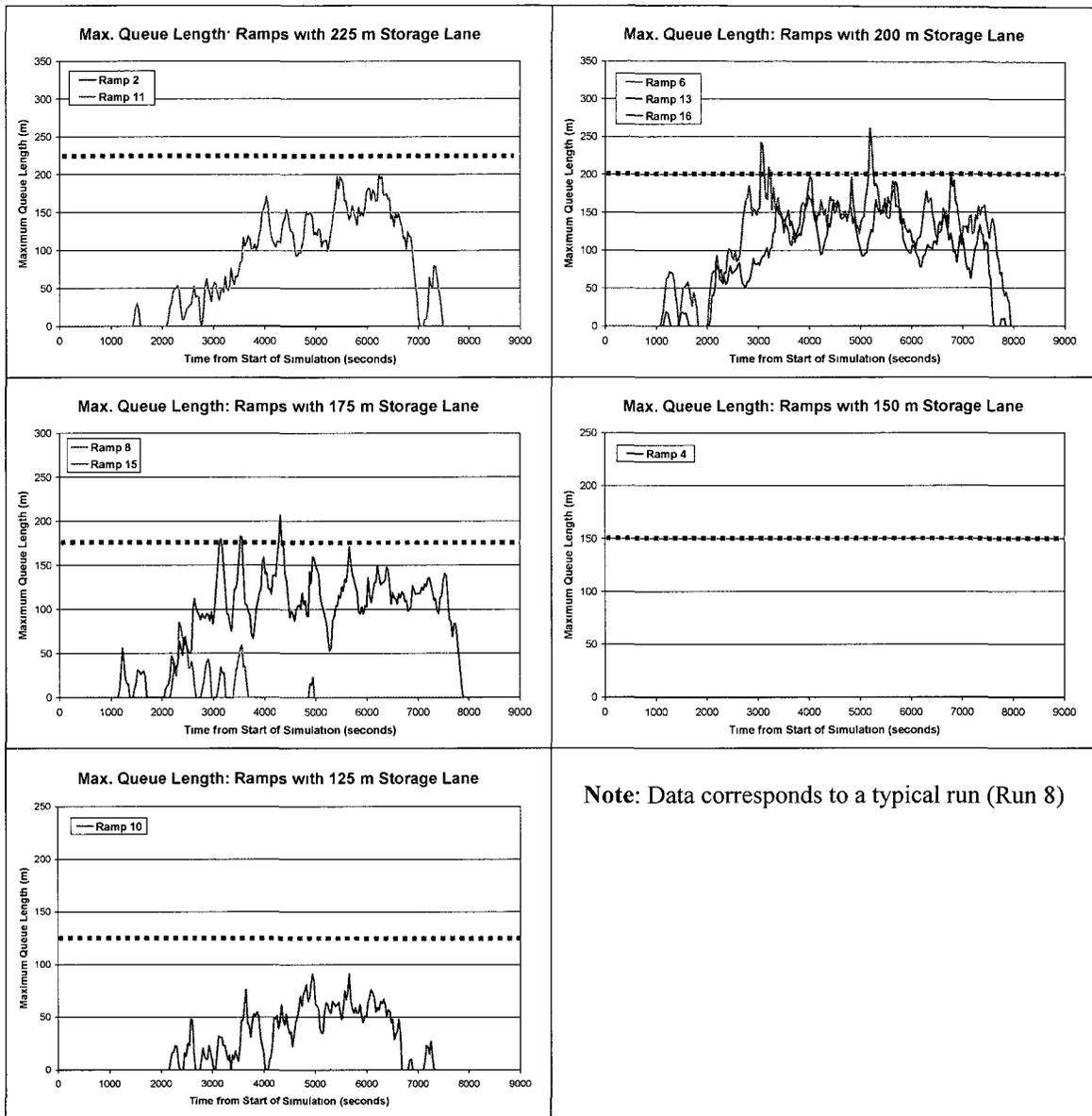


Figure 9-6 Congestion Maps – New Algorithm (Efficiency Only)



**Figure 9-7 Adequacy of Ramp Storage – New Algorithm (Efficiency Only)**

### 9.4.3 Utility based on Efficiency and Equity

Results of the ramp control algorithm with both efficiency and equity objectives included in the utility function can be found in Table 9-4 and Figure 9-8. These results are based on a total of 10 simulation runs. From Table 9-4, it can be seen that the average network travel speed (78.4 km/hr) has declined compared to the efficiency-only scenario (82.3 km/hr), but is still higher than the no control case (72.8 km/hr). On a percentage basis, the efficiency-only scenario was able to achieve a 13% improvement in network performance

compared to the base scenario, while the efficiency+equity scenario achieved only an 8% improvement (all differences are statistically significant). Given these findings, no congestion maps have been prepared for the efficiency+equity scenario since the performance can generally be expected to be between that of the no metering scenario and the scenario with the algorithm maximized for efficiency only.

While the efficiency+equity scenario is not able to achieve the same benefit in network performance, the trade-off is a significant improvement in equity. Whereas some ramps were experiencing average travel times in excess of 200 seconds per vehicle under the efficiency-only scenario, with equity included in the objective function, the maximum ramp travel time drops to just over 100 seconds per vehicle.<sup>28</sup> The improvement in equity is most evident by comparing Figure 9-8 with Figure 9-5. From a review of Figure 9-8, not only have the maximum ramp travel times declined compared to the efficiency-only scenario, but qualitatively, travel times in general appear to exhibit much lower variation between ramps.

To further assess the improvement in equity, a ratio was computed of the highest average ramp travel time in any 15-minute period to the lowest average ramp travel time in the same period. This ratio is analogous to the equity index developed by Meng and Khoo (2010), with the ramp group defined as all ramps along the corridor. Given its formulation, the ratio provides an indication of the relative difference in operating characteristics at the 'best' ramp compared to the 'worst' ramp, and thus provides a measure of the level of spatial inequity experienced by drivers accessing the two facilities. For the efficiency-only scenario, the ratio ranges from approximately 1.4 at the shoulders of the peak period to 12.8 when demand is heaviest. In comparison, the efficiency+equity scenario has a travel time ratio which ranges between 1.4 during the shoulder period to only 2.6 during peak conditions, a significant improvement compared to the efficiency-only case. Again, these findings support the conclusion that the efficiency+equity scenario has a more equitable distribution of ramp delay, which may be more palatable to drivers.

---

<sup>28</sup> The ramp travel time statistics quoted in this section are based on the average results from the 10 simulation runs carried out for each scenario. All figures showing ramp travel times are likewise based on average results.

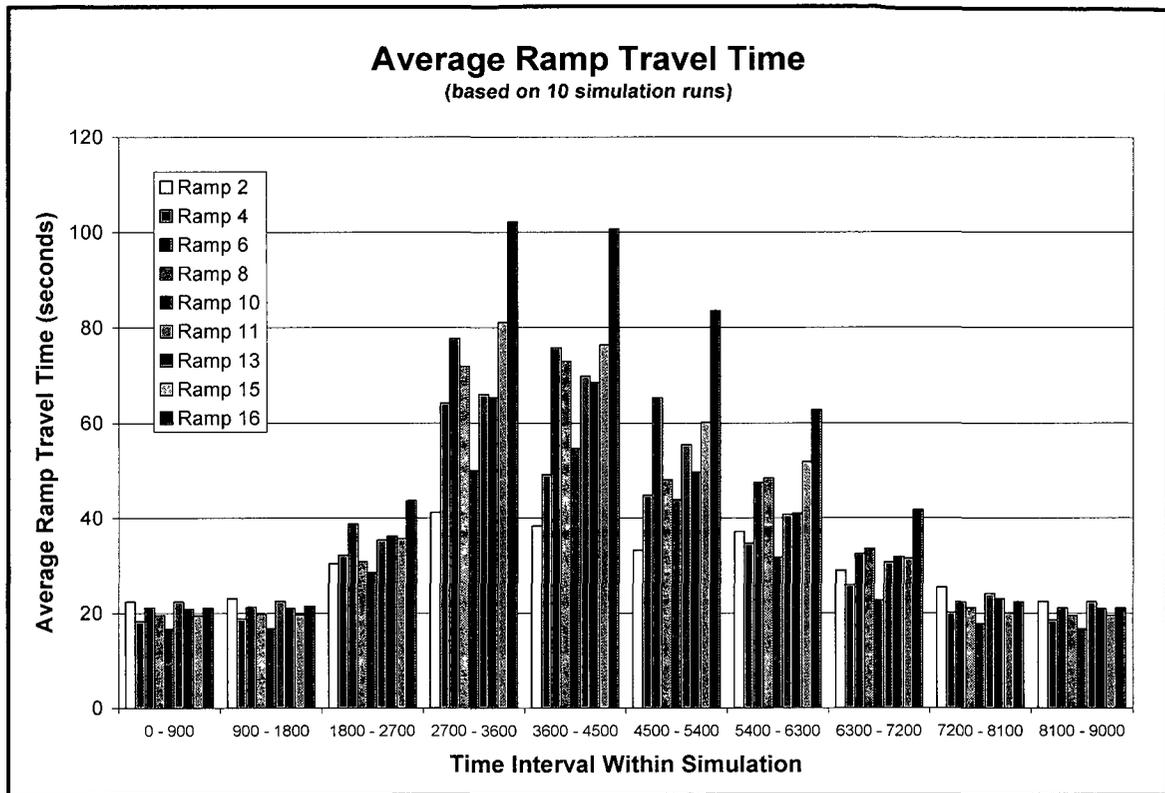
In terms of impacts to the arterial network, queue spill-over is all but eliminated in the efficiency+equity model runs. As a result, no exhibits have been provided showing the ramp queues.

It should be emphasized that the results presented in this section correspond to the specific utility weighting scheme described in Section 7.6. Different weighting schemes will produce different results which may be more or less “optimal” depending on the viewpoint of the individual and the relative importance placed on efficiency and equity objectives.

**Table 9-4 Network Performance Statistics – New Algorithm (Efficiency + Equity)**

Performance Measure	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average
# of vehicles that have left the network	21,323	21,343	21,347	21,331	21 374	21,341	21,334	21 334	21,394	21,343	21,346
# of vehicles in network at end of simulation	522	480	493	539	482	534	499	495	467	496	501
Total distance traveled [km]	149,103	148,742	149,041	148,864	149,153	148,804	148,947	149,027	149 283	149,167	149,013
Total travel time [h]	1821	1995	1928	1841	1966	1885	2046	1788	1811	1956	1904
Average speed [km/h]	81 9	74 6	77 3	80 9	75 9	78 9	72 8	83 4	82 4	76 3	78 4
Total delay time [h]	398	575	506	420	543	466	625	365	387	533	482
Average delay time per vehicle [s]	66	95	83	69	89	77	103	60	64	88	79
Number of stops	28,706	41,462	32,873	28 403	39,632	32,553	51,403	22,206	25,621	40,951	34,381
Average number of stops per vehicle	1 3	1 9	1 5	1 3	1 8	1 5	2 4	1 0	1 2	1 9	1 6
Total stopped delay [h]	19 4	24 6	17 7	19 0	26 2	22 8	33 6	14 5	19 0	28 8	22 6
Average stopped delay per vehicle [s]	3 2	4 1	2 9	3 1	4 3	3 8	5 5	2 4	3 1	4 8	3 7

\* Excludes the first 500 seconds of the simulation while the network is being loaded



**Figure 9-8 Average Ramp Travel Time – New Algorithm (Efficiency + Equity)**

### 9.5 Comparison with the ALINEA Ramp Control Algorithm

To assess the value of the new ramp control algorithm, it is useful to compare its performance against that of other, well-established algorithms. While most existing algorithms have been structured to maximize efficiency, the new ramp control algorithm also includes equity objectives, making direct comparisons impossible. However, by removing the equity dimension from the utility function for the new algorithm, the algorithm is essentially reduced to maximizing efficiency, and should therefore be capable of producing results that are comparable to other algorithms.

Accordingly, the ramp control algorithm ALINEA was selected for comparison against the new ramp control algorithm as applied based on efficiency criteria alone. The following sections provide a brief introduction to ALINEA, its implementation in the VISSIM test network, and the resulting operational performance in relation to the new algorithm.

### 9.5.1 Overview of the ALINEA Algorithm

ALINEA (asservissement linéaire d'entrée autoroutière) is a local, traffic-responsive ramp metering strategy which uses feedback control to determine on-ramp flows (Papageorgiou et al. 1991). Using real-time occupancy data as input, the control law attempts to maintain the freeway occupancy immediately downstream of the entrance ramp at some predefined target level such that congestion is avoided. If the observed occupancy is higher than the target occupancy, the control law acts to decrease the ramp flow; if the observed occupancy is lower than the target occupancy, the control law allows the ramp flow to increase. By reacting smoothly to even slight deviations from the target occupancy, traffic flow is stabilized around the target value (Smaragdis and Papageorgiou 2003). Thus, with the target occupancy set equal to the critical occupancy, the flow downstream of the ramp is maintained near capacity, maximizing freeway throughput.

Using ALINEA, the metering rate for each ramp is calculated at every control interval  $\mathbf{k}$  as follows:

$$r(\mathbf{k}) = r(\mathbf{k}-1) + K_R [\hat{o} - o_{out}(\mathbf{k}-1)]$$

where:

$r(\mathbf{k})$  is the ramp flow to be allowed onto the freeway over the next control interval  $\mathbf{k}$

$r(\mathbf{k}-1)$  is the ramp flow rate observed during the previous control interval  $(\mathbf{k}-1)$

$K_R > 0$  is a regulator parameter (also known as the control gain)

$\hat{o}$  is the target occupancy

$o_{out}$  is the current occupancy as measured during the previous interval

In the above equation, the ramp flows are measured in vph, which is also the units for  $K_R$ , while the occupancy is expressed as a percent. The duration of the control interval depends on the characteristics of the freeway system (such as the data collection frequency), but is often set to between 30 and 60 seconds. To convert the ramp flow into a cycle length for implementation within the signal controller, the following equation can be used:

$$\text{Cycle Length (seconds)} = 3600 / r(\mathbf{k})$$

To ensure the ramp metering rates are reasonable,  $r(\mathbf{k})$  is confined to the interval  $[r_{\min}, r_{\max}]$ , where  $r_{\min}$  corresponds to the most restrictive ramp metering rate allowed within the system, while  $r_{\max}$  corresponds to the flow with the meter turned off. In practical terms, it may be more convenient to first convert the metering rate into a corresponding cycle length, and then impose constraints on the minimum and maximum cycle length. Thus, for example, if the cycle length is allowed to vary between 4 seconds (corresponding to a ramp flow of 900 vph) and 15 seconds (corresponding to a ramp flow of 240 vph), a cycle length calculated to be greater than 15 seconds based on  $r(\mathbf{k})$  would be capped at 15 seconds. Likewise, a cycle length calculated to be less than 4 seconds would indicate that the meter should be turned off<sup>29</sup> (with the ramp flow limited by only the ramp capacity).

### 9.5.2 Rationale for Selecting ALINEA as a Basis for Comparison

Many different ramp control algorithms have been developed which could be used as a basis for comparison against the new algorithm developed in this research. Since it is only feasible to select one algorithm for comparison, it is important that the algorithm selected produce results that are representative of the other algorithms available. From a review of the various options, the ALINEA ramp control algorithm was considered to offer a number of advantages, as follows:

- ALINEA has been tested in simulation studies, and has also been implemented in the field with positive results (Papageorgiou et al. 1997)
- Other research studies involving the development of new ramp control algorithms have used ALINEA as a basis for comparison (Jacob and Abdulhai 2005; Sun and Horowitz 2006). By using the same ‘benchmark algorithm’, it is possible to get a sense of how two algorithms will perform relative to each other even if not compared directly.
- ALINEA is well-documented. All of the information required to implement the algorithm is readily available in the literature.

---

<sup>29</sup> Unless a pre-existing queue exists, in which case it may not be desirable to release all of the vehicles at once. In such situations, a minimum cycle length of 4 seconds could be used until the queue dissipates.

- As noted in Zhang et al. (2001), ALINEA is relatively easy to implement since the only parameters to estimate are the control gain (regulator parameter) and target occupancy.
- Results of comparison tests suggest that ALINEA performs as well as other prominent algorithms under certain conditions.

This latter point is particularly important. Since ALINEA is only a local algorithm, in theory, algorithms applying coordinated control should achieve superior results, particularly under heavy demand. In practice however, ALINEA has been found to perform as well as more sophisticated algorithms, at least in certain cases. For example, as part of a comparative evaluation of different ramp metering algorithms, Zhang et al. (2001) found no significant performance difference between ALINEA, Minnesota's Zone algorithm, Seattle's Bottleneck algorithm, and National Engineering Technologies' SWARM algorithm with one time-step ahead prediction. While such results support the use of ALINEA as a comparison benchmark, it should not be implied that ALINEA performs as well as all other ramp metering algorithms under all conditions.

For moderate congestion, Zhang et al. (2001) suggest that ALINEA is "effective, robust, and flexible" (pg. 8), however, the algorithm may not perform as well in heavy congestion where ramp queue constraints exist. Under coordinated control, such constraints are generally addressed by including upstream ramps in the metering scheme. However, with ALINEA, queue spillback is not considered directly but is instead typically handled by imposing an over-ride mechanism; upstream metering is not initiated until after mainline congestion has propagated back to the ramp in question. As a result, ALINEA may have difficulty balancing freeway and ramp objectives when demand is high (Zhang et al. 2001). Indeed, despite results to the contrary, Zhang et al. (2001) suggest that the coordinated algorithms included in their comparative evaluation could outperform ALINEA if some of the key parameters were better calibrated, or preferably, updated in real-time.

Notwithstanding the above caveats, ALINEA is believed to provide a suitable basis for comparison against the new ramp control algorithm, and was therefore implemented in the VISSIM test network.

### 9.5.3 Implementation in the VISSIM Test Network

To implement ALINEA, only minor changes to the VISSIM test network described in Section 8.2 were required. Detectors were added immediately downstream of each signal stop bar to estimate vehicle flows departing the meter. A detector was also installed at the entrance to each ramp to allow for ramp queue estimates to be developed. In addition, mainline detectors were added to the network to measure freeway occupancy directly downstream of each entrance ramp. For ALINEA to be effective, any traffic congestion triggered by on-ramp flow must be visible in the occupancy data (Papageorgiou et al. 1991). Accordingly, detectors were placed in the two mainline lanes closest to the freeway shoulder at approximately the midway point of the ramp speed change lane. This position should be adequate to quickly capture any traffic issues as they arise within the merge area.

In addition to the network changes described above, the signal control logic was also modified to incorporate the ALINEA control law for computing the ramp metering rates (and corresponding cycle lengths). In implementing the algorithm, a 30 second control interval was assumed.

To reduce the potential for queue spillback onto the arterial network, the signal controller includes a rather rudimentary mechanism for over-riding the calculated cycle length in the event that queue lengths become critical. The over-ride works by imposing a minimum cycle length of 4 seconds once the estimated queue reaches 75% of the available ramp storage.<sup>30</sup> The queue estimate is derived by keeping track of the vehicles entering and exiting the ramp storage area during each control interval as measured by the ramp detectors (assuming an initial queue of zero when the meters are turned off). In practice, occupancy readings from sensors positioned at the critical queue length are often used for detecting spillback, and are expected to yield similar results.

---

<sup>30</sup> Even when the over-ride is in effect, variability in vehicle arrivals may cause the ramp demand to temporarily exceed the release rate, causing the queue to increase. By setting the critical queue length to 75% of the ramp storage, a safety margin is provided to accommodate such variability, allowing the algorithm to prevent spillback before it actually occurs. On high-volume ramps, there is more fluctuation in on-ramp flows, and the available storage space tends to fill up more quickly, requiring action to be taken sooner. For this reason, the storage for Ramps 6 and 16 was set artificially low in the ramp controller to prevent spillback occurrence.

Over-ride mechanisms such as the one employed above may suffer from oscillatory behaviour which results as the ramp controller cycles between the minimum cycle length (which is triggered once the ramp queue reaches its critical length) and a much higher cycle length (which has been calculated by the control algorithm to prevent/minimize mainline congestion). While the over-ride rate is in effect, more traffic is allowed onto the freeway than is desirable from a flow breakdown perspective. As a result, once the ramp queue has dissipated to the point where the over-ride mechanism is no longer needed, even more stringent metering rates may be required to address mainline congestion. The use of such metering rates in turn causes the queue to increase even faster, which again triggers the over-ride mechanism causing the process to repeat.

Admittedly, more elegant queue control mechanisms have been developed which avoid this oscillatory behaviour (see for example Smaragdis and Papageorgiou 2003), however, the approach adopted was found to produce acceptable results using available sensor data and more sophisticated options were therefore not considered.

A VisVAP flow chart illustrating the ALINEA control logic can be found in Appendix M.

#### 9.5.4 Determination of Control Parameters

To implement ALINEA, two parameters must be estimated: the regulator parameter,  $\mathbf{K}_R$ , and the target occupancy,  $\hat{\mathbf{o}}$ .

- The **regulator parameter**,  $\mathbf{K}_R$ , controls the strength of the response; as  $\mathbf{K}_R$  increases, the greater the change in the metering rate from one control interval to the next, and the faster the reaction to deviations from the target occupancy. Comments from the literature suggest that performance of the ALINEA algorithm is not particularly sensitive to the value of  $\mathbf{K}_R$ . In Papageorgiou et al. (1991), the authors claim that from a theoretical perspective, results of the algorithm are insensitive over a wide range of  $\mathbf{K}_R$  values, and imply that real-life calibration of  $\mathbf{K}_R$  may not be strictly necessary. In Smaragdis and Papageorgiou (2003), the authors note that “the same value of  $\mathbf{K}_R$  has been used in all known simulation or field applications of ALINEA without any need for fine-tuning” (pg. 75). According to Papageorgiou et al. (1991), a value of  $\mathbf{K}_R = 70$  vph has been found to yield good results in real-life experiments. In other simulation studies,  $\mathbf{K}_R$  values of 200 vph have been used (Jacob and Abdulhai 2005, Zhang et al. 2001).

- The **target occupancy** is often based on the freeway's critical occupancy (i.e. the occupancy observed when the freeway is operating at capacity). By setting the target occupancy to slightly less than this critical value, the algorithm attempts to stabilize flows at levels as close to maximum as possible without triggering congestion.

In practice, the critical occupancy for a given freeway section can be determined using speed-flow-occupancy curves developed from loop detector data. However, given that the critical occupancy may not necessarily be the optimal value for the target occupancy used in ALINEA, a simulation-based approach was adopted to determine the optimal parameter values for both the critical occupancy, as well as  $K_R$ . By using these optimal values in the simulation runs, the resulting performance outputs represent the upper range of what can realistically be achieved using ALINEA for the highway section in question.

An Excel macro was written to loop through various combinations of parameter values. For each iteration, the macro updates the ramp controller VAP file, initiates the VISSIM simulation, and stores the results of the simulation run. To account for random effects, 10 simulations were carried out for each unique combination of parameters by changing the random seed. The measure of effectiveness used to assess performance was taken as the average system travel speed over the entire simulation run (minus the initial 500 seconds used for loading the network), calculated as the total vehicle-kilometers of travel divided by the total vehicle-hours of travel. This MOE captures the effects of both freeway congestion and ramp delays, but unlike total system travel time, is not affected by any random variations in travel demand between the different simulation runs.

- $K_R$  values ranging from 50 vph to 110 vph were tested. For set values of the target occupancy, a series of two-sample t-tests were conducted to assess the statistical significance of differences in performance outcomes between different values of  $K_R$  (since no hypothesis could be made about the direction of the difference, the two-tailed test was applied). It was found that none of the observed differences were statistically significant at the 5% level (or even the 10% level in the vast majority of cases), implying that results of the ALINEA algorithm are insensitive to  $K_R$  over the range of values tested – a finding that is not altogether surprising given the discussion above. In light of these results, a decision was made to use a value of  $K_R = 70$  vph in the ALINEA algorithm, consistent with the value cited by Papageorgiou et al. (1991).

- Initially, target occupancies ranging from 18% to 26% were tested.<sup>31</sup> Based on the results, it was found that the average system travel speed was increasing with increasing values of the target occupancy. Accordingly, additional tests were conducted for occupancies ranging from 26% to 58%. From these tests, it was determined that network performance generally stabilized once the target occupancy reached 28%; for target occupancies greater than this value, no further improvement in average speed was evident, however, no decline in operational performance was observed either.

Given the results for  $K_R$ , the effect of different target occupancies was examined in detail for the case where  $K_R = 70$  vph (refer to Table 9-5). As before, two-sample t-tests were carried out to assess statistical significance. For target occupancies of 28% or higher, no significant difference in operational performance was found at the 5% significance level. However, a statistically significant difference ( $\alpha = 0.05$ ) was observed when comparing the performance results for the 26% scenario with scenarios involving higher target occupancies. Since the best results were associated with target occupancies in the higher range (i.e. 28% or greater), any of the occupancies in this range could be selected as the optimal value.

Before deciding on a particular target occupancy for use in the ALINEA algorithm, it is worth understanding why target occupancies above a certain threshold (i.e. 28%) all yield similar results. It is hypothesized that this behaviour is a direct result of the queue length constraints imposed within the network. Given these constraints, ALINEA is unable to prevent the onset of congestion at the freeway bottleneck. If flows peak rapidly prior to such congestion, there is only a short transition period where metering can be effective at preventing breakdown. Assuming such conditions prevail in the test network, similar results will be achieved whether metering is triggered prior to breakdown (at a target occupancy near the critical value), or immediately after breakdown occurs (as the occupancy spikes much higher).

While applying a target occupancy greater than the critical occupancy was found to have negligible impact at the bottleneck modelled in the VISSIM test network, in general, it is not desirable to do so in case metering is required at other ramp locations where breakdown can be avoided if metering is initiated soon enough. It is not known with certainty whether the 28% occupancy value at the lower end of the optimal range

---

<sup>31</sup> A number of lower target occupancies were tested to assess whether there would be any benefit to initiating metering at upstream ramps earlier than would otherwise occur in order to reduce flows through the bottleneck. However, no such benefits were observed.

represents the critical occupancy or a slightly lower value. In either event, it was considered to be the most appropriate value for use in the ALINEA algorithm and was therefore selected for implementation. Interestingly, this value is higher than the target occupancies used in other simulation-based applications of ALINEA (Jacob and Abdulhai 2005, Zhang et al. 2001) but is similar to the value cited by Papageorgiou et al. (1991) which was used in a real-world application of ALINEA in Paris.

**Table 9-5 ALINEA Results Under Different Target Occupancies ( $K_R = 70$  vph)**

Target Occupancy (%)	Average Network Travel Speed (km/hr)										
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average
26	76.8	76.8	82.5	75.0	82.7	76.9	76.3	74.5	73.4	75.7	77.1
28	75.4	88.1	84.6	82.9	80.8	78.7	85.6	76.2	81.0	77.9	81.1
30	80.7	89.8	75.5	82.0	85.3	79.4	82.3	80.0	72.3	83.6	81.1
32	84.2	80.9	74.4	77.4	81.1	82.8	83.4	85.4	84.6	84.4	81.9
34	75.5	78.6	76.4	78.0	86.5	78.4	88.6	84.4	76.7	84.4	80.7
36	81.0	78.9	79.5	77.0	81.0	83.2	80.8	84.8	76.6	86.8	81.0
38	71.5	89.7	77.0	82.1	86.7	81.4	85.5	81.6	76.5	85.3	81.7
40	76.5	82.6	78.3	74.5	87.3	78.7	80.8	80.6	81.1	80.7	80.1
42	74.3	86.1	76.7	83.4	78.8	90.7	84.6	81.7	78.2	84.8	81.9
44	83.1	79.9	88.1	79.8	79.8	76.3	80.9	79.7	89.3	90.4	82.7
46	76.9	85.1	82.5	91.2	89.5	78.8	85.3	82.5	87.4	78.4	83.7
50	72.5	85.8	84.5	85.5	82.0	84.4	87.0	79.1	80.2	81.2	82.2
54	79.7	83.9	80.4	79.9	86.8	81.3	87.5	90.9	83.4	79.0	83.3
58	80.2	88.2	84.4	81.7	92.0	79.5	86.5	87.1	83.1	78.0	84.0

**Notes:**

<sup>1</sup> All tests correspond to the case with  $K_R = 70$  vph

<sup>2</sup> Average network travel speed calculated for the full 2.5 hr simulation, minus the initial 500 seconds while the network is being loaded

As the above discussion has illustrated, the ALINEA algorithm is largely insensitive to the values selected for the target occupancy and regulator parameter, as long as the values lie within a certain range. These findings are generally consistent with Zhang et al. (2001), which, after examining various alternatives for the critical occupancy and regulator gain, concluded that there is a wide range of parameter values over which ramp metering performance does not change significantly.

### 9.5.5 Network Performance Under ALINEA Control

The following Tables and Figures (Table 9-6, Figures 9-9 to 9-11) provide a summary of the results of the ALINEA algorithm, based on the optimal parameter values from Section 9.5.4 above (i.e. target occupancy = 28%,  $K_R = 70$  vph). Interpretation of the

figures is as described previously in Section 9.4.1. To account for random variation, a total of 10 simulation runs were carried out.

From a review of the ALINEA results, the following conclusions can be drawn:

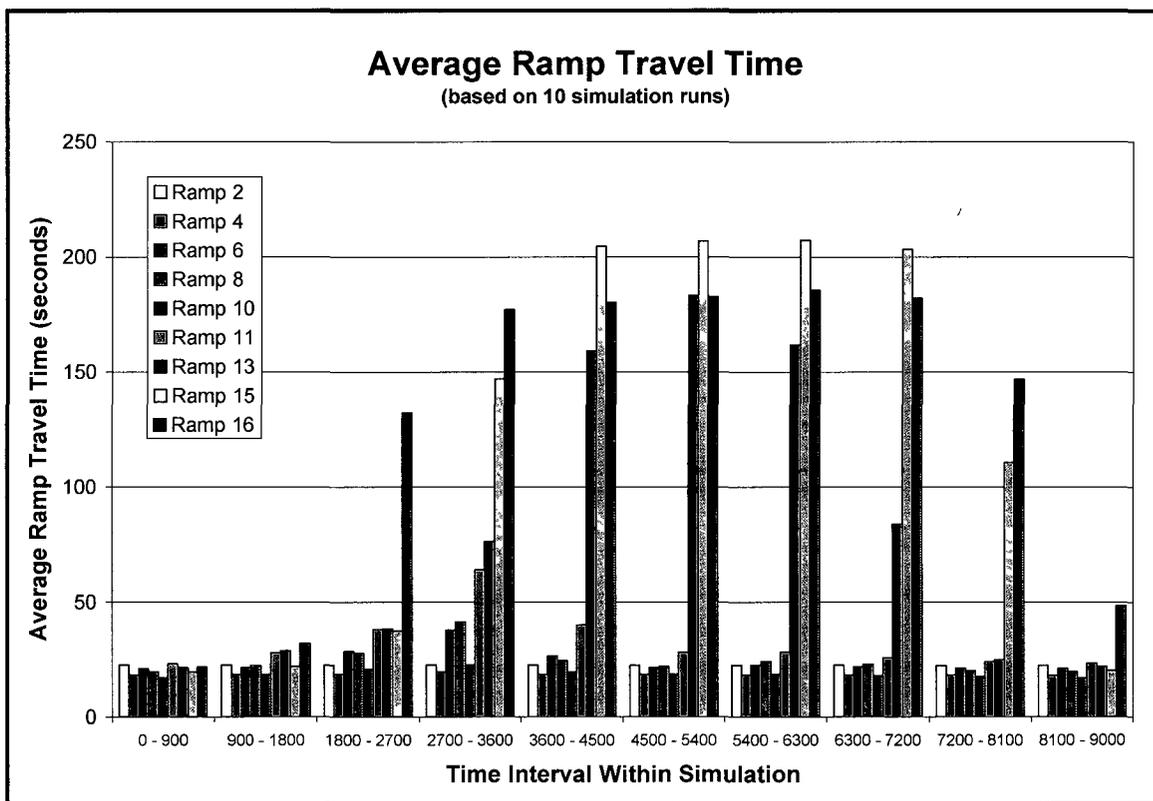
- ALINEA is effective at improving the operational performance of the freeway network. Compared to the “no control case” described in Section 9.4.1, the average network travel speed improves by roughly 11%, from 72.8 km/hr to 81.1 km/hr, a difference which is statistically significant (probability  $< 0.001$ ).
- Under ALINEA control, metering is first initiated at the start of the freeway bottleneck (i.e. Ramp 16) and gradually progresses upstream as congestion spreads. The resulting distribution of ramp delay is not considered equitable, since only vehicles using the first few ramps upstream of the bottleneck experience delay (in the range of 150 to 200 seconds per vehicle during peak intervals).
- While ALINEA is effective at reducing overall system travel time, it does not completely eliminate mainline congestion.
- There is some variability in the results, suggesting that freeway performance may fluctuate somewhat from one day to the next even if demand levels are reasonably similar.
- The over-ride mechanism used to prevent queue spillback works reasonably well. While some spillback was observed (particularly at Ramp 16 at the start of the bottleneck where ramp demand is high), the duration of each spillback event was limited to less than one minute in most cases, with the extent of spillback generally less than 100 m.

For interests sake, performance of the ALINEA algorithm was also examined in detail for the case with the target occupancy set to 58%. Even though the overall system performance (as measured by the average network travel speed) is not statistically different from the 28% target occupancy runs, qualitatively, it would appear that using a target occupancy of 58% results in lower ramp queues, particularly at the ramps farthest from the bottleneck. It is hypothesized that lower ramp queues sometimes improve network performance, and sometimes cause network performance to deteriorate, depending on whether mainline congestion remains stable or increases under less restrictive metering. Overall, the two effects tend to cancel out over multiple runs, which is why the impact of lower ramp queues is not statistically significant.

**Table 9-6 Network Performance Statistics – ALINEA Algorithm**

Performance Measure	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average
# of vehicles that have left the network	21,335	21 335	21 391	21 334	21,340	21 324	21,321	21,340	21,336	21 365	21,342
# of vehicles in network at end of simulation	498	495	469	505	500	525	519	496	515	510	503
Total distance traveled [km]	148,949	149,029	149,281	149 166	149,007	149,017	148,954	148,856	148,921	149 092	149,027
Total travel time [h]	1976	1692	1764	1800	1844	1892	1741	1954	1839	1915	1842
Average speed [km/h]	75.4	88.1	84.6	82.9	80.8	78.7	85.6	76.2	81.0	77.9	81.1
Total delay time [h]	555	270	339	377	423	471	319	534	418	492	420
Average delay time per vehicle [s]	91	44	56	62	70	78	53	88	69	81	69
Number of stops	36,990	15,621	20,605	24,114	26,713	30,632	19,714	35,084	26,037	32,946	26,846
Average number of stops per vehicle	1.7	0.7	0.9	1.1	1.2	1.4	0.9	1.6	1.2	1.5	1.2
Total stopped delay [h]	53.8	24.0	29.2	35.5	38.4	43.6	29.5	50.4	40.9	46.7	39.2
Average stopped delay per vehicle [s]	8.9	4.0	4.8	5.9	6.3	7.2	4.9	8.3	6.7	7.7	6.5

\* Excludes the first 500 seconds of the simulation while the network is being loaded



**Figure 9-9 Average Ramp Travel Time – ALINEA Algorithm**

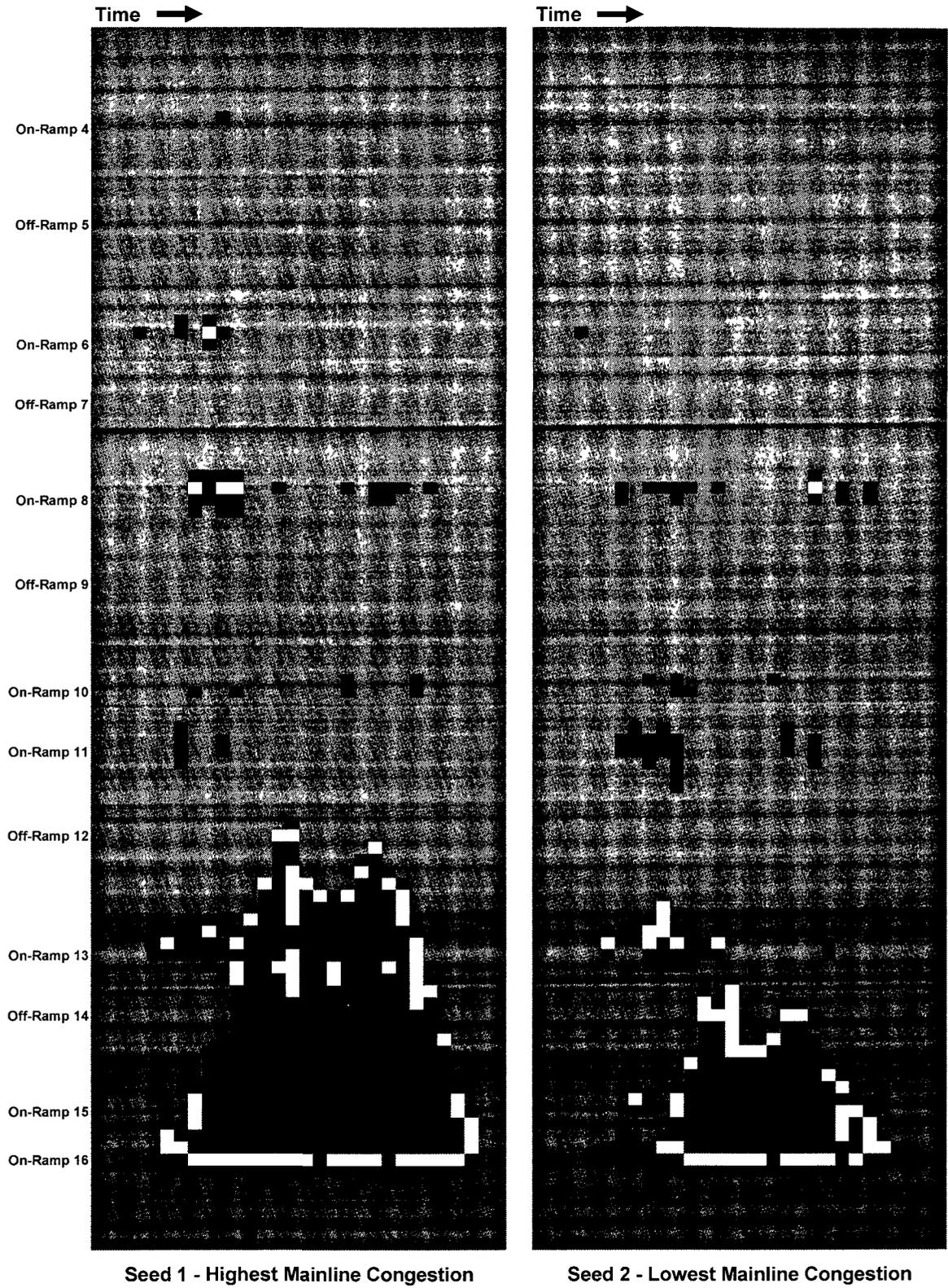
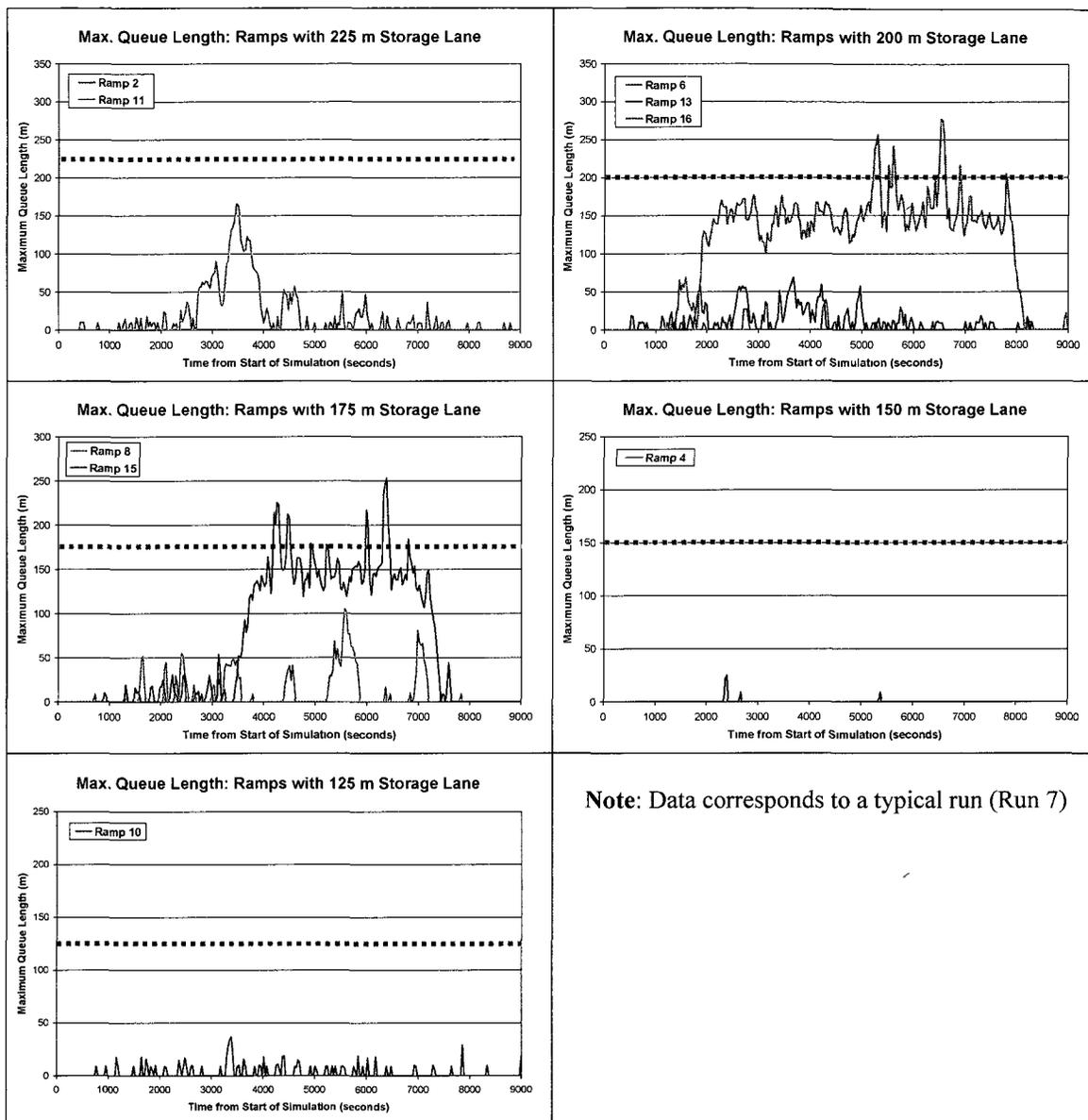


Figure 9-10 Congestion Maps – ALINEA Algorithm



**Figure 9-11 Adequacy of Ramp Storage – ALINEA Algorithm**

### 9.5.6 Comparison of Results

Table 9-7 presents a comparison of the efficiency performance of the ALINEA algorithm and the new ramp control algorithm developed in this research. As before, efficiency is measured in terms of the average network travel speed. It can be seen that although the ALINEA algorithm performs slightly worse than the new algorithm optimized for efficiency only, the difference is not statistically significant. This finding was anticipated; there was no expectation that the new algorithm would perform any better than existing

algorithms on an efficiency basis. The main improvement being investigated was the inclusion of equity in the algorithm. Nonetheless, it is important that the new algorithm be able to achieve results that are comparable to existing algorithms when equity objectives are omitted. Otherwise, the underlying foundations of the algorithm could be questioned. From the results presented, it appears that, when operating in an efficiency-only mode, the algorithm is able to perform as well as the ALINEA algorithm. Notwithstanding this conclusion, there is little value in implementing the new ramp metering algorithm if the only objective is to maximize efficiency, given the number of existing algorithms which already do this task quite well.

The real benefit of the new algorithm lies in its ability to account for equity objectives in calculating the control solution. The figures on average ramp travel time presented earlier for the ALINEA algorithm (Figure 9-9) and the new algorithm (Figures 9-5 and 9-8 for the efficiency-only and efficiency+equity scenarios respectively) clearly illustrate the ability of the new algorithm to manage freeway operations in a more equitable manner, albeit with a loss in overall network efficiency. This trade-off between equity and efficiency has been clearly documented in the literature (refer to Section 3.2.1); the results from the current investigation provide additional evidence that this trade-off exists.

A further comparison of equity impacts is provided in Figure 9-12. When equity is included in the utility function, the new algorithm is able to reduce the maximum travel time experienced at any ramp to just over 100 seconds per vehicle, compared to a maximum travel time of over 200 seconds per vehicle for the other efficiency-based algorithms. Moreover, the ratio of the maximum to minimum average ramp travel time during peak conditions declines substantially when equity is considered to just under 3, compared to 12.6 for the new algorithm maximized for efficiency, and 11.1 for ALINEA control.

It is interesting to note that both the ALINEA algorithm and the new algorithm optimized for efficiency do not completely eliminate mainline congestion. These results suggest that, for the network tested, the lowest system-wide travel time may not necessarily correspond to the case with no congestion. The fact that congestion remains may be a by-

product of the queue length constraints, which do not allow the algorithm to meter hard enough to prevent congestion; or it may simply be a reflection of the fact that beyond a certain threshold, the travel time savings from reduced congestion are more than offset by the increase in ramp delay. Indeed, the upstream exit mechanism by which ramp metering achieves some of its benefits (refer to Appendix B) suggests that there may be little value to be derived in reducing the mainline queue if it is not blocking any upstream exits, or if the off-ramp volumes in the vicinity of the bottleneck are relatively low.

Although Banks (2000) found that minimization of delay and minimization of freeway congestion do not necessarily coincide, he concludes that the conflict between objectives is not very serious, and is likely to have little practical importance in most situations. However, results from the AMOC algorithm for coordinated ramp metering suggest that the introduction of ramp queue length constraints or other restrictions may in fact lead to a situation where delay is minimized with some extent of mainline congestion remaining (Kotsialos and Papageorgiou 2004). Only in the case of no ramp queue constraints were minimum delay and minimum mainline congestion found to coincide.

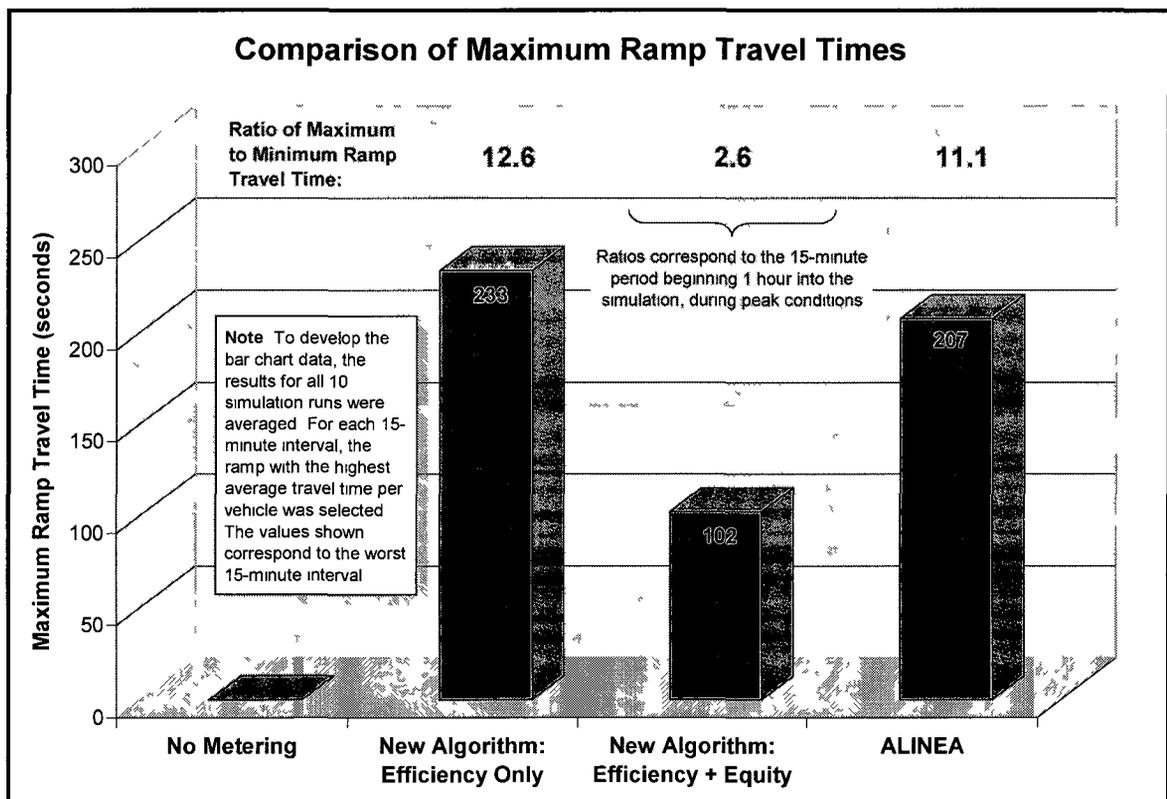
These results highlight the importance of ensuring that the ramp control algorithm (and underlying traffic model) are capable of capturing the effects of congestion, since the most efficient solution may involve some level of mainline queuing, particularly if metering rates are constrained to satisfy ramp queue length criteria. With equity included in the mix, it is even more important to be able to trade-off competing objectives by allowing mainline congestion to occur.

**Table 9-7 Comparison of Average Network Travel Speed Results**

Percentage Difference	No Metering 72.8 km/hr	New Algorithm: Efficiency Only 82.3 km/hr	New Algorithm: Efficiency+Equity 78.4 km/hr	ALINEA 81.1 km/hr
<b>No Metering</b>				
<b>New Algorithm: Efficiency Only</b>	+13% (pr<0.001, 1-tail)			
<b>New Algorithm: Efficiency+Equity</b>	+8% (pr<0.001, 1-tail)	-5% (pr<0.03, 1-tail)		
<b>ALINEA</b>	+11% (pr<0.001, 1-tail)	-1% Not stat. sig.	+3% (pr<0.07, 1-tail)	

**Note:**

Upper value corresponds to the percentage difference between the scenario on the right and the scenario on the top  
 Lower value indicates the statistical significance of the difference In most cases, a one-tail t-test was used since the direction of the difference could be hypothesized in advance (the only exception was the ALINEA algorithm versus the new algorithm optimized for efficiency only, where it was not known in advance which algorithm would be better)



**Figure 9-12 Comparison of Ramp Travel Time Equity**

# 10 ALGORITHM PERFORMANCE IN A REAL-WORLD NETWORK

## 10.1 General Approach

To test the ramp control algorithm under more realistic conditions, a simulation model was developed for the freeway system in Ottawa (Ontario), Canada's national capital. While field trials are ultimately needed to fully evaluate any new control technology, simulation tests provide a low-cost mechanism for testing and refining control measures, allowing potential issues to be addressed prior to field implementation and minimizing the risk of undesirable impacts which could cause the technology to be abandoned. For this reason, simulation tests are considered an essential precursor to any future real-world trials, and were therefore viewed as a logical next step in assessing the performance of the new ramp control algorithm.

## 10.2 Development & Calibration of the Ottawa Simulation Model

The Ottawa traffic simulation model was developed using VISSIM, a commercially available software packaged produced by PTV America. A previous model of the Ottawa area had been developed in INTEGRATION, however, since INTEGRATION lacks a programming interface to implement customized algorithms, a decision was made to convert the model to VISSIM. The key advantages of VISSIM include an extensive range of parameters for fine-tuning model performance, the availability of complementary macroscopic modelling software (VISUM) with built-in conversion tools to facilitate network development and demand estimation, a COM interface which supports custom applications, high-quality graphics for visualization of traffic flow, and dynamic assignment capabilities. This latter feature is particularly important for modelling driver response to control actions such as ramp metering.

The following sections provide a brief overview of the model development process.

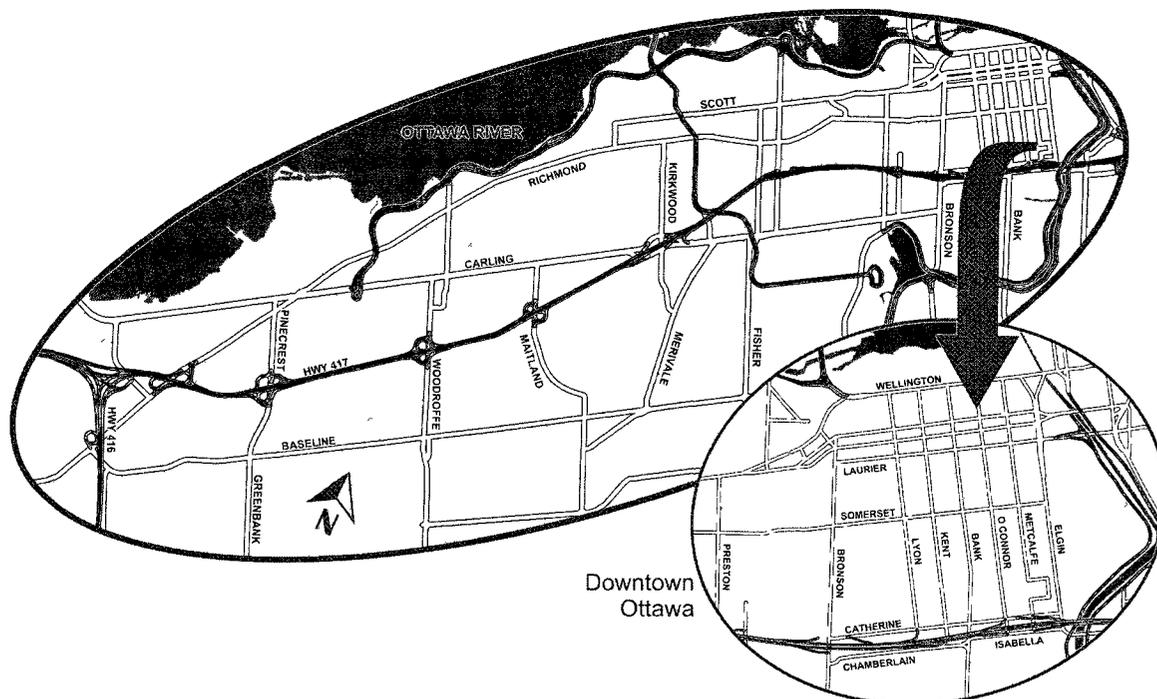
### 10.2.1 Study Area

The study area for the VISSIM model was originally defined to include downtown Ottawa, as well as the adjacent urbanized area to the west. However, to reduce the simulation run-time, a smaller study area was selected for testing the ramp metering algorithm, focusing primarily on the Highway 417 corridor. Highway 417 is Ottawa's only east-west freeway, and plays a key role in meeting the City's mobility needs.

The section of Highway 417 included in the model covers a distance of roughly 15 km, extending from Highway 416 in the City's west end to the Rideau Canal at the eastern limits of the downtown (refer to Figure 10-1). This section currently suffers from traffic congestion, and was therefore felt to be an appropriate candidate for assessing the suitability of the new ramp metering algorithm. Future plans for this location include widening to four mainline lanes per direction between Highway 416 and Carling Avenue, as well as numerous interchange modifications (TSH 2007).<sup>32</sup> While operational measures such as ramp metering may not be sufficient to fully resolve the capacity deficiency, such measures are worth exploring as a means of deferring the very significant capital investment required for road widening. Moreover, since road widening is not considered feasible through the downtown, ramp metering may prove to be a viable alternative for improving freeway performance in this area.

---

<sup>32</sup> Note that the section of Highway 417 to the west of the study area was widened in 2010. In the eastbound direction, a new high-occupancy vehicle (HOV) lane was introduced, as well as an additional lane for general purpose traffic. In the westbound direction, no HOV lane was constructed; instead, the widening involved the addition of two general purpose lanes.



**Figure 10-1 Map of Study Area**

To adequately capture freeway operations and interaction with adjacent roads, all Highway 417 interchanges within the study limits are fully represented in the VISSIM model, including all cross-streets and associated signalized ramp junctions. The model also includes Catherine Street and Chamberlain Avenue / Isabella Street, two major east-west arterials which run parallel to Highway 417 through the downtown. Given their proximity to the highway, these roads may serve as diversion routes for freeway traffic, and may also experience queue spillback from ramp meters should freeway control be introduced.

Within the study limits, Highway 417 generally has between 3 and 4 mainline lanes per direction. The corridor includes both single and multi-lane ramps, and incorporates a variety of interchange configurations. There are no high-occupancy vehicle lanes or dedicated transit lanes on the mainline itself, however, transit-only lanes are provided at some interchanges. As an urban freeway, the speed limit is set at 100 km/hr, similar to practice elsewhere in Ontario.

In developing the VISSIM model, the freeway geometry was coded to reflect conditions in 2004, consistent with the demand assumptions described in Section 10.2.5.

### **10.2.2 Simulation Period**

The VISSIM model corresponds to the afternoon peak period of travel demand since traffic congestion is greatest at this time. Based on travel survey results, the afternoon peak hour occurs between 4:30 p.m. and 5:30 p.m. To capture travel activity over the longer peak period, a two-hour window was modelled, from 4:00 p.m. to 6:00 p.m.

In carrying out the model runs, it was assumed that the first 500 seconds of the simulation were used for loading the network with traffic, loosely based on the time required for a vehicle to travel the entire freeway section at the posted speed limit of 100 km/hr. As a result, the first 500 seconds of the simulation are generally excluded from the network performance statistics.

### **10.2.3 Data Requirements**

The development of a traffic micro-simulation model requires extensive data collection. Some of the key data elements obtained for developing and calibrating the model include:

- GIS road centerline files for creating the skeleton network
- High resolution air photographs showing lane configurations
- Intersection drawings
- Traffic signal phasing and timing plans
- Speed limit and parking regulations from local by-laws
- Traffic count data
- Origin-destination (OD) survey results

The traffic count database that was received for this study contains a record of all the traffic counts conducted in the City of Ottawa since 1995. In general, every signalized intersection in the city is counted at least once every two years, although certain locations may be counted more or less frequently. Using the traffic count database, a number of queries were developed to create the necessary input files for estimating travel demand. In essence, the queries select the most recent count data that is available for a given intersection, excluding any incomplete counts or counts with road closures or restrictions. Since the most recent data contained in the database is from 2004, this year was selected as the base year for the analysis. This base year corresponds well with the date of the most recent OD survey, which was conducted in 2005.

#### **10.2.4 Model Development – Network Specification**

Since network coding is extremely time-consuming in VISSIM, the initial network coding was carried out using the software VISUM, a macroscopic modelling package which works in conjunction with VISSIM. While VISSIM uses a link-connector format to represent the network, VISUM uses a link and node representation which is much easier to code. GIS files imported into VISUM formed the basis for the network, with network details added manually using high-resolution air photographs as a reference layer.

Once the coding was finalized in VISUM, the network was converted to a VISSIM format using the built-in conversion routine. A process of manual editing was then carried out to address any coding issues arising from the conversion process.

#### **10.2.5 Model Development – Travel Demand Estimation**

The travel inputs for the VISSIM model were developed using three primary sources of information:

1. 2005 Origin-Destination (OD) survey for the National Capital Region<sup>33</sup>
2. City of Ottawa traffic count database
3. Preliminary Design and Environmental Assessment Study for Highway 417, from Highway 416 to Anderson Road (TSH 2007)

Development of the travel inputs was carried out using the macroscopic VISUM model described in the previous section. As a starting point, the City of Ottawa's EMME model for the National Capital Region was used to extract the 2005 OD survey data for the study area of interest using the traversal matrix procedure. This procedure essentially develops a new OD matrix for the smaller study area which replicates the travel behaviour captured in the larger matrix, using access points into the smaller study area as new origin-destination nodes.

---

<sup>33</sup> The National Capital Region includes both Ottawa (located on the south side of the Ottawa River in Ontario), and Gatineau (located on the north side of the Ottawa River in Quebec). Combined, the National Capital Region has a population of roughly 1.2 million people, and is Canada's fifth largest metropolitan area (Source: Statistics Canada Summary Tables, Population of Census Metropolitan Areas, <http://www40.statcan.gc.ca/101/cst01/demo05a-eng.htm>).

The OD matrix resulting from this procedure was then modified to match the more detailed traffic zone system in VISUM, which was developed to support the microscopic traffic modelling in VISSIM. Given the size of the EMME traffic zones, several network access points are needed for each zone to ensure that traffic is loaded onto the network as realistically as possible within the simulation environment, particularly at signalized intersections.

Once the OD survey data was extracted from the EMME model for the study area of interest and modified to correspond to the VISUM traffic zone system, the resulting OD matrix was imported into VISUM and a process of further refinement was carried out. As part of this process, the TFlowFuzzy feature in VISUM was used to adjust the OD data to better match the observed traffic at selected locations.

The traffic count data used in this process was extracted from the City of Ottawa traffic count database using specially developed queries to link the selected data with the corresponding road segment in VISUM for use in the TFlowFuzzy algorithm. As with the OD survey data, all count data used in the analysis corresponds to the afternoon peak hour of a typical weekday.

While Highway 417 is under the jurisdiction of the Ontario Ministry of Transportation, certain elements of the highway are included in the City of Ottawa's traffic count program. The City conducts intersection turning movement counts at signalized ramp junctions, as well as Classification and Occupancy counts on certain mainline sections of the highway.<sup>34</sup> As a result, much of the traffic count data needed for Highway 417 was available from the City's traffic count database. For those locations not included in the database, traffic count information was obtained from the Preliminary Design / Environmental Assessment study for Highway 417.<sup>35</sup>

---

<sup>34</sup> Note that the Classification and Occupancy data was only used for the final model calibration.

<sup>35</sup> The report includes observed data from traffic count stations (from 2000/2001), as well as derived data based on upstream and downstream conditions. Only data from traffic count stations was used for model development & calibration.

In working with TFlowFuzzy, no effort was made to adjust the count data to a common base year or season.<sup>36</sup> Since the algorithm modifies the demand inputs to better match the observed counts across the entire network, any anomalies in the count data are automatically dealt with, resulting in a set of ‘balanced’ traffic flows. In interpreting the calibration results, this fact should be kept in mind; the modelled flows will never match the observed flows exactly due to random variation in the observed data and differences in when the data was collected.

The use of TFlowFuzzy to modify the OD matrix has two important benefits:

1. Improved distribution of demand to the various network access points – The process of converting from the EMME traffic zone system to the more detailed traffic zone system used for micro-simulation is subject to considerable error. By using TFlowFuzzy, the assumed volume entering the network at each access point will more closely match observed conditions.
2. Ability to account for truck traffic – Since the OD survey only captures personal travel, heavy vehicle activity is ignored. By using TFlowFuzzy, the network flows will be adjusted to more closely match the observed traffic, which includes all motorized modes.

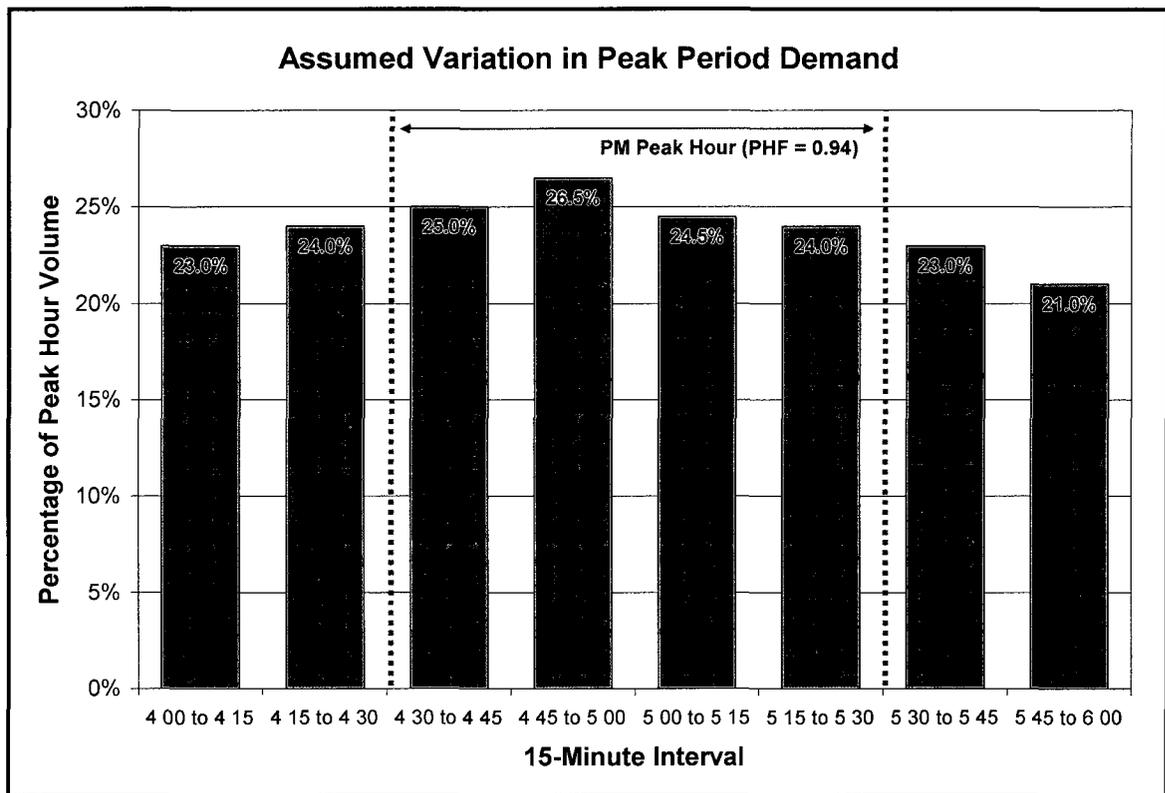
Once the results of the TFlowFuzzy procedure were deemed to be acceptable, a new OD matrix was generated for the somewhat smaller study area modelled in VISSIM. This task was accomplished using VISUM’s subnetwork generator, which, similar to the traversal matrix procedure in EMME, considers the paths used by vehicles in the original study area, and generates new zones at the interface where traffic enters or leaves the subnetwork such that the resulting OD matrix for the ‘clipped’ network replicates the flows in the larger network. The resulting OD matrix was then exported into a format that could be read by VISSIM for micro-simulation of traffic flows.

To account for variation in demand over the 2-hour simulation period, the peak hour matrix was factored to create a series of matrices, each 15 minutes in duration. The assumed distribution is shown in Figure 10-2. This distribution has a peak hour factor of

---

<sup>36</sup> Most of the count data used in the analysis was collected in 2003/2004 during the months of May, June, July, and August. Evidence from the traffic count data suggests that overall traffic growth in the years prior to 2004 was not significant (although specific corridors may have seen larger variations).

0.94, which is roughly comparable to the average peak hour factor for those links having an observed peak hour flow of a least 500 vehicles per hour.



**Figure 10-2 Assumed Variation in Demand During the Peak Period**

In the process of calibrating the model, a small number of manual adjustments were made to the OD data to better replicate the observed counts. The results of the calibration exercise are described in the following section.

In terms of heavy vehicle traffic on the road network, it was assumed that roughly 2.5% of the total travel demand was comprised of heavy vehicles, loosely based on trends observed in the traffic count data for the afternoon peak hour, as well as observations from the 2007 Interprovincial Roadside Truck Survey (TRANS 2011). For Highway 417 through volumes, a 5% truck percentage was assumed, based on results presented in the Preliminary Design & Environmental Assessment study for the corridor (TSH 2007).

### 10.2.6 Model Calibration / Validation

The appropriateness of the various model parameters was confirmed as part of the calibration exercise. A summary of the assumptions adopted in the Ottawa model can be found in Appendix P.

For the most part, the same parameter values used in the VISSIM test network described in Section 8.2 were also used in the Ottawa model, with one or two significant exceptions:

- The CC1 parameter in the freeway car following model was increased from 0.9 seconds to 1.1 seconds. This parameter represents the desired headway time between successive vehicles to maintain a minimum safety distance. As a result, lowering this value had the effect of reducing the freeway capacity.
- The capacity was further reduced by modelling temporary lack of attention, during which time drivers do not respond to a preceding vehicle (except for emergency braking). The duration of each incident was set at 1 second, while the probability was set at 1%.

With no freeway sensor data available for the corridor, it is difficult to estimate an appropriate capacity value. Nonetheless, a reduction in the VISSIM default capacity was found to be the only way to replicate both the observed traffic volumes and congestion. While increasing the through traffic on Highway 417 had the effect of producing the desired congestion, the downstream flows exceeded the observed counts. Thus, rather than artificially increasing the travel demand, parameter changes were introduced to reduce the freeway capacity, which again triggered the desired congestion, but also produced downstream traffic volumes which more closely matched the observed counts.

As input to the calibration phase, path and cost files were generated for use in the traffic assignment. The dynamic assignment process in VISSIM is an iterative process, as drivers “learn” the best routes between origin-destination nodes. The route choice decision is based on a discrete choice model (specifically, a variant of the Logit model) which determines the proportion of drivers assigned to a given route based on the generalized cost of the route compared to the other available choices. For the Ottawa model, convergence of the assignment was assumed to occur when the difference in travel time on all origin-destination paths was less than 20% between any two successive

iterations. Path and cost files were initially generated at the start of the calibration phase, and were subsequently updated as calibration proceeded. Once the calibration results were deemed to be acceptable, the same path and cost files were used for all subsequent model runs as recommended in the VISSIM manual.<sup>37</sup> Convergence results for the final set of path and cost files can be found in Appendix P.

During the calibration process, the model parameters were adjusted to better replicate observed conditions. In isolated instances, origin-destination flows were modified to provide a better fit with the observed counts. To further improve the model results, additional enhancements were undertaken:

- In certain cases, drivers were observed to exit and then re-enter the freeway at the same interchange (specifically, at the Parkdale Avenue interchange and the Carling Avenue interchange). While such behaviour may occur in reality to a limited extent, a decision was made to close such paths within the simulation model to prevent significant numbers of vehicles from making these manoeuvres.
- It was noted that the simulation model has difficulty modelling lane changes on the arterial network where there is significant weaving activity. In reality, drivers act cooperatively so that lane blockages are quickly resolved. Within the model, vehicles waiting for a lane change are unable to move forward, blocking the intersection so that no one is able to proceed and causing significant congestion. To address such behaviour, a special “weaving link” behaviour type was defined. For these links, the maximum deceleration for cooperative braking was set at  $-6 \text{ m/s}^2$  to encourage more cooperative braking. In addition, in terms of lateral behaviour, drivers were assumed not to observe vehicles in adjacent lanes, allowing them to proceed with a lane change regardless of other vehicles’ lateral positioning.

Since no real-world sensor data was available for calibrating the parameters of the Ottawa model, the reliability of the model was assessed by examining the simulated traffic volumes and congestion patterns.

- First, the simulated traffic volumes were compared to the observed traffic counts to ensure that the model was replicating the observed data with a reasonable level of accuracy.
- Second, the patterns of traffic congestion were noted, and compared to observed freeway conditions as recorded in traffic camera snapshots. In addition, a

---

<sup>37</sup> However, new cost and path files were generated for the ramp metering scenarios to capture the impact of ramp delays on vehicle routing through the network.

comparison was also made with the Highway 417 simulation results reported in the Preliminary Design and Environmental Assessment Study for the corridor.

Volume III of FHWA's Traffic Analysis Toolbox provides an example of calibration targets that were adopted by the Wisconsin DOT for simulating the Milwaukee freeway system (Dowling et al. 2004). The targets are based on guidelines developed in the United Kingdom, and have also been adopted by Caltrans (Dowling et al. 2002). Given their widespread use, similar targets were used in assessing the performance of the Ottawa model. Table 10-1 presents the results of the volume calibration; additional details can be found in Appendix P.

The calibration was assessed for peak hour conditions, based on a total of 10 simulation runs. As shown in Table 10-1, the calibration results are generally considered acceptable. While the model does not meet the calibration criteria for high volume links with flows greater than 2700 vph, there are only 12 links in the category. These links generally correspond to Highway 417, which is subject to both recurrent and non-recurrent congestion. As a result, significant day-to-day variation in the count data is anticipated, and any mis-match between the simulated and observed volumes was attributed to this variability.

Since turning movement counts are more difficult to replicate within a simulation environment, the level of error is generally expected to be higher. While the calibration results were assessed for link volumes only (consistent with Dowling et al. 2004), the model was also found to replicate turning movements with a reasonable level of accuracy, as shown in Appendix P.

**Table 10-1 Peak Hour Volume Calibration Results**

	<b>Measure</b>	<b>Criteria</b>	<b>Acceptance Target</b>	<b>Simulation Results</b>	<b>Target Met?</b>
<b>Individual Links</b>	<i>Difference between Modelled and Observed link flows</i>				
	• Flows less than 700 vph (68 links)	Difference within 100 vph	> 85% of cases meet criterion	Range: 88% - 94% Average: 91%	✓
	• Flows between 700 vph and 2700 vph (81 links)	Difference within 15%	> 85% of cases meet criterion	Range: 85% - 90% Average: 89%	✓
	• Flows greater than 2700 vph (12 links)	Difference within 400 vph	> 85% of cases meet criterion	For all runs, 67% of links passed	✗
	GEH Statistic for individual link flows	GEH < 5	> 85% of cases meet criterion	Range: 83% - 88% Average: 86%	✓
<b>All Links</b>	Sum of all link flows	Within 5% of sum of all link counts	Criterion met	Difference < 1% for all runs	✓
	GEH Statistic for sum of all link flows	GEH < 4	Criterion met	Range: 0.3 - 3.3 Average: 1.1	✓

**Notes:**

1. Criteria & acceptance targets based on Dowling et al. (2004)
2. GEH statistic computed as follows:

$$GEH = \sqrt{\frac{(E - V)^2}{(E + V)/2}}$$

where E = model estimated volume and V = field count

As part of the calibration exercise, an effort was also made to ensure that the congestion patterns simulated in the VISSIM model were similar to those observed in real-life. Table 10-2 presents a summary of Highway 417 congestion based on observations from traffic cameras (recorded in September, 2006), and also findings from the operational assessment and INTEGRATION modelling that was carried out as part of the Highway 417 Preliminary Design and Environmental Assessment Study (TSH 2007). Since the VISSIM simulation model has a 2004 base year, it is roughly comparable to conditions somewhere in between the 2001 and 2011 results presented in the TSH report.

From a review of Table 10-2, it is clear that the various sources of information do not always agree on the nature and extent of Highway 417 mainline queuing, likely due to day-to-day variability in traffic conditions and limitations of the assessment techniques (including slightly different base years). Nonetheless, certain trends emerge:

- Westbound on Highway 417, congestion regularly occurs between Carling / Maitland Avenue and somewhere in the downtown. There is some evidence that

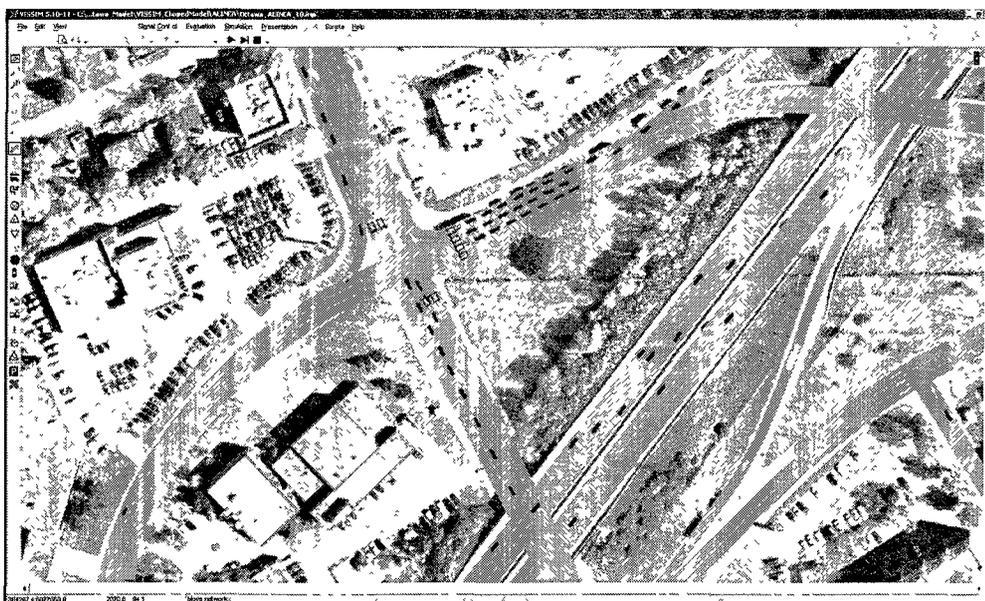
congestion may extend further in both directions (i.e. to Woodroffe Avenue in the west, and beyond the VISSIM model limits in the east)

- Eastbound on Highway 417, bottlenecks outside the eastern limits of the VISSIM model create congestion which extends some distance back through the downtown

In general, the VISSIM results for the westbound direction agree with the trends presented above. Congestion is initiated at the Woodroffe Avenue on-ramp and Carling Avenue on-ramp, with the Woodroffe Avenue queue sometimes extending as far back as Carling Avenue, creating one continuous queue. Overall, the westbound queue spills back to between Bronson Avenue and Metcalfe Street (both located in downtown Ottawa), depending on the model run.

In the eastbound direction, the VISSIM model shows traffic moving freely with no major bottlenecks or slowdowns. Although such results are not consistent with the trends presented above, they are considered to be reasonable given that the VISSIM model does not include the major bottlenecks to the east of the downtown which are the source of the eastbound congestion.

A congestion map showing the VISSIM simulation results for the congested westbound direction of Highway 417 is provided in Appendix P for a typical model run.



*Snapshot of the Ottawa Model at the Carling Avenue Interchange*

**Table 10-2 Highway 417 Congestion Patterns during the Afternoon Peak Period****Traffic Camera Snapshots from September 25<sup>th</sup>, 26<sup>th</sup>, and 27<sup>th</sup>, 2006***Westbound*

- Congestion through the downtown west to at least Carling Avenue and possibly Maitland Avenue
- Further west, volumes continue to be heavy, but traffic seems to be moving better (may be observing shockwave behaviour)



*South side of Hwy 417 near Booth Street looking eastbound*

*Eastbound*

- Congestion heading out of the downtown (more extensive on September 27<sup>th</sup> with queuing to approximately Rochester Avenue)
- Elsewhere within the study area, traffic appears to move well

**Highway 417 Preliminary Design and Environmental Assessment Study (TSH 2007)****Traffic Operations Study - Estimated Performance for Existing (2001) & Future (2011) Conditions***Westbound*

- Analysis of existing conditions shows localized congestion at the Maitland Avenue interchange, with the westbound ramp merge operating at LOS F. The basic freeway segment between Carling Avenue and Maitland Avenue operates at LOS E, indicating unstable, near capacity conditions. By 2011, congested operations are projected for all but one of the mainline sections between the Carling Avenue ramp diverge and Maitland Avenue ramp merge.
- Existing near-capacity operation of the weaving section east of Metcalfe Street is projected to continue through 2011 and beyond.
- Congestion anticipated at the Woodroffe Avenue and Pinecrest Road interchanges by 2011.

*Eastbound*

- Existing congestion east of Metcalfe Street is projected to remain through 2011 and beyond.
- Basic freeway segment between Maitland Avenue and Carling Avenue projected to operate near capacity by 2011.

**Highway 417 Preliminary Design and Environmental Assessment Study (TSH 2007)****INTEGRATION Simulation Model Results***Westbound*

- Model results
  - 2001: Congestion from Bronson Avenue to Woodroffe Avenue
  - 2011: Congestion from the eastern limits of the VISSIM model to downstream of the Woodroffe Avenue on-ramp
- General observations
  - Westbound Highway 417 from Bronson Avenue to Woodroffe Avenue experiences recurring congestion in 2001 due to high traffic volumes and significant weaving. By 2011, this congestion is expected to extend eastward to the VISSIM model limits.

*Eastbound*

- Model results
  - 2001: Congestion from upstream of the Rochester Street off-ramp to the VISSIM model limits
  - 2011: Congestion throughout the entire VISSIM model limits
- General observations
  - Major weaving sections on Highway 417 outside the VISSIM model limits (i.e. east of the downtown) generate significant congestion in the eastbound direction. Results of the INTEGRATION modelling suggest that this congestion may back up all the way to Highway 416 in the future.

### 10.3 Implementation of the Ramp Metering Algorithm

The ramp metering algorithm was tested on the westbound section of Highway 417 between downtown Ottawa and Pinecrest Road. This section experiences significant traffic congestion during the afternoon peak hour, and is therefore a suitable candidate for ramp metering control.<sup>38</sup> In total, the ramp metering limits cover a distance of roughly 12 km, and include 10 on-ramps and 9 off-ramps.

In implementing ramp metering control within the simulation environment, an effort was made to be as realistic as possible in terms of the detector placement, stop bar position, and other ramp metering elements. It was generally assumed that no changes would be made to the freeway geometry, with the following two exceptions:

- To increase the amount of ramp storage, all ramps within the metering limits were assumed to have two lanes for storing queued vehicles: the existing travel lane, and the ramp shoulder. However, it is recognized that use of the ramp shoulder for storage purposes may, in some cases, require pavement widening.
- To further increase the amount of ramp storage at the Carling and Bronson on-ramps, the ramp-arterial junction was reconfigured. While the modified geometry is considered reasonable given the adjacent land use, no attempt was made to confirm the feasibility of the changes. An opportunity to allow queue storage on the side street connecting to the Rochester on-ramp was also identified, however, this option was not pursued given the significant changes in traffic circulation that this would entail.

Although the above changes were found to improve the effectiveness of ramp metering within the Highway 417 corridor, storage constraints continued to be a limiting factor in the overall congestion benefit that could be achieved through metering control. Many of the ramps in downtown Ottawa are extremely short, and have little storage capacity even when expanded to two lanes. Given the physical restrictions in the downtown core, further expansion of the storage capacity was deemed to be infeasible.

To implement the new ramp control algorithm within the VISSIM model, the Highway 417 corridor was divided into segments ranging in length from roughly 350 m to 650 m. These segment lengths are sufficient to ensure that no vehicle can cross more than one

---

<sup>38</sup> Since the model does not show congestion in the eastbound direction (refer to Section 10.2.6), there is little value to be gained by testing the performance of the algorithm in this direction.

segment boundary during a single 10 second time step. Since Highway 417 has different operating characteristics than those assumed in the hypothetical test network, appropriate assumptions were adopted for a number of algorithm parameters such as the freeway capacity under congested flow. However, it was assumed that the same flow breakdown relationships developed previously could be applied to the Ottawa model, despite differences in the freeway cross-section and driver behaviour.<sup>39</sup> Since the flow breakdown model was only developed for on-ramps with an auxiliary speed change lane, it was further assumed that any ramps with a lane addition would have a negligible probability of breakdown – an assumption which is generally consistent with observations from the model calibration.

Additional information on the implementation of ramp metering in the Ottawa model can be found in Appendix P. Appendix P also provides a summary of the various assumptions that were adopted in applying the ramp control algorithm.

#### **10.4 Summary of Key Results**

The introduction of ramp metering on Highway 417 was found to reduce both the duration and extent of mainline congestion, regardless of the algorithm employed. However, even with ramp metering control, significant congestion remains, primarily due to ramp storage constraints which limit the amount of metering that can be carried out. In applying ramp control, it was found that in some model runs, the congestion at the Woodroffe on-ramp could be totally eliminated, in other runs, the congestion re-appeared. These results suggest that traffic operations along the corridor are subject to considerable variability due to the effects of over-saturation.

Results from the simulation tests are presented in Table 10-3 for the average network travel speed. In general, the results are similar to those obtained previously with the VISSIM test network. When optimized for efficiency only, the new algorithm is able to achieve an improvement in network performance which is only slightly worse than that observed with the ALINEA algorithm. Since the performance difference between the two

---

<sup>39</sup> As noted in Section 10.2.6, different driver behaviour parameters were used in the VISSIM model for Highway 417 to better replicate the observed traffic volumes and congestion patterns.

algorithms is not statistically significant (based on a two-sided t-test with  $\alpha = 0.05$ ), it is reasonable to conclude that the performance of the new algorithm is roughly comparable to that of ALINEA when optimized for efficiency alone.

When equity is included in the utility function for the new algorithm, the improvement in network travel speed declines. This finding is consistent with the previous results, which found a trade-off between efficiency and equity. In this case however, the drop in efficiency performance is less dramatic. From a review of the ramp travel time results, it would appear that this may be due to the fact that the algorithm is not able to achieve as great an improvement in equity. Figure 10-3 illustrates the average ramp travel time along the corridor with new algorithm in operation, with the upper part of the figure corresponding to the scenario with equity excluded, and the bottom portion corresponding to the scenario with equity included.

A review of Figure 10-3 shows a substantial decline in the ramp travel at the Woodroffe on-ramp with the inclusion of equity in the utility function. However, a much smaller benefit is achieved at the Maitland and Carling on-ramps, which also experience relatively high delays. It is hypothesized that this finding may be the result of the limited ramp storage capacity on ramps closer to the downtown, and the fact that all of the ramps were already being metered reasonably hard under the efficiency-only scenario (as reflected by the ramp queue lengths). Ramps not being metered particularly hard in the maximum efficiency scenario are either downstream of the bottleneck, or have a storage length that precludes substantial metering. Thus, there is limited opportunity to spread the delay more equitably, and the corresponding impact on network efficiency is less severe.

Other key observations include the tendency of the algorithm to reduce the ramp delay on all ramps along the corridor, even ones with relatively low delay,<sup>40</sup> and the difference in equity performance during the initial and latter portions of the simulation. With regards to this final point, it would appear that when demand (and congestion) are lower, the algorithm is able to achieve a more significant equity improvement, with less variation in the ramp delays.

---

<sup>40</sup> It is anticipated that this behaviour may be a reflection of the utility functions used in the algorithm. To assess the effect of different functional forms and weights, further investigation is required.

It is also worth noting that, under the efficiency only scenario, the algorithm did not appear to meter certain on-ramps as hard as would reasonably be expected. It is hypothesized that this may be a reflection of the limited storage capacity of these ramps, the metering rate over-ride procedure used to deal with spillback (which may have been triggered too early), and demand prediction errors (which cause the algorithm to select a higher metering rate than necessary to avoid spillback). Improvements to the queue prediction methodology would allow the algorithm to maintain a more stable queue length, ensuring that the available storage space is used as efficiently as possible.

In terms of spillback occurrence, the performance of the new algorithm was generally found to be acceptable once appropriate measures were put in place to deal with high-volume ramps such as Carling and Maitland (refer to Appendix P). While small spillback episodes were occasionally observed, for the most part, such episodes lasted for only a few minutes. On one or two of the model runs, larger spillback episodes were recorded in the simulation outputs, however, it is believed that the extent of impact may have been over-stated due to limitations of the methodology used to calculate queues within the VISSIM software.<sup>41</sup>

**Table 10-3 Comparison of Algorithm Results for Average Network Travel Speed**

<b>Algorithm</b>	<b>Average Network Travel Speed<sup>1</sup></b>	<b>Improvement Compared to the Base Case</b>
Base Case (no ramp metering)	59.8 km/hr	--
New Algorithm (Efficiency Only)	64.9 km/hr	8.5%
New Algorithm (Efficiency + Equity)	63.4 km/hr	6.0%
ALINEA	66.6 km/hr	11.5%

<sup>1</sup> Based on results for the entire simulation period, excluding the first 500 seconds while the network is being loaded

<sup>41</sup> Once a ramp queue spreads upstream of an adjacent intersection, the most critical intersection approach queue is included in the ramp queue calculation, making it difficult to distinguish which part of the queue may be due to spillback, and which part is reflective of normal queuing behaviour at the intersection.

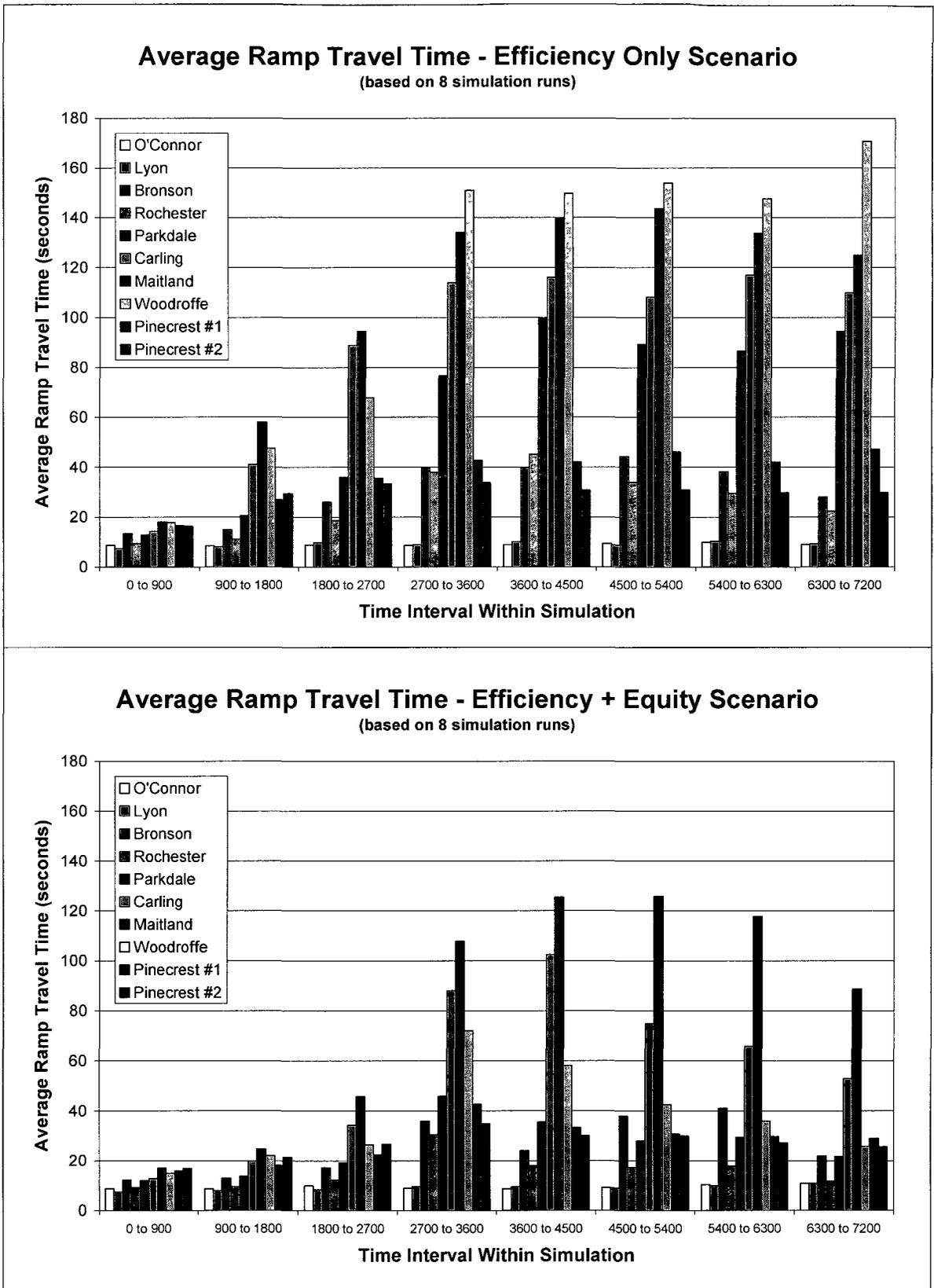


Figure 10-3 Comparison of Algorithm Results for Average Ramp Travel Time

## 11 CONCLUSIONS

This thesis has described the development of a new ramp control algorithm which better meets the needs of system operators and users by considering both equity and efficiency in the control problem. To capture the uncertainty inherent in real-world traffic systems, the algorithm was implemented within the framework of a dynamic Bayesian decision network. Key components of the algorithm include the freeway traffic model used for traffic state estimation and predictive control; the particle filter used in belief updating; and the module for determining the optimal control solution on the basis of the expected utility. The algorithm was coded in MATLAB and integrated into the micro-simulation software VISSIM for testing within a simulation environment.

To confirm the validity of the proposed approach, off-line testing of the freeway traffic model which underlies the algorithm was carried out. From the results presented, it can be concluded that the model is generally well-suited to modelling freeway operations under both congested and uncongested conditions where evidence is available at regular intervals. Even when the probability of flow breakdown is low, the model can accurately predict freeway performance by incorporating evidence from traffic sensors.

In subsequent on-line experiments using a typical freeway test network, it was determined that the proposed ramp control algorithm generally functions as intended, addressing both equity and efficiency objectives in accordance with the underlying utility function. In terms of results, the version of the algorithm maximized for efficiency was able to achieve a 13% gain in the average network travel speed compared to the no metering scenario – a level of improvement comparable to that attained with the popular ALINEA algorithm. With the utility function modified to include equity objectives, the algorithm was only able to achieve an 8% improvement in network travel speed, however, the ramp delays were substantially more equitable. With efficiency as the sole objective, the average ramp travel time was found to exceed 200 seconds per vehicle in certain cases; the inclusion of equity caused the maximum threshold to drop to just over 100 seconds per vehicle – a sizeable improvement. These findings provide additional support for the trade-off between efficiency and equity documented in the literature.

To further establish the validity of the proposed approach, the algorithm was also tested using a simulation model of the Highway 417 freeway in Ottawa, Ontario. The results generally confirm the previous findings from the VISSIM test network. However, while a trade-off was observed between equity and efficiency, the magnitude of the trade-off was lower. It is hypothesized that this may be due to the greater level of traffic congestion on Highway 417, and the limited amount of ramp storage available, which makes it difficult for ramp metering to reduce congestion, and provides less opportunity for distributing the ramp delays more equitably. Of course, different utility weightings could produce different results.

Overall, the results of the algorithm tests appear promising. The new ramp control algorithm incorporates a number of innovations which underscore its potential merit:

- **Its Bayesian structure allows sources of uncertainty to be modelled explicitly, not only in tracking mode but also when carrying out predictive control.** The resulting control solutions can thus be tailored to reflect the risk-tolerance level of the system operator – something which most algorithms can only address in an ad hoc fashion if at all.
- **Its novel approach to modelling flow breakdown as a stochastic process better reflects real-world behaviour.** The breakdown model developed in this research demonstrates both the feasibility of the approach, and its potential application to freeway modelling and control.
- **Its modular framework supports any number of future enhancements and modifications;** any of the underlying components (the freeway traffic model, inference process, solution algorithm, etc.) can be easily swapped out for something different or new components added in. Indeed, the algorithm framework was specifically designed to support the introduction of new features, such as incident detection or integration with the arterial network.
- **Its unique utility formulation allows equity and efficiency objectives to be traded off directly** in determining the optimal ramp metering rates, addressing a major gap in current ramp metering approaches. Transportation agencies have the flexibility of defining utility in a way that best meets their requirements, in accordance with local attitudes and values. This ability to tailor the algorithm is of significant benefit, providing a mechanism to pro-actively manage the system in a way that reduces opposition from the public.

From the results of this research, it can be concluded that the proposed approach is feasible, and that dynamic Bayesian decision networks can provide an appropriate

framework for structuring traffic control problems. However, the results also demonstrate that further work is needed to address specific issues prior to field implementation. Opportunities for improvement are described in the Recommendations section which follows.

## 12 RECOMMENDATIONS

### 12.1 Overview

This thesis represents the first step in developing a new ramp metering algorithm within the framework of a dynamic Bayesian decision network – an algorithm which addresses both efficiency and equity in accordance with the research objectives. Since the primary goal was to assess the feasibility of the proposed approach, the main emphasis was on developing the core algorithm functionality – the freeway traffic model, the inference module, the demand prediction module, and the solution module.

In developing these components, a methodology was sought which would achieve the intended functionality in a straightforward a way as possible, recognizing that future enhancements may be warranted in subsequent phases of work. Likewise, a decision was made to defer all ‘non-essential’ features until after the current research was complete, once the feasibility of the approach could be confirmed and the merit of proceeding with further trials was clear.

In light of the above, a secondary, but no less important goal was to identify areas of the algorithm which could be improved upon in future work, and develop recommendations to enhance performance and expand the scope of application. The following discussion provides an overview of the key recommendations arising from this research effort.

### 12.2 Potential Future Enhancements

Recommendations for future enhancements fall into two broad categories: recommendations aimed at improving the performance of the existing algorithm, and recommendations aimed at expanding the algorithm to include new functionality.

#### *Improvements to the Existing Algorithm – Modelling of Freeway Operations<sup>42</sup>*

- **Estimating the initial freeway state** – Within the Freeway Traffic Model, the initial freeway state is represented by a probability distribution derived from traffic data. Moving forward, it is recommended that options be explored for

---

<sup>42</sup> Many of these recommendations have been summarized from Section 7.4.

improving the accuracy of the initial distribution and/or refining it through an initialization process.

- **Predicting future demand** – The current demand prediction module uses flow observations from the previous two minutes to estimate future demand over the prediction horizon. Since the accuracy of the Freeway Traffic Model is closely tied to the accuracy of the demand predictions, it is recommended that more sophisticated demand estimation techniques be explored to improve prediction accuracy.
- **Modelling traffic diversion** – As part of improving the demand predictions, improved methodologies are needed for modelling traffic diversion at on- and off-ramps. At on-ramps, diversion to the arterial network may occur due to ramp metering delays, reducing the on-ramp demand. At off-ramps, drivers may divert from the freeway in response to variable message signs or other forms of traveller information which warn of downstream congestion. If such diversion is ignored, the projected demand entering/exiting the freeway may be over- or under-stated.
- **Regulating the ramp queue length** – The accuracy of the demand predictions can have a major impact on the ability of the algorithm to regulate ramp queues, particularly when storage constraints are controlling the metering rate. To minimize spillback onto the arterial network, a rudimentary queue over-ride feature was included in the ramp controller. However, more sophisticated over-ride strategies have been developed which may prove more efficient. There is also opportunity to provide additional evidence on the ramp queue to the Freeway Traffic Model to improve tracking behaviour, for example using the vehicle speed at the queue detector to estimate the current queue length as in the approach developed by Sun and Horowitz (2006).
- **Modelling ramp capacity** – In the current version of the model, it is assumed that the ramp capacity is adequate to accommodate the observed demand. In future versions of the algorithm, it is recommended that this constraint be removed. In the case of off-ramps, this may require developing a mechanism to deal with ramp queues which spread back onto the freeway mainline due to insufficient intersection capacity at the ramp-arterial junction.
- **Estimating the average travel speed** – The speed relationship used in the current version of the model does not account for the impact of merging and diverging behaviour. However, turbulence at ramp junctions can influence vehicle speeds in the vicinity of the ramp. By adjusting the speed model to account for such impacts, model performance may be enhanced.
- **Testing for freeway recovery** – To predict when freeway congestion has dissipated, the Freeway Traffic Model reassesses the probability of flow breakdown whenever the cell density approaches the critical density observed when congestion first appeared. Given the limitations of this approach,

opportunities for improving on the recovery test should be explored as the model is further refined.

- **Modelling freeway queues** – The current version of the Freeway Traffic Model employs a relatively simplistic approach to modelling traffic queues which essentially ignores shockwave behaviour unless evidence is available from traffic sensors to refine the queue estimates. While ramp metering is intended to reduce freeway congestion, it may not be possible to avoid congestion altogether, and it is therefore recommended that an improved methodology be developed for modelling freeway shockwaves and the resultant traffic queues.
- **Updating the congested capacity (queue discharge flow)** – The approach adopted for updating the congested capacity was shown to improve the accuracy of the Freeway Traffic Model considerably. Given the sensitivity of the model to this value, additional tests are needed to ensure that the assumptions adopted work well under real-world conditions where the capacity may vary even over the course of the peak period due to environmental factors or other variables. Moreover, it is recommended that other options for estimating the real-time capacity be explored as well to address the methodological issues outlined in Section 7.4 and see if further improvements are possible.
- **Updating other key parameters** – While the Freeway Traffic Model currently includes a mechanism to update the probability distribution for the congested capacity over time (to account for the effect of weather, etc.), the probability distribution for all other parameters is assumed to be fixed. Moving forward, the appropriateness of this assumption should be assessed, and where warranted, changes introduced to allow the default distribution to be updated based on the observed evidence.
- **Modelling flow breakdown** – One of the key elements of the Freeway Traffic Model is the relationship used to predict flow breakdown as a function of prevailing geometric and traffic conditions. To improve the breakdown relationship, it is recommended that the following activities be carried out:<sup>43</sup>
  - Develop probability of breakdown functions using real-world data to confirm the appropriateness of the approach and identify any issues which may need to be addressed
  - Determine the optimal position of the cell boundaries in the Freeway Traffic Model from a flow breakdown perspective
  - Expand the breakdown model to cover a range of freeway cross-sections
  - Examine the impact of weather and other variables on flow breakdown (such as lane width, roadside environment, heavy vehicle proportion, etc.)
  - Explore other types of models that could be used to predict flow breakdown

---

<sup>43</sup> Additional discussion on these recommendations can be found in Section 7.4.1.

Since the probability of breakdown function used in the Freeway Traffic Model only predicts flow breakdown at freeway merges, the model is expected to perform poorly if flow breaks down at other locations, such as off-ramps or lane drops. As a result, it is also recommended that new relationships be developed to model the breakdown phenomenon at such locations.

Likewise, adjustments are needed to ensure that the Freeway Traffic Model is sensitive to flow breakdown due to freeway incidents. Although incidents may be relatively rare, they are a normal part of freeway operations, and it is important that the Freeway Traffic Model be able to track freeway performance when incidents do occur. Such capability does not necessarily require that the model detect or predict incidents, only that it be able to account for the impact of incidents when they arise.<sup>44</sup>

- **Testing different underlying traffic models** – One of the key strengths of the proposed framework is its flexibility; one module can be easily swapped for another as long as it fits within the Bayesian formulation of the problem. While the Freeway Traffic Model developed in this research was found to perform well, the use of other underlying traffic models should also be investigated to assess which models perform best from a ramp metering perspective.
- **Modelling of mixed traffic flows** – The current version of the algorithm considers the traffic stream to be relatively homogeneous. However, it is conceivable that the model could be structured to consider cars and trucks as separate vehicle classes, similar to the approach adopted by Ngoduy (2008). Of course, such a modification would only be warranted on freeways with a high proportion of trucks, where a significant difference exists between the operating characteristics of the various modes.

### *Improvements to the Existing Algorithm – Real-Time Inference*

- The approach to real-time inference adopted in this research involved the use of a relatively simple particle filter. However, the literature on particle filters is vast, and much more sophisticated inference approaches have been developed, including rao-blackwellized particle filters (where some variables are analytically marginalized out using standard algorithms such as the Kalman filter) and unscented particle filters (which use an unscented Kalman filter as the proposal distribution rather than the transition prior) to name just two. It is therefore recommended that alternative filtering techniques be explored in order to improve

---

<sup>44</sup> For example, the model could be ‘informed’ of an incident by having a system operator manually enter the incident details, or relying on the results of an incident detection algorithm. Where such information is not available, the algorithm should still be able to detect non-recurrent congestion based on the readings received from traffic sensors. However, for maximum effectiveness, the Freeway Traffic Model must allow for the possibility of flow breakdown at any cell boundary, not only at freeway merges which is currently the case. Moreover, additional freeway sensors are desirable to detect incidents more quickly and pinpoint the corresponding bottleneck more precisely.

the performance of the Freeway Traffic Model, particularly in tracking mode. With better filtering techniques, it may be possible to improve the accuracy of the model estimates and/or reduce the number of particles, improving the estimation speed.

### *Improvements to the Existing Algorithm – Finding the Optimal Solution*

- The main obstacle to implementing the new ramp control algorithm in the field is the solution speed. Currently, the algorithm is not able to find a control solution within the real-time constraints of the problem. However, the extent of the deficiency is not considered to be insurmountable. Options for addressing this issue include optimizing the MATLAB code (or converting it to a more efficient language), providing additional computing power, or implementing a more efficient solution algorithm. This latter option in particular is worth exploring. It is entirely feasible that better solution algorithms may exist which could improve the accuracy and speed of the solution module, enhancing the overall performance of the ramp control algorithm.

### *Addition of New Features*

- **Incident detection** – Since traffic incidents are random events, they can be represented using Bayesian networks. As a result, it should be relatively straightforward to expand the Bayesian network for the ramp control algorithm to include an incident detection module which detects freeway incidents given speed and flow evidence from traffic sensors.
- **Applicability to high-volume ramps** – In the current version of the algorithm, it is assumed that only one vehicle is allowed to proceed per green indication. This type of metering has a maximum capacity of roughly 900 vph. With a few minor changes to the algorithm, other metering schemes could be accommodated (such as bulk or tandem metering), expanding the range of application to high-volume ramps.
- **Inclusion of variable message signs as control parameters** – While ramp meters represent one form of freeway control, improvements in freeway performance can also be achieved by providing information to drivers, for example, via variable message signs. Such information can be used to warn drivers of downstream bottlenecks, encouraging traffic diversion to less congested routes where additional capacity may be available. As a result, there is considerable merit in including variable message signs as control parameters to be optimized in the ramp metering algorithm. Should this option be pursued, it will also be necessary to upgrade the Freeway Traffic Model to better predict off-ramp diversion in response to traveller information.

- **Integration with the arterial network** – The framework for the new ramp control algorithm was specifically developed to allow for arterial network integration as a future enhancement. It is anticipated that such integration could be achieved via a feedback mechanism whereby conditions on the arterial network are used to adjust ramp metering rates on the freeway, and conditions on the freeway are used to adjust signal timing parameters on arterial roads to better manage the effects of traffic diversion resulting from ramp delays, freeway incidents, and other factors. A more sophisticated approach would be to incorporate arterial traffic control within the same framework as the ramp control algorithm so that all control parameters are determined simultaneously in an effort to optimize system-wide traffic flow from both equity and efficiency perspectives. For integration to be truly effective, the Freeway Traffic Model may need to be refined to better predict diversion behaviour resulting from control activities. In addition, to allow drivers to make informed route-choice decisions, variable message signs would ideally be deployed at all major decision points.

While freeway-arterial integration can take a number of different forms with varying levels of complexity, a reasonable first step is to ensure that the ramp control algorithm accounts for conditions on parallel routes to minimize negative impacts. In the most simplistic approach, the algorithm would operate in a strictly responsive mode; should conditions on designated arterial routes deteriorate below a given threshold,<sup>45</sup> the algorithm would automatically over-ride the metering rate by a certain pre-specified amount based on the location of the ramp in relation to the arterial deficiency and the magnitude of the ramp delays which may be triggering traffic diversion. Such an approach avoids the need to predict how drivers will respond to the adjusted rates,<sup>46</sup> but implies that arterial performance outweighs freeway performance. The risk lies in the time lag which occurs between noting the deficiency, taking a control action and observing the results of that action.<sup>47</sup>

Within the framework of the ramp control algorithm, it is also possible to compute the ‘optimal’ ramp metering rates adjusted for arterial conditions, considering the trade-off

---

<sup>45</sup> The established threshold should account for prevailing arterial conditions, so that arterials which already suffer from a poor level of service do not unnecessarily restrict metering efforts. Where such conditions exist, an acceptable strategy might be to allow ramp metering up to the point where diversion begins.

<sup>46</sup> Presumably, as ramp metering rates are adjusted, drivers will divert accordingly, particularly if information on ramp delays is made available via variable message signs. Such changes in behavior will be captured in the traffic sensor data used to compute the ramp metering rates in subsequent intervals.

<sup>47</sup> Not only do such delays highlight the need to adjust metering rates before minimum service standards are compromised (to allow time for changes in travel behaviour to take effect), but also the importance of establishing an over-ride strategy which minimizes inefficient “cycling” between higher and lower metering rates.

between arterial and freeway performance. However, doing so requires information on not only the current arterial state, but also the anticipated change as metering rates are adjusted.

In both the approaches described above, integration with the arterial network is achieved solely by adjusting the ramp metering rates on the freeway to respond to arterial conditions. In this type of arrangement, a separate traffic management system is required for the arterial network. Ideally, this system would employ some form of semi or fully actuated control, allowing signal timing parameters to be adjusted in real-time in response to traffic diversion from the freeway.

In implementing freeway-arterial integration, one of the major challenges lies in defining how arterial performance should be measured. The integration strategy should also be sensitive to cost constraints and the reality that it may not be feasible to upgrade all intersections along a given route to support integration requirements.

### *Next Steps*

Moving forward, a number of steps are needed before the algorithm can be successfully deployed in an existing network. For such deployment to become reality, partnerships will be needed between the research community tasked with further development of the algorithm, and transportation agencies seeking innovative approaches to managing the operation of their freeway infrastructure. Once such partnerships have been established, the next steps include the following:

1. Refine the model as necessary based on the recommendations provided above
2. Select a corridor to be used for field trials and conduct additional simulation studies to refine the key parameters
3. Implement the algorithm in the field and further fine-tune the parameters as required to meet the stated objectives

These tasks are considered essential for the effective deployment of the algorithm in real-world applications – the ultimate objective of this research work.

## 13 REFERENCES

- Abdulhai, B., and L. Kattan. 2004. *Network monitoring and traffic management & control: State of the art and state of the practice in Canada*. Discussion Paper prepared for the ATLANTIC Project, April 2004.  
<http://www.crt.umontreal.ca/atlantic/pdf/1-2Final.pdf>
- Alkadri, M. 1998. "Ramp metering: A Systems approach pilot survey of acceptability by freeway users." *ITE Journal*. August: 75-80.
- Arizona Department of Transportation (Arizona DOT). 2003. *Ramp meter design, operations, and maintenance guidelines*. Prepared by ITS Engineers and Constructors, Inc. for the Arizona DOT, Transportation Technology Group, August 2003. <http://www.azdot.gov/Highways/TTG/PDF/RampMeter-DesignGuide-0803.pdf>
- Atta-Armah, Richard. 1994. *Ramp metering methodology: The Ottawa-Carleton case*. M. Eng. Thesis, Carleton University, Ottawa, Canada.
- Banks, J. 1991. "Two-capacity phenomenon at freeway bottlenecks: A Basis for ramp metering?" *Transportation Research Record* 1320: 83-90.
- Banks, J. 2000. "Are minimization of delay and minimization of freeway congestion compatible ramp metering objectives?" *Transportation Research Record*. 1727: 112-119.
- Banks, J. 2002. "Review of empirical research on congested freeway flow." *Transportation Research Record* 1802: 225-232.
- Banks, J. 2005. "Metering ramps to divert traffic around bottlenecks: Some elementary theory." *Transportation Research Record*. 1925: 12-19.
- Bellemans, T., B. De Schutter, and B. De Moor. 2006a. "Model predictive control for ramp metering of motorway traffic: A Case study." *Control Engineering Practice*. 14: 757-767.
- Bellemans, T., B. De Schutter, G. Wets, and B. De Moor. 2006b. "Model predictive control for ramp metering combined with extended Kalman filter-based traffic state estimation." In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 406-411.
- Bertini, R., and A. Myton. 2005. "Use of performance measurement system data to diagnose freeway bottleneck locations empirically in Orange County, California." *Transportation Research Record* 1925: 48-57.
- Bertini, R., and S. Malik. 2004. "Observed dynamic traffic features on freeway section with merges and diverges." *Transportation Research Record* 1867: 25-35.

- Boel, R., and L. Mihaylova. 2004. "Modelling freeway networks by hybrid stochastic models." In *Proceedings of the 2004 IEEE Intelligent Vehicles Symposium*, 182-187.
- Boel, R., and L. Mihaylova. 2006. "A Compositional stochastic model for real time freeway traffic simulation." *Transportation Research Part B*. 40: 319-334.
- Brooks, S. P. 1998. "Markov Chain Monte Carlo method and its application". *The Statistician*. 47(1): 69-100.
- The BUGS Project Website. 2008. <http://www.mrc-bsu.cam.ac.uk/bugs/>
- California Center for Innovative Transportation (CCIT). 2001. *Ramp metering*. Posted on the ITS Decision Website at: <http://www.calccit.org/itsdecision/>
- Caltrans. 2000. *Ramp meter design manual*. Prepared by the Traffic Operations Program in cooperation with Design and Local Programs and the Department of California Highway Patrol, January 2000.  
[http://www.dot.ca.gov/hq/traffops/systemops/ramp\\_meter/RMDM.pdf](http://www.dot.ca.gov/hq/traffops/systemops/ramp_meter/RMDM.pdf)
- Cambridge Systematics, Inc. 2001. *Twin Cities ramp meter evaluation: Final report*. Prepared for the Minnesota Department of Transportation.
- Casella, G., and E. I. George. 1992. "Explaining the Gibbs sampler." *The American Statistician*. 46(3): 167-174.
- Cassidy, M. J., and S. Ahn. 2005. "Driver turn-taking behavior in congested freeway merges." *Transportation Research Record* 1934: 140-147.
- Chaudhary, N. A., and C. J. Messer. 2002. "Freeway on-ramp design criteria for ramp meters with excessive queue detectors." *Transportation Research Record*. 1796: 80-85.
- Cheng, J., and M. J. Druzdzel. 2000. "AIS-BN: An Adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks." *Journal of Artificial Intelligence Research*. 13(2000): 155-188.
- Chib, S., and E. Greenberg. 1995. "Understanding the Metropolis-Hastings algorithm." *The American Statistician*. 49(4): 327-335.
- Chu, L., H. X., Liu, and W. Recker. 2004. "Using microscopic simulation to evaluate potential intelligent transportation system strategies under nonrecurrent congestion." *Transportation Research Record*. 1886: 76-84.
- Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. 1999. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Series on Statistics for Engineering and Information Science. New York: Springer-Verlag.

- Daganzo, C. 1994. "The cell transmission model: A Dynamic representation of highway traffic consistent with the hydrodynamic theory." *Transportation Research Part B*. 28(4): 269-287.
- Davis, G., N. Nihan, M. Hamed, and L. Jacobson. 1990. "Adaptive forecasting of freeway traffic congestion." *Transportation Research Record*. 1287: 29-33.
- Delcan, CH2MHill, and Golder Associates. 2006. *Lane allocation position paper*. Program Level Technical Report. Prepared for the British Columbia Gateway Program, February 2006. [http://www.th.gov.bc.ca/gateway/reports/pdr-supplane\\_allocation\\_tech\\_rpt\\_feb2006.pdf](http://www.th.gov.bc.ca/gateway/reports/pdr-supplane_allocation_tech_rpt_feb2006.pdf)
- Delquié, P. 2008. *A Tutorial guide to using ASSESS*. <http://faculty.insead.edu/delquie/ASSESS.htm>
- Dowling, R., A. Skabardonis, and V. Alexiadis. 2004. *Traffic analysis toolbox*. Volume III: Guidelines for applying traffic microsimulation software. Prepared for the Federal Highway Administration, Office of Operations by Dowling Associates, Oakland, California, June 2004. FHWA Report No.: FHWA-HRT-04-040.
- Dowling, R., J. Holland, and A. Huang. 2002. *Guidelines for applying traffic microsimulation modeling software*. Prepared for the California Department of Transportation by Dowling Associates, Oakland, California, September 2002.
- Doucet, A., N. de Freitas, and N. Gordon. 2001. "An Introduction to sequential Monte Carlo methods." In *Sequential Monte Carlo Methods in Practice*. Editors: Doucet, A., N. de Freitas, and N. Gordon. Series: Statistics for Engineering and Information Science. Springer: 1-14.
- Edwards, W., R. F. Miles Jr., and D. von Winterfeldt, editors. 2007. *Advances in decision analysis: From foundations to applications*. New York, New York: Cambridge University Press.
- Elefteriadou, L., R. Roess, and W. McShane. 1995. "Probabilistic nature of breakdown at freeway merge junctions." *Transportation Research Record*. 1484: 80-89.
- Evans, J., L. Elefteriadou, and N. Gautam. 2001. "Probability of breakdown at freeway merges using Markov chains." *Transportation Research Part B*. 35: 237-254.
- Foo, S., 2006. *Advanced traffic management systems information infrastructure: The ITS Centre and Testbed (ICAT) platform*. A thesis submitted for the degree of Doctor of Philosophy, Graduate Department of Civil Engineering, University of Toronto.
- Forbes, J., T. Huang, K. Kanazawa, and S. Russell. 1995. "The BATmobile: Towards a Bayesian automated taxi." *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1878-1885.

- Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. *Bayesian data analysis*. 2<sup>nd</sup> Edition. Boca Raton, Florida: Chapman & Hall/CRC.
- Gomes, G., and R. Horowitz. 2006. "Optimal freeway ramp metering using the asymmetric cell transmission model." *Transportation Research Part C*. In press.
- Gordon, L. R. 1996. "Algorithm for controlling spillback from ramp meters." *Transportation Research Record*. 1554: 162-171.
- Gordon, N., D. Salmond, and A. Smith. 1993. "Novel approach to nonlinear / non-Gaussian Bayesian state estimation." *IEE Proceedings-F*. 140(2): 107-113.
- Guo, H., and W. Hsu. 2002. "A Survey of algorithms for real-time Bayesian network inference". Presented at the joint *AAAI/KDD/UAI 2002 Workshop on Real-Time Decision Support and Diagnosis Systems*, Edmonton, Alberta, Canada, July 29, 2002.
- Hall, F., and K. Agyemang-Duah. 1991. "Freeway capacity drop and the definition of capacity." *Transportation Research Record*. 1320: 91-98.
- Hall, F., V. Hurdle, and J. Banks. 1992. "Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways." *Transportation Research Record*. 1365: 12-18.
- Hegy, A, D. Girimonte, R. Babuska, and B. De Schutter. 2006. "A Comparison of filter configurations for freeway traffic state estimation." Proceedings of the *IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, September 2006, 1029-1034.
- Hellinga, B., and M. Van Aerde. 1995. "Examining the potential of using ramp metering as a component of ATMS." *Transportation Research Record*. 1494: 75-83.
- Huang, T., D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russel, and J. Weber. 1994. "Automatic symbolic traffic scene analysis using belief networks." Proceedings of the *12th National Conference on Artificial Intelligence*, 966-972.
- Institute of Transportation Engineers (ITE). 1984. "Displays for metered freeway entrance ramps." Proposed Recommended Practice by ITE Technical Committee 4M-11. *ITE Journal*. April: 14-18.
- ITS Canada. No date. *ITS applications in Canada*. Internet Site. <http://www.itscanada.ca/english/applications.htm>
- Jacob, C., and B. Abdulhai. 2005. "Integrated traffic corridor control using machine learning." Proceedings of the *IEEE International Conference on Systems, Man and Cybernetics*, 3460-3465.

- Jacobson, E. L., and J. Landsman. 1994. "Case studies of U.S. freeway-to-freeway ramp and mainline metering and suggested policies for Washington State." *Transportation Research Record*. 1446: 48-55.
- Jensen, F. V. 2001. *Bayesian networks and decision graphs*. Statistics for Engineering and Information Science. New York: Springer-Verlag.
- Kanazawa, K., D. Koller, and S. Russell. 1995. "Stochastic simulation algorithms for dynamic probabilistic networks." Proceedings of the 11<sup>th</sup> Annual Conference on Uncertainty in Artificial Intelligence, 346-351.
- Karimi, A., A. Hegyi, B. De Schutter, J. Hellendoorn, and F. Middelham. 2004. "Integrated model predictive control of dynamic route guidance information systems and ramp metering." Proceedings of the 2004 IEEE Intelligent Transportation Systems Conference, Washington D.C., 491-496.
- Keeney, R. L., and H. Raiffa. 1993. *Decisions with multiple objectives: Preferences and value tradeoffs*. New York, New York: Cambridge University Press.
- Kerner, B., and S. Klenov. 2006. "Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic." *Physica A*. 364: 473-492.
- Klein, L. A., and M. R. Kelley. 1996. *Detection technology for IVHS*. Volume 1: Final Report. Performed by Hughes Aircraft Company for the U.S. Department of Transportation, Federal Highway Administration, Office of Engineering & Highway Operations R&D. FHWA Report No.: FHWA-RD-95-100. [http://ntl.bts.gov/lib/jpodocs/repts\\_te/6184.pdf](http://ntl.bts.gov/lib/jpodocs/repts_te/6184.pdf)
- Korb, K. B., and A. E. Nicholson. 2004. *Bayesian artificial intelligence*. Series in Computer Science and Data Analysis. Boca Raton, Florida: Chapman & Hall/CRC.
- Kosmatopoulos, E., M. Papageorgiou, D. Manolis, J. Hayden, R. Higginson, K. McCabe, and N. Rayman. 2006. "Real-time estimation of critical occupancy for maximum motorway throughput." *Transportation Research Record* 1959: 65-76.
- Kotsialos, A., and M. Papageorgiou. 2004. "Efficiency and equity properties of freeway network-wide ramp metering with AMOC." *Transportation Research Part C*. 12(2004): 401-420.
- Kotsialos, A., M. Papageorgiou, C. Diakaki, Y. Pavlis, and F. Middelham. 2002. "Traffic flow modeling of large-scale motorway networks using the macroscopic modeling tool METANET." *IEEE Transactions on Intelligent Transportation Systems*. 3(4): 282-292.

- Kotsialos, A., M. Papageorgiou, M. Mangeas, and H. Haj-Salem. 2002. "Coordinated and integrated control of motorway networks via non-linear optimal control." *Transportation Research Part C*. 10: 65-84.
- Kuhne, R., R. Mahnke, I. Lubashevsky, and J. Kaupuzs. 2002. "Probabilistic description of traffic breakdowns." *Physical Review E*. 65: 1-13.
- Kwon, J., and K. Murphy. 2000. *Modeling freeway traffic with coupled HMMs*. Technical Report. University of California, Berkeley.
- Lee, C., B. Hellenga, and K. Ozbay. 2006. "Quantifying effects of ramp metering on freeway safety." *Accident Analysis and Prevention*. 38: 279-288.
- Levinson, D. 2002. "Identifying winners and losers in transportation." *Transportation Research Record*. 1812: 179-185.
- Levinson, D. 2003. "Perspectives on efficiency in transportation." *International Journal of Transport Management*. 1 (2003): 145-155.
- Levinson, D., K. Harder, J. Bloomfield, and K. Carlson. 2006. "Waiting tolerance: Ramp delay vs. freeway congestion." *Transportation Research Part F*. 9 (2006): 1-13.
- Lindley, D. V. 1985. *Making decisions*. 2<sup>nd</sup> Edition. Toronto, Ontario: John Wiley & Sons.
- Logical Decisions. 2009. *Logical Decisions v6.2*. Help Documentation. <http://www.logicaldecisions.com>
- Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. "WinBUGS -- A Bayesian modelling framework: Concepts, structure, and extensibility." *Statistics and Computing*. 10: 325-337.
- MacCarley, C. A., S. P. Mattingly, M. G. McNally, D. Mezger, and J. E. Moore II. 2002. "Field operational test of integrated freeway ramp metering / arterial adaptive signal control: Lessons learned in Irvine, California." *Transportation Research Record*. 1811: 76-83.
- Marshall, K. T., and R. M. Oliver. 1995. *Decision making and forecasting with emphasis on model building and policy analysis*. Toronto, Ontario: McGraw-Hill, Inc.
- Martin, P. T., and Y. Feng. 2003. *Detector technology evaluation*. Department of Civil and Environmental Engineering, University of Utah Traffic Lab, Salt Lake City, Utah, November 2003.
- The MathWorks, Inc. 2010. *MATLAB Help*. Version R2010a.

- McLean, T., C. Brader, C. Diakaki, and M. Papageorgiou. 1998. "Urban integrated traffic control in Glasgow, Scotland." Presented at the *1998 IEE Conference on Road Transport Information and Control*, Conference Publication 454: 243-249.
- Meng, Q., and H. L. Khoo. 2010. "A Pareto-optimization approach for a fair ramp metering." *Transportation Research Part C*. 18(2010): 489-506.
- Middleton, D., R. Longmire, and S. Turner. 2007. *State of the art evaluation of traffic detection and monitoring systems*. Volume I – Phases A & B: Design. Final Report 627(1). Prepared by the Texas Transportation Institute, Texas A&M University for the Arizona Department of Transportation in cooperation with the U.S. Department of Transportation, Federal Highway Administration, October, 2007.
- Mihaylova, L., R. Boel, and A. Hegyi. 2007. "Freeway traffic estimation within particle filtering framework." *Automatica*. 43: 290-300.
- Minka, T. 2007. *Bayesian inference in dynamic models – An Overview*.  
<http://research.microsoft.com/~minka/papers/dynamic.html>
- Minnesota Department of Transportation (Mn DOT). 2001. *Road design manual (metric)*. Design Standards Unit, Office of Technical Support.  
<http://www.dot.state.mn.us/design/rdm/index.html>
- Minnesota Department of Transportation (Mn DOT). 2009. *Traffic engineering manual*. Office of Traffic, Safety, and Technology, October 2009.  
<http://www.dot.state.mn.us/trafficeng/publ/tem/index.html>
- Murphy, K. 1998. *A Brief introduction to graphical models and Bayesian networks*.  
<http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>
- Murphy, K. 2002. *Dynamic Bayesian networks*. Draft book chapter posted on-line at:  
<http://www.cs.ubc.ca/~murphyk/Papers/dbnchapter.pdf>
- Neapolitan, R. 2004. *Learning Bayesian networks*. Upper Saddle River, New Jersey: Pearson Prentice Hall.
- Needham, C., J. Bradford, A. Bulpitt, and D. Westhead. 2007. "A Primer on learning in Bayesian Networks for computational biology." *PLoS Computational Biology*. 3(8): 1409-1416.
- Nevada Department of Transportation (Nevada DOT). 2006. *HOV/Managed lanes and ramp metering design manual*. Prepared by Parsons Brinckerhoff for the Nevada DOT in cooperation with the Regional Transportation Commission of Southern Nevada, March 2006.  
[http://www.nevadadot.com/reports\\_pubs/HOV/pdfs/HOV\\_DesignManual.pdf](http://www.nevadadot.com/reports_pubs/HOV/pdfs/HOV_DesignManual.pdf)

- Ngoduy, D. 2008. "Applicable filtering framework for online multiclass freeway network estimation." *Physica A*. 387: 599-616.
- Norsys Software Corporation. 2007. *Netica 3.24*. Help Documentation.
- O'Brien, A. 2000. "New ramp metering algorithm improves systemwide travel time." *TR News*. 209: 38-39.
- Ontario Ministry of Transportation (MTO). 2010. *About COMPASS*. Internet Site. <http://www.mto.gov.on.ca/english/traveller/trip/compass-sio.shtml>
- Ozbay, K., I. Yasar, and P. Kachroo. 2006. "Improved online estimation methods for a feedback-based freeway ramp metering strategy." Proceedings of the *IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, September 2006, 412-417.
- Ozguven, E., and K. Ozbay. 2008. "Nonparametric Bayesian estimation of freeway capacity distribution from censored observations." *Transportation Research Record*. 2061: 20-29.
- Papageorgiou, M., H. Hadj-Salem, and F. Middelham. 1997. "ALINEA local ramp metering: Summary of field results." *Transportation Research Record*. 1603: 90-98.
- Papageorgiou, M., H. Hadj-Salem, and J-M. Blosseville. 1991. "ALINEA: A Local feedback control law for on-ramp metering." *Transportation Research Record*. 1320: 58-64.
- Papamichail, I., M. Papageorgiou, V. Vong, and J. Gaffney. 2010. "Heuristic ramp-metering coordination strategy implemented at Monash freeway, Australia." *Transportation Research Record*. 2178: 10-20.
- Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, California: Morgan Kaufmann Publishers, Inc.
- Persaud, B., S. Yagar, D. Tsui, and H. Look. 2001. "Breakdown-related capacity for freeway with ramp metering." *Transportation Research Record*. 1748: 110-115.
- Persaud, B., S. Yagar, and R. Brownlee. 1998. "Exploration of the breakdown phenomenon in freeway traffic." *Transportation Research Record*. 1634: 64-69.
- Ristic, B., S. Arulampalam, and N. Gordon. 2004. *Beyond the Kalman filter: Particle filters for tracking applications*. Boston, Massachusetts: Artech House.
- Russell, S. J., and P. Norvig. 1995. *Artificial intelligence: A Modern approach*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

- Sarvi, M., M. Kuwahara, and A. Ceder. 2004. "Freeway ramp merging phenomena in congested traffic using simulation combined with a driving simulator". *Computer-Aided Civil and Infrastructure Engineering*. 19(2004): 351-363.
- Smaragdis, E., and M. Papageorgiou. 2003. "Series of new local ramp metering strategies." *Transportation Research Record*. 1856: 74-86.
- Smith, S. A., and C. Perez. 1992. "Evaluation of INFORM: Lessons learned and application to other systems." *Transportation Research Record*. 1360: 62-65.
- Son, B., T. Kim, H. J. Kim, and S. Lee. 2004. "Probabilistic model of traffic breakdown with random propagation of disturbance for ITS application." In *Knowledge-based Intelligent Information and Engineering Systems*. Lecture Notes in Artificial Intelligence. 3215: 45-51.
- Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2003. *WinBUGS user manual*. Version 1.4. January 2003.
- Stephanedes, Y., and E. Kwon. 1993. "Adaptive demand-diversion prediction for integrated control of freeway corridors." *Transportation Research Part C*. 1(1): 23-42.
- Sun, X., and R. Horowitz. 2006. "Set of new traffic-responsive ramp-metering algorithms and microscopic simulation results." *Transportation Research Record*. 1959: 9-18.
- Sun, X., L. Munoz, and R. Horowitz. 2003. "Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control." In *Proceedings of the 42<sup>nd</sup> IEEE Conference on Decision and Control*, 6333-6338.
- Taylor, C., D. Meldrum, and L. Jacobson. 1998. "Fuzzy ramp metering: Design overview and simulation results." *Transportation Research Record*. 1634: 10-18.
- Tian, Z. Z., K. Balke, R. Engelbrecht, and L. Rilett. 2002. "Integrated control strategies for surface street and freeway systems." *Transportation Research Record*. 1811: 92-99.
- Totten Sims Hubicki Associates (TSH). 2007. *Highway 417 (Ottawa Queensway) from Highway 416 Easterly to Anderson Road Preliminary Design Study and Environmental Assessment: Transportation Environmental Study Report*. Prepared for the Ontario Ministry of Transportation, January 2007.
- TRANS Committee. 2011. *2007 Interprovincial roadside truck survey: Summary of results*. National Capital Region, Canada, February 2011.
- Transportation Research Board (TRB). 2000. *Highway capacity manual*. Washington, DC: TRB, National Research Council.

- Transport Canada. 2006. *Transport Canada releases first, systematic analysis of cost of urban traffic congestion in Canada*. News Release Dated March 22, 2006. <http://www.tc.gc.ca/mediaroom/releases/nat/2006/06-h006e.htm>
- Transports Québec. 2007. *Feux de gestion d'accès*. Internet Site. [http://www.mtq.gouv.qc.ca/portal/page/portal/grand\\_public/vehicules\\_promenade/reseau\\_routier/signalisation/signaux\\_lumineux/feux\\_gestion\\_acces](http://www.mtq.gouv.qc.ca/portal/page/portal/grand_public/vehicules_promenade/reseau_routier/signalisation/signaux_lumineux/feux_gestion_acces)
- U.S. Department of Transportation (U.S. DOT). 1995. *Ramp metering status in North America: 1995 Update*. Washington, DC: U.S. DOT, Federal Highway Administration. [http://ntl.bts.gov/lib/jpodocs/repts\\_pr/3725.pdf](http://ntl.bts.gov/lib/jpodocs/repts_pr/3725.pdf)
- U.S. Department of Transportation (U.S. DOT). 2003. *Freeway management and operations handbook*. Washington, DC: U.S. DOT, Federal Highway Administration, Office of Transportation Management. FHWA Report No.: FHWA-OP-04-003. [http://ops.fhwa.dot.gov/freewaymgmt/publications/frwy\\_mgmt\\_handbook/index.htm](http://ops.fhwa.dot.gov/freewaymgmt/publications/frwy_mgmt_handbook/index.htm)
- U.S. Department of Transportation (U.S. DOT). 2004. *Manual on uniform traffic control devices*. 2003 Edition with Revision No. 1 Incorporated, dated November 2004 (HTML). Washington, DC: U.S. DOT, Federal Highway Administration, Office of Transportation Operations. <http://mutcd.fhwa.dot.gov>
- U.S. Department of Transportation (U.S. DOT). 2006a. *Coordinated freeway and arterial operations handbook*. McLean, VA: U.S. DOT, Federal Highway Administration, Turner-Fairbank Highway Research Center. FHWA Report No.: FHWA-HRT-06-095. <http://www.fhwa.dot.gov/publications/research/operations/its/06095/>
- U.S. Department of Transportation (U.S. DOT). 2006b. *Ramp management and control handbook*. Washington, DC: U.S. DOT, Federal Highway Administration, Office of Transportation Management. FHWA Report No.: FHWA-HOP-06-001. [http://ops.fhwa.dot.gov/publications/ramp\\_mgmt\\_handbook/manual/manual/default.htm](http://ops.fhwa.dot.gov/publications/ramp_mgmt_handbook/manual/manual/default.htm)
- U.S. Department of Transportation (U.S. DOT). 2008. *ITS deployment statistics database*. Hosted on-line by the U.S. DOT, Research and Innovative Technology Administration. <http://www.itsdeployment.its.dot.gov>
- Van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan. 2000. *The Unscented particle filter*. Technical Report, Cambridge University Engineering Department.
- Van Katwijk, R., and P. Van Koningsbruggen. 2002. "Coordination of traffic management instruments using agent technology." *Transportation Research Part C*. 10(2002): 455-471.
- Vlahogianni, E., J. Golias, and M. Karlaftis. 2004. "Short-term traffic forecasting: Overview of objectives and methods." *Transport Reviews*. 24(5): 533-557.

- Wang, F. Y. 2003. "Integrated intelligent control and management for urban traffic systems." *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*. Vol. 2. 1313-1317.
- Wang, Y., and M. Papageorgiou. 2005. "Real-time freeway traffic state estimation based on extended Kalman filter: A General approach." *Transportation Research Part B*. 39: 141-167.
- Wang, Y., M. Papageorgiou, A. Messmer, P. Coppola, A. Tzimitsi, and A. Nuzzolo. 2009. "An adaptive freeway traffic state estimator." *Automatica*. 45:10-24.
- Wu, J., and G. L. Chang. 1999. "Heuristic method for optimal diversion control in freeway corridors." *Transportation Research Record*. 1667: 8-15.
- Wu, J., M. McDonald, and K. Chatterjee. In press. "A detailed evaluation of ramp metering impacts on driver behaviour." *Transportation Research Part F*.
- Xin, W., P. G. Michalopoulos, J. Hourdakis, and D. Lau. 2004. "Minnesota's new ramp control strategy: Design overview and preliminary assessment." *Transportation Research Record*. 1867: 69-79.
- Victoria Transport Policy Institute (VTPI). 2011. "Equity evaluation." *Online TDM Encyclopedia*. <http://www.vtpi.org/tdm/tdm13.htm>
- Yin, Y., H. Liu, and H. Benouar. 2004. "A note on equity of ramp metering." *Proceedings of the IEEE Intelligent Transportation Systems Conference*, Washington, D.C., October 2004, 497-502.
- Yongjun, Z., L. Wenjun, S. Bin, and J. Yanhua. 2009. "An Unscented particle filter approach to estimating real-time traffic state." *Proceedings of the 2009 International Conference on Measuring Technology and Mechatronics Automation*, 471-474.
- Yudkowsky, E. No date. *An Intuitive explanation of Bayesian reasoning*. <http://yudkowsky.net/bayes/bayes.html>
- Zhang, L., and D. Levinson. 2004a. "Optimal freeway ramp control without origin-destination information." *Transportation Research Part B*. 38(2004): 869-887.
- Zhang, L., and D. Levinson. 2004b. "Some properties of flows at freeway bottlenecks." *Transportation Research Record* 1883: 122-131.
- Zhang, L., and D. Levinson. 2005. "Balancing efficiency and equity of ramp meters." *Journal of Transportation Engineering*. 131(6): 477-481.
- Zhang, M., T. Kim, X. Nie, W. Jin, L. Chu, and W. Recker. 2001. *Evaluation of On-Ramp Control Algorithms*. California Path Program, Institute of Transportation Studies, University of California, Berkeley.

# **APPENDIX A**

## **DESIGN OF RAMP METERING SYSTEMS**

## A. DESIGN OF A RAMP METERING SYSTEM

The design of a ramp metering system is influenced by many factors, including the objectives of the transportation agency, prevailing traffic and geometric conditions, and local policy and standards. The *Ramp Management and Control Handbook* developed by the U.S. Federal Highway Administration (FHWA) provides guidance and recommended practice for managing and controlling traffic on freeway ramps (U.S. DOT 2006b). In particular, the handbook provides an excellent introduction to the concepts and issues relevant to the design of effective ramp metering systems, and forms the basis for much of the discussion which follows. Other key references include the FHWA's *Freeway Management and Operations Handbook* (U.S. DOT 2003) which contains a chapter on ramp management and control, and the U.S. *Manual on Uniform Traffic Control Devices* (MUTCD) which provides warrants for ramp meter installation and design standards for freeway entrance ramp control signals (U.S. DOT 2004).

### A.1 Typical Ramp Metering Installation

Figure A-1 illustrates the typical configuration of a single-lane on-ramp with a ramp meter. In general, the length of the ramp and position of the meter should allow for:

- Sufficient distance for a stopped vehicle at the meter to accelerate and attain safe merging speed (acceleration distance)
- Sufficient room for vehicles approaching the meter to safely stop behind the queue of vehicles waiting to enter the freeway (safe stopping distance)
- Sufficient storage for vehicles queued at the meter to prevent spillback onto the arterial system (storage distance) (Chaudhary and Messer 2002)

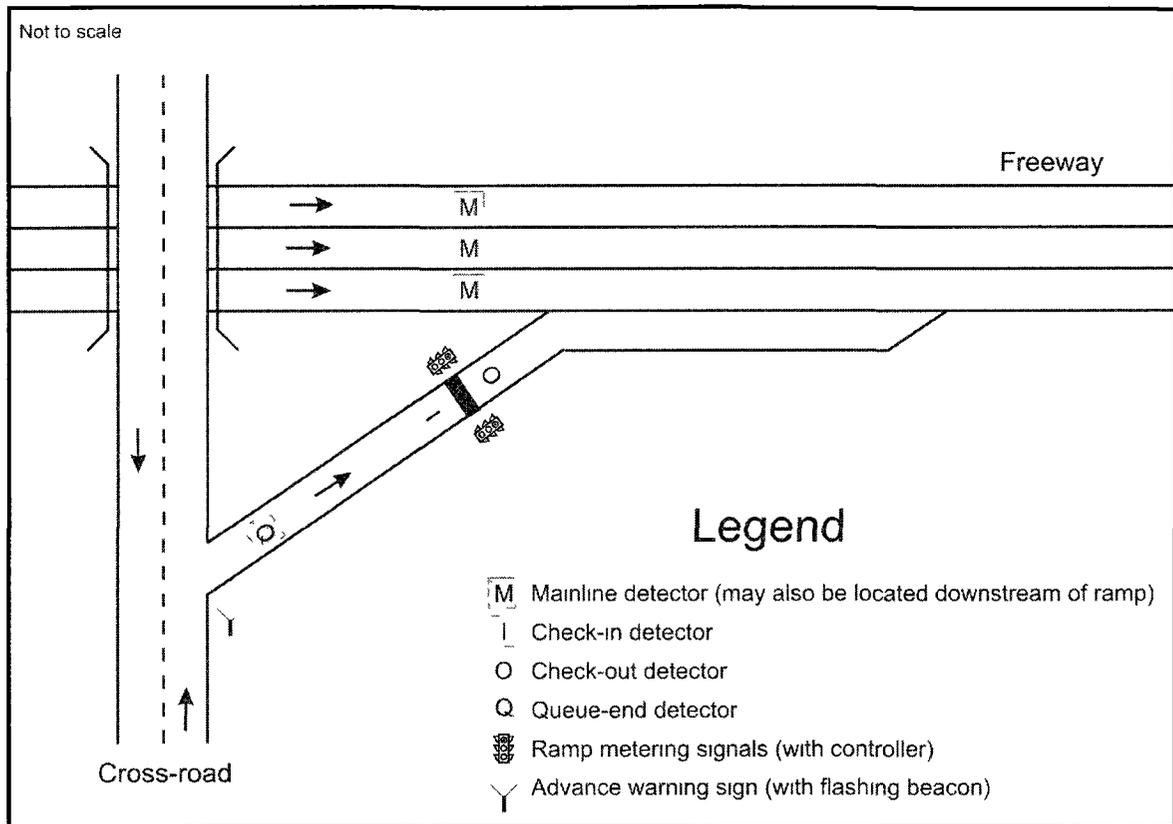
Depending on the control strategy, a number of detectors are needed to track vehicle movement through the metered zone:

- Demand (or check-in) detectors are installed on the ramp immediately in front of the stop bar. These detectors sense a vehicle's presence at the stop bar and initiate the green signal in accordance with the metering algorithm.
- Passage (or check-out) detectors are installed just beyond the stop bar. These detectors monitor the number of vehicles that enter the freeway, including any vehicles that violate the signal display. Since passage detectors track the number of vehicles crossing the stop bar during the green phase, they can be used to determine when the green phase should be terminated; once the requisite number of vehicles have been detected, a green-to-red signal change is triggered.
- Queue detectors are used to prevent ramp queues from spilling back onto adjacent roads, disrupting operations. One or several detectors may be employed to monitor queue growth and identify when queued vehicles are in danger of exceeding the available storage capacity. When excessive queue lengths are

detected, the metering rate is typically increased (allowing more vehicles to enter the freeway) until the queue length is reduced to an acceptable level.

- Mainline detectors provide information on freeway occupancy, speed, and volume. This information is used to determine the ramp metering rates in traffic-responsive systems. At isolated, non-coordinated ramps, data from mainline detectors immediately upstream or downstream of the ramp are used to establish the local traffic-responsive metering rate; in system-wide implementations, data from each mainline detector along the corridor is transferred to a centralized controller for determining coordinated metering rates. Note that depending on the metering algorithm, mainline detectors may be required upstream or downstream of the entrance ramp, or both.
- Exit ramp detectors (not shown in Figure A-1) are often installed to provide traffic flow data for the metering algorithm, particularly in the case of system-wide traffic-responsive systems. Likewise, detectors may also be installed at entrance ramps without ramp meters to provide a complete picture of corridor traffic volumes for use in determining optimal metering rates at a system-wide level.
- Merge detectors (not shown in Figure A-1) are used to detect vehicles stored in the merging area of the ramp. These detectors are most commonly used in locations where the geometry of the merge area is poor. If merging problems are detected, the controller may hold back vehicles at the ramp meter.

Other important features that must be addressed in the design of a ramp metering system include pavement markings, signage, and traffic enforcement areas. The type and placement of signal displays is also a key consideration, as well as the selection of controller and communications equipment. Specific guidance on each of these design issues can be found in the *Ramp Management and Control Handbook* referenced above (U.S. DOT 2006b). The Institute of Transportation Engineers' proposed recommended practice for metered entrance ramp displays also provides a summary of some of the key issues (ITE 1984).



Adapted From: *Ramp Management and Control Handbook*, Figure 10-2 (U.S. DOT 2006b)

**Figure A-1 Typical Ramp Metering Installation**

## A.2 Flow Control

Vehicle flow through the ramp meter can be characterized according to the number of lanes on the ramp and the number of vehicles from each lane allowed to enter the freeway during the green interval.

- **Single lane ramp, one vehicle per green** – Under this type of flow control, vehicles enter the freeway one vehicle at a time. When a vehicle is detected at the ramp meter, a green phase is initiated in accordance with the current metering rate. Once the vehicle travels over the passage detector, the green phase is terminated. Given the minimum time required to process a vehicle, this type of metering has a capacity of roughly 900 vehicles per hour. If a higher throughput is needed, consideration must be given to either installing an additional lane (tandem metering) or allowing multiple vehicles to go at a time (bulk metering).
- **Single lane ramp, multiple vehicles per green** – Under this approach, also known as platoon or bulk metering, two or more vehicles are allowed to enter the freeway during each green interval. Since longer cycle lengths are required to accommodate the additional vehicles, the maximum throughput under this

approach increases by only 200 to 400 vehicles per hour compared to the more restrictive strategy of allowing only one vehicle per green.

- **Multi-lane ramp, one vehicle per lane per green** – Under this type of flow control, also called tandem or two-abreast metering, two (or more) vehicles are allowed to enter the freeway during the green interval, one from each lane. To reduce conflicts during merging, the green interval for each lane is often offset so that vehicles are released in a staggered manner. In cases of exceptionally high demand, tandem metering may be combined with bulk metering, allowing multiple vehicles to enter the freeway from each lane during the green interval.

The selection of an appropriate flow control strategy is closely linked to the number of lanes on the ramp, which in turn depends on the ramp width (and any widening constraints), the presence of shoulders (for temporary storage of queued vehicles during ramp meter operation), and the provision of transit or high-occupancy vehicle (HOV) ramp meter by-pass lanes. Together, the lane configuration and associated flow control influence the maximum flow rate onto the freeway. The choice of flow control and ramp cross-section should therefore be determined based on the anticipated traffic volume, associated queue storage requirements, and ramp length available to accommodate queuing. Table A-1 provides a general guide to selecting the lane configuration and ramp meter release rate based on the ramp volume.

**Table A-1 Selection of an Appropriate Lane Configuration and Control Strategy**

<b>If ramp volume is...</b>	<b>...then consider this number of metered lanes</b>	<b>...with this release rate</b>
<1000 veh/h	One lane	One vehicle per green
900 – 1,200 veh/h	One lane	Two vehicles per green
1,200 – 1,600 veh/h	Two lanes	One vehicle per green
1,600 – 1,800 veh/h	Two lanes	Two vehicles per green

*Adapted From: Ramp Management and Control Handbook, Table 10-1 (U.S. DOT 2006b)*

### ***Minimum Metering Rate***

The minimum ramp metering rate is a function of the maximum time allowed between successive green phases. Since violations tend to increase as the cycle length increases, there is a practical limit to the time interval between one vehicle departure and the next. According to the *Ramp Management and Control Handbook* (U.S. DOT 2006b), motorists who wait longer than 15 seconds for a green signal begin to believe that the ramp meter is not working properly, leading to decreased compliance. Thus, assuming a maximum cycle length of 15 seconds, the minimum ramp metering rate is 240 vph, based

on a single-lane ramp with one vehicle allowed per green (3600 sec/hr divided by 15 sec/cycle).

### ***Maximum Metering Rate***

The maximum discharge rate represents the maximum number of vehicles that can be processed when ramp metering is in effect, and in some ways, can be considered the “capacity” of the metered ramp. The maximum metering rate depends on the minimum time required for a full cycle. Typically, 4 seconds are needed to process a single vehicle, including the green, yellow, and all-red phases (2.5 seconds of yellow / all red plus 1.5 seconds of green). Based on this minimum cycle length, the maximum discharge rate for a single metered lane is roughly 900 vph (3600 sec/hr divided by 4 sec/cycle).

### **A.3 Ramp Metering Strategies**

Ramp metering strategies can be categorized according to their responsiveness to changing traffic conditions, and extent of coordination among individual ramps.

- Responsiveness
  - **Fixed-time systems** use pre-set metering rates based on ‘typical’ conditions, whereas **traffic responsive systems** use real-time data to determine the control parameters.

In a fixed-time system, static metering rates are specified for each ramp, calculated using historical data. Typically, different rates are used for different times of the day to reflect daily traffic patterns, with the meters activated according to pre-set schedules. In comparison, traffic responsive systems calculate metering rates in real-time using data from traffic sensors. Metering is initiated in response to observed traffic conditions unless superseded by policy considerations.
- Coordination
  - Under **local (or isolated) control**, the ramp metering rates used at each ramp are determined independently, based solely on the traffic conditions immediately upstream or downstream of the ramp. This type of control is most appropriate for isolated safety and operational deficiencies that are confined to a single location.
  - Under **system-wide (or coordinated) control**, ramp metering rates are calculated centrally based on conditions over an extended area. As a result, the metering rates applied at individual ramps can be coordinated to address broader operational issues.

While fixed-time metering systems are the simplest and least expensive to install, they cannot respond to real-time traffic conditions, including non-recurrent congestion. As a result, such systems are most suited to locations with predictable traffic patterns. However, even in cases where congestion can be accurately predicted, metering rates based on historical data will often be slightly higher or lower than required due to random fluctuations in traffic flow, leading to unnecessary ramp delays or freeway congestion. To reduce such impacts, it is important that fixed-time metering rates be updated frequently.

In contrast, traffic responsive strategies calculate ramp metering rates dynamically using information from freeway detectors that monitor traffic conditions in real-time. These strategies are therefore responsive to both recurring and non-recurring congestion, and can adapt to changes in the traffic environment, providing enhanced performance compared to their fixed-time counterparts.

Both fixed-time and traffic responsive systems can be applied locally or system-wide. Local traffic responsive systems calculate metering rates based on conditions in the immediate vicinity of the ramp, optimizing flow over a localized area without regard for interaction between upstream and downstream segments. In comparison, system-wide traffic responsive strategies are designed to improve operational performance over an extended freeway section, and thus require information from multiple ramp and freeway detectors to fully capture current conditions along the entire section length. Since travel inputs are known over a wide area, coordinated systems can predict potential issues before they arise and take appropriate action, rather than operating in a strictly responsive mode. Not surprisingly, these types of systems have the most complex hardware requirements, and are therefore the most expensive to install. In the event of a communications failure, system-wide ramp metering strategies typically revert to local control to ensure continuous operation of the system.

Table A-2 provides a summary of the advantages and disadvantages of the various ramp metering strategies. In general, system-wide traffic responsive systems are considered to offer the greatest potential for optimizing freeway performance within a congested corridor.

**Table A-2 Advantages and Disadvantages of Different Ramp Metering Approaches**

<b>Metering Approach</b>	<b>Advantages</b>	<b>Disadvantages</b>
Pre-Timed (Local & System-Wide)	<ol style="list-style-type: none"> <li>1. No mainline detection devices are needed.</li> <li>2. No communication with a TMC is required.</li> <li>3. Simple hardware configuration compared to other approaches.</li> <li>4. Provides safety benefit by breaking up platoons of vehicles entering the freeway.</li> <li>5. Can effectively relieve recurring congestion if it is fairly constant day-after-day.</li> </ol>	<ol style="list-style-type: none"> <li>1. Requires frequent observations so rates can be adjusted to changing traffic conditions.</li> <li>2. Often results in over restrictive metering rates leading to unneeded ramp queuing and delays (unless metering at demand is employed), which could affect arterial operations as well.</li> <li>3. Not responsive to unusual conditions, such as non-recurring congestion, which in turn can lead to public dissatisfaction.</li> </ol>
Local Traffic Responsive	<ol style="list-style-type: none"> <li>1. Ability to better manage freeway congestion than pre-timed metering approaches (especially for non-recurring congestion).</li> <li>2. Operating costs are lower than pre-timed (due to automatic, rather than manual, meter adjustments), so the extra investment upfront may pay itself off over time.</li> </ol>	<ol style="list-style-type: none"> <li>1. Higher capital and maintenance costs than pre-timed.</li> <li>2. Increased maintenance needs because of mainline detection.</li> <li>3. Reactive versus proactive. In other words, improvements are made after the fact, rather than before problems occur.</li> <li>4. Doesn't consider conditions beyond the adjacent freeway section, making it difficult to optimize conditions for a downstream bottleneck.</li> </ol>
System-Wide Traffic Responsive	<ol style="list-style-type: none"> <li>1. Provides optimal metering rates based on real-time conditions throughout the system or corridor.</li> <li>2. Some algorithms, such as the fuzzy logic algorithm, have the ability to address multiple objectives (e.g., freeway congestion and ramp queues).</li> </ol>	<ol style="list-style-type: none"> <li>1. Requires mainline detection (both downstream and upstream detectors).</li> <li>2. Requires communication to central computer.</li> <li>3. Requires technical expertise for calibrating and implementing system.</li> <li>4. More expensive than local traffic responsive in implementation resources needed and communications maintenance.</li> </ol>

Source: *Ramp Management and Control Handbook*, Table 5-4 (U.S. DOT 2006b)

#### A.4 Queue Management

Queue management is an important component of any ramp metering installation. As described in the *Ramp Management and Control Handbook* (U.S. DOT 2006b), the general approach to queue management is to either provide sufficient storage for worst-case queues, or install queue detectors to determine when queue lengths are becoming problematic, and adjust the metering rate accordingly. Where sufficient storage space is unavailable, queue detection and mitigation are essential to prevent queued vehicles from spilling back onto the arterial network.

In addition to preventing impacts to adjacent roads, queue length constraints may also serve as a surrogate means of ensuring that ramp delays do not become excessive. Whether incorporated as part of the metering algorithm or implemented using a more crude over-ride technique, such queue length constraints can have a significant impact on the operation of the ramp metering system, and its ability to control congestion on the freeway mainline. For this reason, it is often beneficial to coordinate queue management between adjacent ramps; if the metering rate at one ramp must be reduced to prevent spillback, it may be possible to adjust the rate at nearby ramps so that the overall number of vehicles accessing the freeway is unaffected.

# **APPENDIX B**

## **RAMP METERING MECHANISMS**

## **B. RAMP METERING MECHANISMS**

There are four primary mechanisms by which ramp metering reduces freeway congestion and delay (Banks 2000):

- Upstream exit mechanism
- Bottleneck flow mechanism
- Diversion mechanism
- Improved merging operations and safety

An understanding of these mechanisms is crucial to developing an effective ramp metering strategy. A brief description of each mechanism is provided in the following sections.

### **B.1 Upstream Exit Mechanism**

In the upstream exit mechanism, delay is reduced by expediting traffic flow to exits upstream of the freeway bottleneck.

When the freeway is congested, vehicle queues often extend over a significant distance, blocking access to off-ramps. As a result, vehicles wishing to exit the freeway are prevented from doing so, and are forced to join the mainline queue instead. This in turn causes the queue to extend over an even greater distance, exacerbating the situation. By introducing ramp metering, mainline congestion is reduced, and vehicles exiting the freeway are no longer delayed by off-ramp blockages. Offsetting this reduction in delay is an increase in travel time imposed at the ramp meters. Overall, a net reduction in delay can be achieved by introducing a ramp metering scheme which ensures that as much of the ramp delay as possible is borne by the bottleneck flow (Banks 2000).

In understanding how the upstream exit mechanism works, it is useful to consider the mainline queue which exists in the absence of ramp metering. This queue can be divided into two components: those vehicles destined through the bottleneck, and those vehicles destined to exits upstream of the bottleneck. Under a ramp metering scheme which exploits the upstream exit mechanism, the queue of vehicles destined through the freeway bottleneck is essentially moved to the metered ramps, preventing freeway flow from breaking down. Since delay is merely transferred from one location to another, the net change is zero. The benefit is derived from the remaining vehicles who are bound for exits upstream of the bottleneck. These vehicles were previously delayed in the mainline queue, but with ramp metering, are now able to exit the freeway without delay.

Clearly, if some of the vehicles destined to exits upstream of the bottleneck are subject to ramp delays, the overall benefit of the ramp metering strategy will be reduced. Thus, it is important to limit metering to those ramps where a significant portion of the ramp flow is bound for the bottleneck. This implies that metering will most often be restricted to those ramps immediately upstream of the bottleneck; the further a ramp is from the bottleneck, the more likely that drivers exit the freeway before reaching the critical section.

Given the above discussion, metering priority should be based on the fraction of ramp flow destined to the bottleneck (Banks 2005). For similar reasons, it is important to limit the number of ramps included in the metering scheme. For each additional ramp subject to control, the greater the potential for unnecessarily delaying drivers who exit the freeway upstream of the bottleneck, offsetting any benefits from reduced mainline congestion. The efficiency of the metering strategy therefore depends on the storage capacity of the metered ramps and the queue length policy of the transportation agency. If ramp queue lengths are constrained, more meters must be added to the metering scheme to achieve the desired reduction in mainline congestion. This in turn impacts how the metering system functions, since the overall level of delay reduction is sensitive to the number of metered ramps. As Banks (2000) found, a ramp metering strategy which minimizes freeway congestion does not necessarily minimize delay if metering must be extended beyond the first few ramps upstream of the bottleneck in order satisfy constraints related to the minimum metering rate or maximum queue.

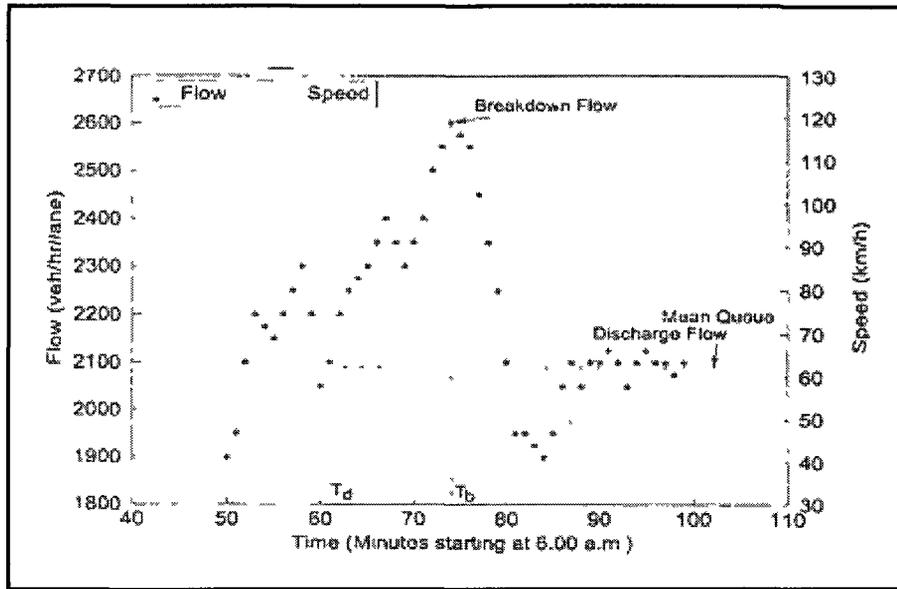
Since ramp metering strategies based on the upstream exit mechanism tend to meter only those ramps closest to the bottleneck, such strategies are often criticized as being inequitable.

## **B.2 Bottleneck Flow Mechanism**

There is evidence that the maximum flow rate through a freeway bottleneck depends on the type of traffic flow (Banks 1991, Zhang and Levinson 2004b, Bertini and Myton 2005, Hall and Agyemang-Duah 1991). Under uncongested conditions, freeway capacity tends to be higher; once flow breaks down, throughput is reduced (refer to Figure B-1). Thus, a ramp metering system which prevents freeway demand from exceeding capacity may actually increase traffic flow through the bottleneck.

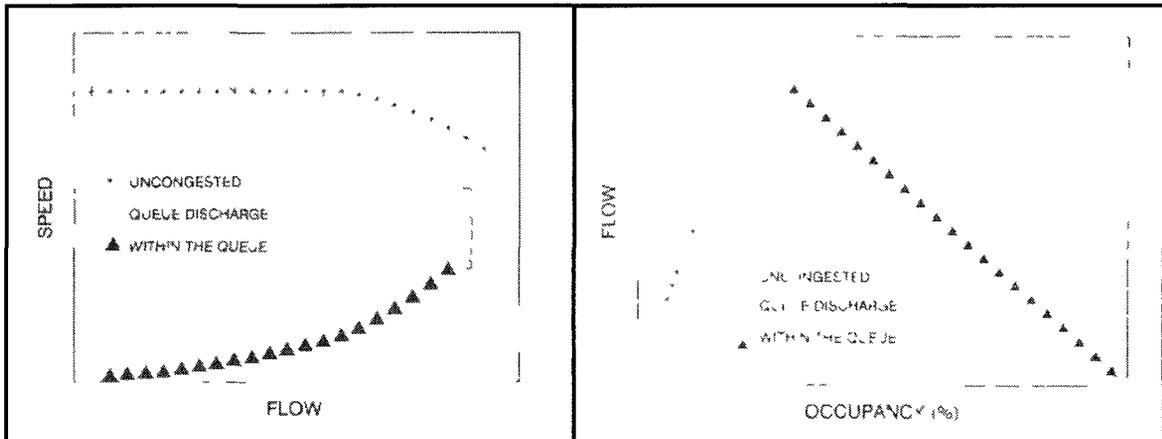
The U.S. Department of Transportation's 1995 Ramp Metering Status Update cites evidence from a number of studies that showed an increase in peak period freeway volumes after the introduction of ramp metering. The report claims that these findings were not random occurrences, and can be attributed "to flow rates higher than those that occur under ... 'breakdown' conditions" (pg. 21).

When the freeway is operating in an uncongested state, the relationship between traffic flow and density is straightforward; as flow increases, density increases in a roughly linear relationship (refer to Figure B-2). At the same time, the average speed declines, albeit by a relatively small amount. As demand levels approach capacity, flow becomes less stable, increasing the probability of breakdown. Given the stochastic nature of traffic flow, the transition from a stable state to a congested state is not fixed at one specific flow level, but instead can occur over a range of values, up to a maximum point. With the onset of congestion, the capacity of the bottleneck becomes equal to the queue discharge flow, which, on average, has been found to be lower than the maximum bottleneck flow rate under uncongested conditions. The drop in bottleneck throughput that occurs with the onset of congestion provides the motivation for ramp metering algorithms that prevent freeway flow from breaking down.



Source: U.S. DOT 2006b, Section 5.3.1

**Figure B-1 Bottleneck Throughput Under Congested & Uncongested Flow**



Source: Hall et al. 1992, pg. 13-14

**Figure B-2 Generalized Speed-Flow-Occupancy Relationships**

In order to exploit the bottleneck flow mechanism, the ramp metering algorithm must maintain flows below the uncongested capacity, but above the congested capacity – a feat which, according to Banks (1991, 2000), requires great precision, and which may be difficult to achieve in practice. The success of such a strategy depends on the ability of the algorithm to accurately predict the flow which triggers breakdown. However, as noted above, flow along a given freeway section does not consistently break down at a set value; breakdown may occur at any point within a certain range, complicating efforts to maintain flows as high as possible without causing congestion. To overcome this challenge, Persaud et al. (2001) developed a “probability-of-breakdown” function which gives the probability of flow breakdown as a function of the traffic volume immediately

downstream of the ramp merge. This function, based on data from the QEW freeway in Mississauga, was successfully tested using Monte Carlo simulation techniques to assess its suitability for use in ramp metering applications.

In another example, Kosmatopoulos et al. (2006) examined techniques for real-time estimation of critical occupancy for inclusion in the ALINEA ramp metering algorithm. Under the proposed approach, the critical occupancy is updated in real-time to reflect changes in weather, traffic composition, or freeway control, enhancing the ability of the algorithm to maximize freeway flow. According to the authors, critical occupancy has been found to be less sensitive than capacity, which makes it ideally suited to ramp metering algorithms which attempt to prevent flow breakdown. However, the use of critical occupancy in coordinated predictive-type algorithms (i.e. algorithms that predict how traffic will respond to different control scenarios) is limited due to the difficulty in accurately forecasting the occupancy at each point in space where breakdown may occur.

Many ramp metering strategies are based on the premise of keeping freeway flow less than capacity to prevent the formation of mainline congestion. Obviously, the intent is to keep flow as high as possible without triggering breakdown. It is unclear whether such strategies fully exploit the bottleneck flow mechanism, or whether the maximum flow rate is set conservatively low, so that any ramp metering benefits are derived primarily from the upstream exit and diversion mechanisms.

In investigating possible ways to exploit the two-capacity phenomenon at freeway bottlenecks, Banks (1991) concludes that the optimal strategy is to delay metering as long as possible and then meter at a fairly high rate. Otherwise, if metering is initiated too early, or if the metering rate is set too low, ramp metering may actually be counter-productive. The challenge lies in accurately predicting when breakdown is likely to occur, and maintaining arrival flows within a very narrow range between the two capacity values.

The bottleneck flow mechanism provides an opportunity to improve freeway operations. Through the use of ramp metering, freeway congestion can be prevented, allowing traffic flow to be maintained at higher levels. However, ramp metering in itself may have an impact on mainline capacity. Banks (1991) hypothesizes that metering increases bottleneck capacity by eliminating merge conflicts as a cause of flow breakdown. Thus, even if a metering system is unable to fully exploit the bottleneck flow mechanism, it may still result in an increase in traffic throughput by breaking up platoons of on-ramp vehicles which reduce the capacity of the merge area.

### **B.3 Diversion Mechanism**

While many research articles allude to the ability of ramp metering to induce traffic diversion, few attempt to quantify these impacts in any meaningful way. As noted by Banks as recently as 2005, “traffic diversion has long been recognized as a promising possibility but has been the subject of very little theoretical discussion” (pg. 12). In an attempt to lay the groundwork for such a discussion, Banks sets out some “elementary

theory” on ramp metering and traffic diversion. An overview of the key concepts is provided below.

In general, ramp meters are associated with two forms of traffic diversion:

1. Diversion of traffic from the freeway to the arterial network to avoid ramp delays
2. Diversion of traffic from the arterial network to the freeway to take advantage of reduced mainline congestion

Through the use of ramp metering, a reduction in delay can be achieved by encouraging traffic to divert around the freeway bottleneck. The resulting improvement in mainline operations can lead to a further reduction in delay, by encouraging vehicles destined to areas upstream of the bottleneck to use the freeway, which without congestion, often represents a much faster travel route.

As noted by Banks (2005), the potential impacts of ramp metering on route choice are “quite complicated” (pg. 12). This is particularly true in complex networks with extensive metering systems, numerous alternative routes, and diverse origin-destination patterns.

Ramp metering encourages diversion by altering the travel time on different components of the network, impacting the relative attractiveness of the various routes between each origin-destination pair. Ramp delays increase, freeway delays decline. The travel time on surface streets may also be impacted due to traffic diversion or spillback from freeway on-ramps. The overall impact experienced by a particular driver depends on the portion of the total trip length spent on each facility. A driver using the freeway for only a short distance may find that the ramp delay exceeds the travel time benefit from reduced mainline congestion, and that diversion to an alternate route is preferred. For longer distance trips, the ramp delay is only a small portion of the overall trip duration, and it is far less likely that a parallel arterial route will offer a shorter alternative.

For diversion of traffic around a freeway bottleneck, ramp metering impacts route choice by creating ramp delays. When the delay is equal to the difference in travel time between the freeway route and the alternative route, traffic begins to divert. The queue length which triggers this diversion is known as the “equilibrium queue length”.

When relying on traffic diversion to relieve mainline congestion, a few issues are worth noting, as summarized by Banks (2005):

- Strategies promoting diversion are incompatible with constraints on ramp queue lengths; if the maximum queue length is less than the equilibrium queue length, no diversion will occur; if it is greater, the maximum queue will never be attained and the constraint will be irrelevant
- If the equilibrium queue is so long that it blocks access to the decision point, no diversion will take place (by the time drivers reach the decision point, the freeway path will always be shorter)

- Metering to encourage traffic diversion will generally be limited to a relatively few number of ramps in the vicinity of the bottleneck (the further from the bottleneck, the greater the difference in travel time between the freeway / arterial routes, and the greater the ramp delay / equilibrium queue length needed to trigger diversion)
- In analyzing diversion, it is important to consider what travel time information is likely to be available to motorists and how this information may influence route choice

That traffic diversion reduces overall system delay is not immediately apparent; drivers choosing to divert to an alternative route are actually worse off from a travel time perspective, otherwise, they would have already diverted to the new route without needing an 'incentive' from ramp metering. However, although these drivers may suffer, overall, delay is reduced since the cumulative output of the system has shifted earlier in time. Intuitively, if a driver diverts rather than wait in a ramp queue, he no longer delays drivers arriving behind him, allowing more drivers to get out of the system sooner (for a more elegant explanation, refer to Banks 2005).

In general, the equilibrium queue length will vary depending on the destination of the drivers using the ramp, the level of demand, the ramp metering rate, the level of freeway congestion, and the characteristics of the alternative route, in particular, whether the route is sensitive to traffic flow, and if so, whether it is under- or over-saturated. Although it may be assumed for simplicity that all drivers merely divert around the bottleneck and enter the freeway at the next interchange, in reality, diversion patterns may be much more complex.

If travel times on the alternate route are fixed, and ramp metering rates are constant, a stable equilibrium queue will eventually form on the metered ramps. Once this queue has been established, drivers who would normally use the ramp in the absence of metering divide themselves between the ramp and the diversion route such that the equilibrium queue length is maintained (i.e. the flow entering the ramp equals the ramp metering rate).

In cases where the alternative route is sensitive to traffic flow, the equilibrium queue length will vary throughout the peak period, however, it will do so in a relatively stable manner if it is assumed that travel time equilibria are approximate, and that route choice decisions are based on reasonable expectations of travel time from drivers' past experiences. For diversion routes that are sensitive to increases in traffic flow, total system delay is impacted in two ways: 1) the presence of diverted traffic increases delay to other users of the route, and 2) increased travel time on the alternative route increases the level of ramp delay needed to induce diversion, reducing the effectiveness of the diversion mechanism as system outputs are shifted later in time.

Recognizing the potential benefits of the diversion mechanism, Banks (2005) developed a metering strategy which maintains flow through the bottleneck at its maximum level while diverting as much traffic as possible to an alternative route. According to Banks,

the optimal strategy is to meter as few ramps as possible, using the most restrictive metering rates allowed. The use of such metering rates will tend to reduce the length of the equilibrium queue, and the time it takes to form. Ramps are added to the metering scheme based on the time required for diverted vehicles to reach the system exit. Based on this scheme, ramps closest to the bottleneck will tend to be metered first, similar to ramp metering strategies based on the upstream exit mechanism.

From his review, Banks (2005) provides the following assessment of the role of traffic diversion in improving network performance:

*Diversion of traffic around freeway bottlenecks is in many ways the most promising of the system-delay-reducing ramp metering mechanisms because a) there may be significant amounts of unused capacity on routes parallel to freeway bottlenecks, b) the amount of traffic that can be diverted is apt to be large compared with the flow to exits immediately upstream of the bottleneck, c) the increase in output during periods of congestion is much less dependent on precise control of flow approaching the freeway bottleneck than in the case of schemes attempting to increase flow through the bottleneck, and d) the prospect of diversion provides a 'safety valve', permitting highly unbalanced metering schemes to function without objectionable ramp queue sizes and delays. Finally, ramp metering schemes based on diversion around freeway bottlenecks are robust in the sense that delay will be significantly reduced by any scheme (optimal or not) in which a reasonably large amount of traffic can be diverted during periods when the system is congested (pg. 19).*

#### **B.4 Improved Merging Operations & Safety**

There is a general recognition in the literature that one of the benefits of ramp metering is its ability to break up vehicle platoons at entrance ramps, smoothing traffic operations in the vicinity of the merge and improving safety. For example, Chaudhary and Messer (2002) claim that one of the three main objectives of ramp metering is to break up platoons of vehicles released from an upstream traffic signal, in order to provide a safe merge operation at the freeway entrance. Despite this recognition, however, much of the literature focuses on the congestion relief provided by ramp meters; any benefits from reduced turbulence in the merge area are considered merely a positive by-product of the metering process.

In general, the safety benefits of ramp metering are well-documented. Several studies have shown a reduction in crash rates following the introduction of ramp controls, particularly for rear-end and side-swipe collisions (see U.S. DOT 1995). However, while results from empirical data have been used to confirm that safety benefits from ramp metering do in fact exist, little attempt has been made to understand the underlying crash reduction mechanism.

Since collisions are a major cause of non-recurrent congestion, it follows that any reduction in freeway crashes due to ramp metering will also produce a corresponding reduction in delay. However, it is difficult to estimate the potential magnitude of this benefit due to uncertainty surrounding the impact of ramp metering on collision frequency, severity, and duration. As a result, safety benefits are often overlooked since too little is known about the crash reduction mechanism for quantitative analysis of potential benefits and costs (Banks 2000); very few studies have explicitly considered the safety benefits of ramp metering in a quantitative way (Lee et al. 2006).

Other than ensuring that good design practices are followed in the installation of ramp metering systems, little emphasis has been given to the relationship between ramp metering and safety. The general opinion seems to be that most safety benefits are derived from the decision to install ramp metering, and not from the specifics of the implementation itself.

However, collision statistics indicate that certain ramp metering projects outperform others from a safety perspective. As a result, it is important to understand the key factors that influence collision occurrence in the vicinity of the merge, and how ramp metering influences driver behaviour, in order to maximize potential safety benefits through improved ramp metering design and operation.

As a step towards this objective, a number of researchers have investigated driver behaviour at freeway on-ramps. For example, Sarvi et al. (2004) used simulation techniques combined with a driving simulator to study freeway ramp merging under congested conditions. Cassidy and Ahn (2005) observed that queued vehicles from an on-ramp and queued mainline traffic enter a congested merge in a more or less fixed ratio that is independent of the merge outflow. While such studies provide insight into driver behaviour at freeway merges in general, few have specifically examined the behavioural effects of ramp metering. To improve knowledge in this area, Wu et al. (in press) used instrumented vehicles and roadside video cameras to examine how ramp metering impacts driver behaviour, such as speed, headway, acceleration/deceleration, gap acceptance, and merge distance. The study was conducted for a section of the M27 in South England. It was found that ramp metering was associated with a significant

### How it Works

Freeway congestion that forms at or immediately upstream of merge areas is often a result of large platoons of vehicles entering the freeway from a ramp. These vehicles must compete for gaps in mainline traffic, limiting motorists' ability to focus on traffic in front of them. Adding to this problem are geometric deficiencies that make weaving operations at the ramp-freeway merge point more complex. Such deficiencies include horizontal and vertical curves, closely spaced ramps, and inadequate acceleration or deceleration distances. As a result, rear-end, sideswipe, and lane change collisions may occur on the freeway or ramp...

To a large extent, collisions attributed to merging problems can be reduced by breaking up platoons so vehicles are not forced to compete for the same gaps in mainline traffic. If repeated on a system-wide basis, the overall operation of the freeway may be stabilized, and crashes that result from stop-and-go driving behavior may be reduced.

*Source: Ramp Management & Control Handbook, Section 1.1.5, U.S. DOT (2006b)*

increase in the number of freeway lane changes from the shoulder lane to the adjacent lane in advance of the merge junction, resulting in greater gaps in the shoulder lane for use by merging traffic. This in turn was accompanied by a reduction in both the merge speed and merge distance exhibited by vehicles entering the freeway.

The authors hypothesize that this behaviour can be attributed to the way in which ramp meters impact vehicle acceleration on the ramp approaching the merge junction. With ramp metering on, vehicles entering the freeway must accelerate from a stopped position. In the case of the ramp under investigation, the distance from the ramp meter to the beginning of the merge is insufficient for drivers to accelerate to the same speed as observed without ramp metering, causing vehicles to arrive at the merge with “significantly” lower speeds. Upon observing the slower moving vehicles on the on-ramp, some vehicles in the shoulder lane opt to move into the adjacent lane to avoid deceleration and delay. As a result, gap sizes in the shoulder lane increase, facilitating the merge operation, and allowing on-ramp vehicles to enter the freeway at lower speeds, over a shorter merge distance. While the study by Wu et al. represents an important step in assessing the impact of ramp metering on driver behaviour, it is unclear if similar results would be obtained at ramps with longer acceleration distance between the ramp meter and point of merge.

While it is important to understand how ramp metering influences driver behavior, it is equally important to understand how such changes in behavior may influence safety. Although collision statistics provide valuable data for assessing safety, it is often difficult to obtain a sufficient sample size to ensure statistically significant results. To overcome the limitations of collision data, researchers have identified surrogate measures of crash risk which can be used to study the safety benefits of operational improvements within a simulation environment. In one of the few examples involving ramp metering, Lee et al. (2006) used a real-time crash prediction model to investigate the safety effects of a local traffic-responsive ramp metering strategy. While other studies on ramp metering and safety have been conducted (as cited in Lee et al. 2006), the extent of research in this area has been limited.

Although more research is needed to fully define the relationship between ramp metering and safety, transportation agencies have proceeded to develop ramp metering strategies which specifically exploit the crash reduction mechanism to prevent collisions and minimize non-recurrent congestion and delay.

While most ramp metering applications are concerned primarily with restricting on-ramp flow to improve mainline operations, in some cases, ramp metering is introduced solely to improve merge operations in the vicinity of the ramp without regard for mainline capacity or demand. In this situation, the ramp metering rate is set equal to the ramp demand; all vehicles approaching the freeway are allowed to enter with minimal delay. Known as “non-restrictive” metering or “metering at demand”, this type of metering is most often introduced to reduce mainline collisions by breaking up platoons of vehicles entering the freeway. (U.S. DOT 2006b). Metering at demand may also delay the onset of freeway congestion by increasing the capacity of the merge.

Gap acceptance (or merge) control strategies are also intended to smooth traffic flow. In this type of control, metering is based on real-time occupancy measurements taken upstream of the ramp. When a suitable gap is detected in the shoulder lane, a green signal is triggered at the ramp meter, allowing the waiting vehicle to accelerate and safely merge into freeway traffic. The intent is to facilitate the merging operation, allowing high numbers of vehicles to safely enter the freeway with minimal disruption to mainline flow.

The challenge with such an approach lies in predicting when the gap will arrive at the merge location, and how long it will take the driver on the ramp to accelerate to freeway speeds. To implement gap acceptance control, it is generally assumed that all drivers have a similar level of aggressiveness in terms of gap acceptance and acceleration behavior. It is further assumed that once a gap is detected, no lane changes occur on the mainline which might cause the gap to disappear. In reality, these assumptions are not always valid, and as a result, gap acceptance control “has been plagued with difficulties” due to the instability of measured gaps, and the non-uniformity of vehicle acceleration behavior (CCIT 2001). Moreover, it has been found that gap acceptance control may result in more restrictive metering rates than necessary, and may have higher than normal violations.

# **APPENDIX C**

## **INTRODUCTION TO BAYESIAN NETWORKS**

## C. INTRODUCTION TO BAYESIAN NETWORKS

### C.1 Overview

This appendix provides an introduction to dynamic Bayesian decision networks. Much of the discussion which follows is based on the work by Korb and Nicholson (2004) and Jensen (2001). Both references provide an excellent overview of the topic, with many practical examples. Murphy's (1998) on-line introduction to graphical models and Bayesian networks is also a good starting point while Yudkowsky's (no date) website on Bayesian reasoning provides an intuitive explanation of Bayes' law and its application.

For a more in-depth treatment of the topic, the reader is referred to Judea Pearl's (1988) seminal work, *Probabilistic Reasoning in Intelligent Systems*. The more recent work by Cowell et al. (1999), *Probabilistic Networks and Expert Systems*, is also an excellent technical reference, particularly with regards to exact inference in Bayesian networks.

### C.2 Bayesian Networks as a Framework for Modelling Uncertainty

A Bayesian network is a graphical model which captures the uncertainty inherent in real-life systems. Also known as probabilistic graphical models or belief networks, Bayesian networks can include both discrete and continuous variables. They can model systems of any size and complexity, incorporating both static and dynamic processes involving linear and non-linear relationships.

Within a Bayesian network, random variables are represented by nodes, while direct dependencies between variables are represented by arcs. The arcs in a Bayesian network are directional – they run from one node to another. In the language of Bayesian networks, the former is known as the 'parent' node, while the latter is called the 'child' node.

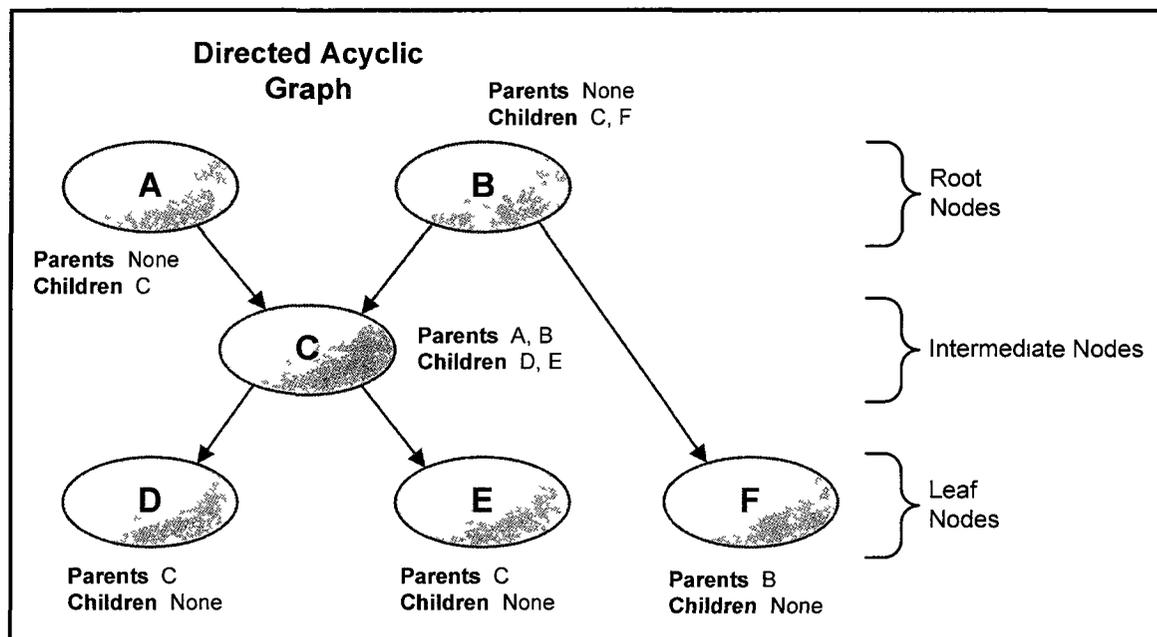
In a directed chain of nodes, a node is an ancestor of another node if it comes earlier in the chain, and a descendent if it comes later in the chain. A node without any parents is referred to as a 'root' node, while a node with no children is called a 'leaf' node. All

"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering -- uncertainty and complexity -- and in particular they are playing an increasingly important role in the design and analysis of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity -- a complex system is built by combining simpler parts.

Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms."

Michael Jordan (1998) as cited in Murphy (1998)

other nodes are considered intermediate nodes. A directed cycle occurs when a sequence of directed arcs leads back to a common node. Since directed cycles are prohibited in Bayesian networks, Bayesian networks are also known as directed acyclic graphs or DAGs.



**Figure C-1 A Simple Bayesian Network**

In general, the topology of the network should capture the relationships between the variables that comprise the system or process of interest. Of particular importance are the query nodes representing the status of certain key aspects of the system which cannot be observed, and the observation nodes, for which evidence may be available. Variables may be used to describe a certain attribute or feature of the system, an event, an action, or a hypothesis. A variable is considered relevant if it is observable and can be used to refine beliefs about other unknown variables, if its value must be known in order to take some action or report some result, or if it is an intermediate or internal variable that helps to express other relationships in the model (Norsys 2007). Both discrete and continuous variables are allowed. For each variable included in the network, it is important to define a set of values that the variable can assume. Since a variable can take on exactly one of these possible values, the values must be mutually exclusive and exhaustive; while the actual value of the variable may be uncertain, it must fall within this pre-defined set.

In developing the network, two nodes are connected if one causes or affects the other, with the direction of the effect indicated by the arc. More specifically, an arc from node **X** to node **Y** indicates that **X** causes **Y**, that **X** partially causes or predisposes **Y**, that **Y** is an imperfect observation of **X**, that **X** and **Y** are functionally related, or that **X** and **Y** are statistically correlated (Norsys 2007). While Bayesian networks often have a causal structure, with arcs representing causal dependencies, this is not strictly necessary as long as all probabilistic dependencies are captured. Indeed, it is common practice to transform Bayesian networks to carry out probabilistic reasoning. While researchers may disagree

on the role of causality in Bayesian networks (Korb and Nicholson 2004), from a practical implementation perspective, it is often useful to think in terms of causality. By structuring networks to reflect cause-effect relationships, network development and interpretation is straightforward and intuitive. Jensen (2001) claims that causal networks are simpler with fewer links and more stable conditional probabilities, but concedes that identifying the correct causal relationship between two correlated variables is not always easy.

Once the structure of a Bayesian network has been established, the relationships between connected nodes are quantified through the use of conditional probability distributions: given a particular instantiation of the parent nodes, what is the probability of the child node taking on each of its possible values? The probabilities incorporated in the Bayesian network can be derived from a number of sources, ranging from empirical studies to subjective estimates elicited from experts.

The process of constructing a Bayesian network requires researchers to study all of the possible variables and inter-connections between variables that are relevant to the problem at hand, allowing them to develop a greater understanding of the underlying processes and relationships, and perhaps causing them to view the problem in new and useful ways. The graphical model provides a focal point for discussion, facilitating cross-disciplinary collaboration, and serves as a framework for visualizing and communicating complex relationships. However, Bayesian networks are much more than just a convenient means for developing a graphical representation of a complex system – they also provide an elegant mathematical construct for modelling such systems. The structure of a Bayesian network encodes information on the relationships between variables – information that can be used for inference and decision-making under uncertainty. The presence of an arc between two nodes (or groups of nodes) implies a direct dependency; the absence of an arc is no less meaningful, expressing a conditional (in)dependence contingent on the network configuration and available information (refer to Figure C-2).

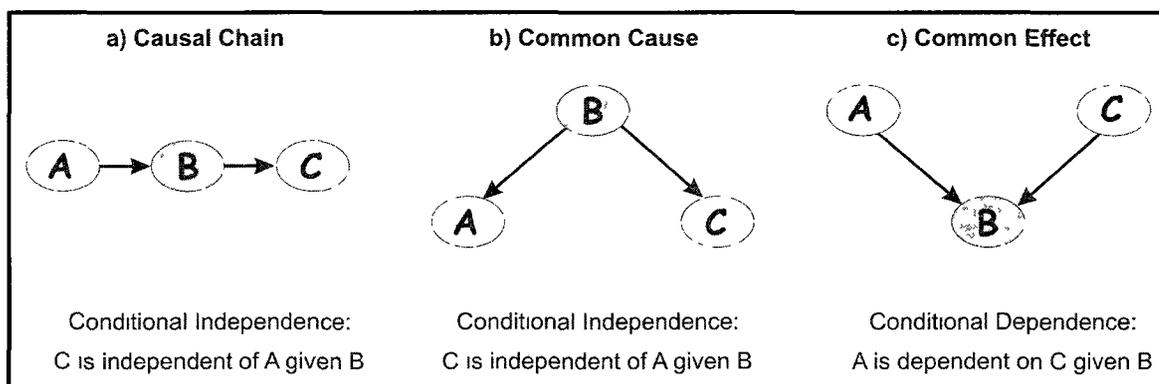
In the ‘causal chain’ shown in part a) of Figure C-2, Node **C** is independent of Node **A** given information about Node **B**. Such independence implies that:

$$P(C | A \wedge B) = P(C | B)$$

In other words, knowledge of Node **A** has no impact on our beliefs about Node **C** if evidence is available for Node **B**. Obviously, if no evidence is available for Node **B**, information on Node **A** will influence our beliefs about Node **C**, underlying a dependency conditional on Node **B**.

Likewise, two nodes having a common cause are conditionally independent given evidence about the common cause. If there is no information about the common cause, the two nodes are dependent. In Figure C-2, part b), information about Node **A** will influence our beliefs about Nodes **B** and **C** assuming no evidence is available for Node **B**. However, if evidence is available for Node **B**, Nodes **C** and **A** are independent – knowledge of one has no impact on the other, given the evidence for Node **B**.

In contrast, nodes which produce a common effect give rise to an opposite form of conditional independence. In Figure C-2, part c), Node **A** is dependent on Node **C** given information about the common effect **B**. However, if there is no information about Node **B**, Nodes **A** and **C** are independent. In essence, it is the lack of knowledge about Node **B** which produces independence, whereas in the other two cases (causal chains and common effects), it is the presence of knowledge about Node **B** that blocks the relation between Nodes **C** and **A**, such that information about Node **A** is no longer relevant to Node **C**, and vice versa.



**Figure C-2 Conditional Dependence & Independence Relationships**

Based on the above discussion, it follows that a node is independent of its ancestors, given its parents. This simple yet powerful characteristic of Bayesian networks allows us to completely specify a joint probability distribution using only the conditional probabilities specified in the Bayesian network. Indeed, by exploiting the conditional independencies implied by the network structure, a more compact representation of the joint probability distribution is possible.

The joint probability that a set of random variables will take on certain values can be expressed as  $P(\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2, \dots, \mathbf{X}_n = \mathbf{x}_n)$ , or more compactly, as  $P(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ . From the rules of probability theory, it follows that:

$$\begin{aligned}
 P(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= P(\mathbf{x}_1) \times P(\mathbf{x}_2 | \mathbf{x}_1) \dots \times P(\mathbf{x}_n | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-1}) \\
 &= \prod_i P(\mathbf{x}_i | \mathbf{x}_1, \dots, \mathbf{x}_{i-1})
 \end{aligned}$$

In a Bayesian network, the value of a particular node is conditional only on the values of its parents' nodes. As a result, the joint probability can be expressed more compactly as:

$$P(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_i P(\mathbf{x}_i | \text{Parents}(\mathbf{X}_i))$$

Thus, the conditional independence assumptions encoded in the network structure allow a more compact representation of the joint probability distribution over the random variables.

As the above discussion implies, Bayesian networks offer several advantages. They have been used in applications ranging from medical diagnosis, weather forecasting, and modelling of biological processes to traffic monitoring, design of automated vehicles, and ecological applications. NASA uses Bayesian networks to provide advice on possible failures of the space shuttle's propulsion system, while Microsoft has experimented with Bayesian networks in interactive software systems (see Korb and Nicholson 2004 for references to these and other applications).

The strength of Bayesian networks lies in their ability to simplify probabilistic reasoning, determine appropriate decisions under uncertainty, and explain the outcome of stochastic processes (Korb and Nicholson 2004). The graphical representation of complex inter-relationships provides a basis for communication and collaboration, while the structure of Bayesian networks, and their ability to take advantage of independencies between variables, has fostered the development of efficient computational algorithms for probabilistic inference and learning causal models – two of the most common uses of Bayesian networks today.

An introduction to probabilistic inference in Bayesian networks is provided in Section C.4, while an application involving learning is described in Chapter 7 of the main document. A simple example of a Bayesian network can be found in Appendix D.

### C.3 Dynamic Bayesian Networks

Bayesian networks provide a graphical representation of the relationships between variables which comprise a random system or domain. If some of the relationships in the model are temporal, the network is referred to as a Dynamic Bayesian Network (DBN)<sup>1</sup>. Such networks allow us to model stochastic processes as the system changes over time.

Within a DBN, the state of the system at each time step is represented by a graphical model comprised of “intra-slice” arcs which capture the relationships between variables at a particular point in time. The structure of this model is generally considered to be constant, and does not change from one time step to the next. Successive time steps are modelled using “inter-slice” or “temporal” arcs. These arcs define relationships between the same and different variables over time. For example, the position and speed of an object at one time step will influence its speed and position in the next time step.

Together, the following elements fully define a DBN:

- Intra-slice topology (i.e. the topology within a time slice)
- Inter-slice topology (i.e. the topology between time slices)
- Conditional probability tables for the first time slice (will have no parents from a previous time slice)

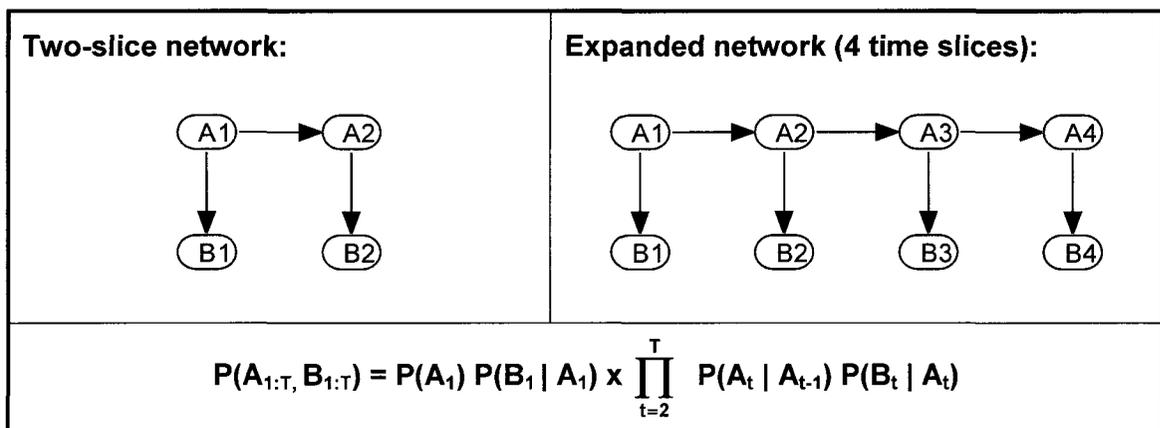
---

<sup>1</sup> Some authors (i.e. Murphy 1998) suggest that Dynamic Bayesian Networks should more appropriately be called Temporal Bayesian Networks, since it is assumed that the model structure does not change over time.

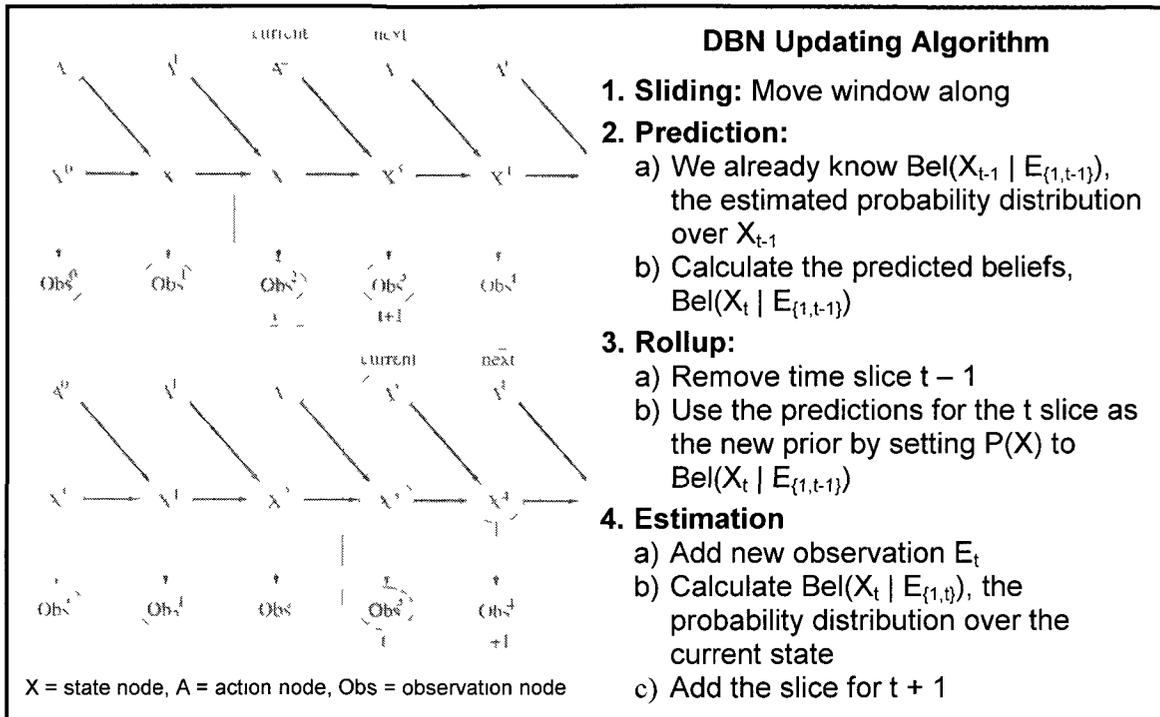
- Conditional probability tables for subsequent time slices (may have parents from the current or previous time slices)

Many DBNs adhere to the Markov assumption, and are structured so that no arcs span more than one time step. Implicit in this representation is the assumption that the state of the system depends only on its state in the previous time step and any action taken upon it. Thus, a DBN can be fully specified by describing the initial state distribution and using a two-slice Bayes' net to define the network structure and conditional probability distributions. This compact representation is possible since the network structure is assumed to remain constant over time. It is also generally assumed that the model is time-invariant, and that the network parameters do not change from one time-step to the next. However, periodic non-stationarities can be captured by adding hidden nodes to the model to represent the current "regime" (Murphy 1998).

To carry out reasoning, the DBN must be unrolled (refer to Figure C-3). In the general case, evidence from all previous time steps up to and including the current one is used to update beliefs for the full DBN, including nodes which fall in future time steps. While standard inference algorithms can be used to compute the new probabilities, in most cases, the network becomes too large. In this situation, analysis of the DBN is often restricted to a sliding "window" of time slices. As time moves forward, older time slices are dropped, and new ones are added. Once a time slice is dropped, any evidence from that time slice is no longer directly available. Instead, the evidence is summarized in the updated probability distributions for the nodes which are carried forward in the sliding window. In essence, evidence is used to update the beliefs of the nodes in the active window. The posterior probability distributions for these nodes are then taken as the new prior distributions, the oldest time slice is dropped, a new time slice is added (with new evidence) and the process is repeated. A simple algorithm for updating a DBN is provided in Figure C-4 for the common case of a two time-slice sliding window.



**Figure C-3 Specification and Expansion of a Simple DBN**



Source: Korb and Nicholson, 2004, pg. 108-109

**Figure C-4 DBN Update Process – Sliding Window with 2 Time-Slices**

Dynamic Bayesian Networks can be considered a generalization of other well-known dynamic models, such as Hidden Markov Models and State Space Models. Rather than combine variables into a single vector, DBNs represent the hidden state as a set of random variables. Observations are treated in a similar manner, facilitating the creation of graphical models which capture conditional independencies between variables which would otherwise be obscured (Murphy 2002).

According to Murphy (1998), the simplest kind of DBN is a Hidden Markov Model (HMM), which has one discrete hidden node and one discrete or continuous observed node per time slice. A simple HMM is illustrated in Figure C-5 below, along with several variants. A State Space Model (or Linear Dynamical System), has the same structure as an HMM, but all the nodes are assumed to be continuous, with linear-Gaussian distributions. The hidden state is represented by a single vector-valued random variable. In a Switching State Space Model, the system can jump between different operating modes or regimes, and the model thus incorporates both discrete and continuous hidden variables. When working with Hidden Markov Models and State Space Models, it is common practice to define a transition model and observation model, in addition to the initial state distribution.

In contrast to the more specialized models described above, the hidden state in a DBN is represented by a set of random variables, which can be either discrete or continuous, with no restriction on the distribution form. Observations are treated in a similar manner, with the network structure represented by a two-slice Bayes' net. In this more general case, the

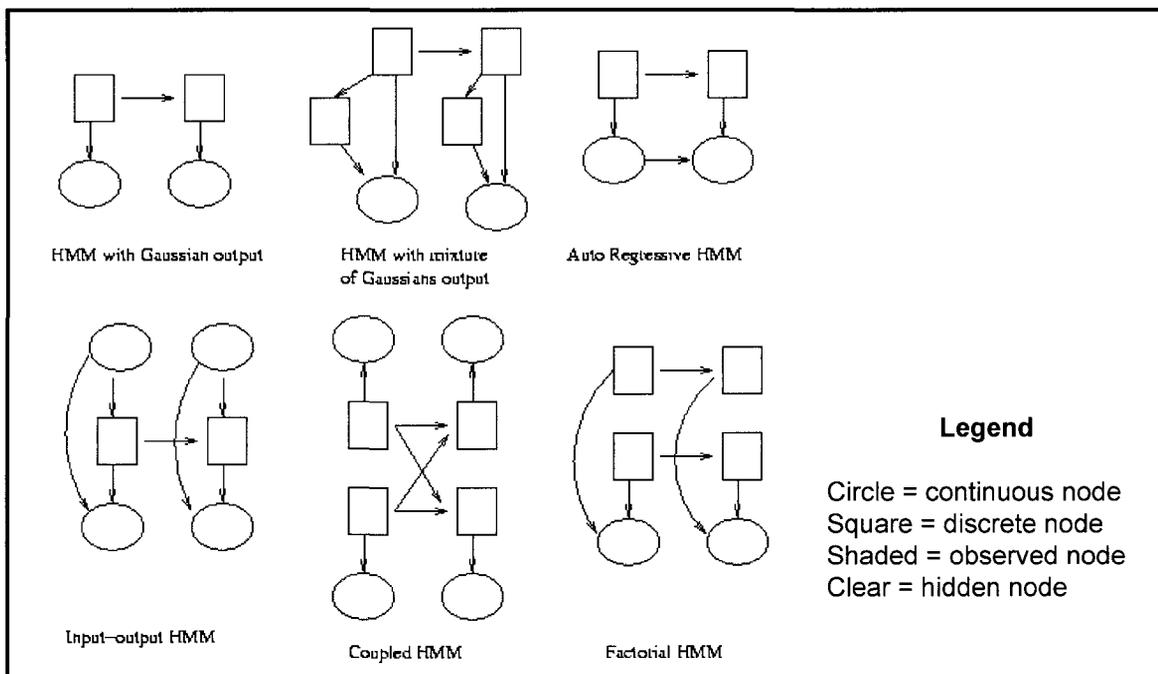
observation and transition models can be defined as a product of the conditional probability distributions in the two-slice network:

$$P(\mathbf{X}_t | \mathbf{X}_{t-1}) = \prod_i^N P(X_t^i | \text{Pa}(X_t^i))$$

where  $X_t^i$  is the  $i$ 'th node in time slice  $t$ , which may be hidden or observed, and  $\text{Pa}(X_t^i)$  is the parents of  $X_t^i$  from the same or previous time slice (assuming a first-order Markov model). The joint distribution for a sequence of time steps is obtained by unrolling the network and taking the product of the conditional probability distributions:

$$P(\mathbf{X}_{1:T}^{1:N}) = \prod_i^N P(X_1^i | \text{Pa}(X_1^i)) \times \prod_{t=2}^T \prod_{i=1}^N P(X_t^i | \text{Pa}(X_t^i))$$

Initial state distribution                      Conditional distributions



Source: Murphy, 1998

**Figure C-5 Hidden Markov Models**

By modelling the temporal relationship between variables explicitly, DBNs allow us to reason about changes over time. While many of the standard inference techniques can be used to carry out probabilistic inference in DBNs, in some cases, algorithms have been developed for certain classes of DBNs which exploit the model structure. For example, on-line filtering of a State-Space Model with linear-Gaussian nodes can be carried out using the Kalman filter from classical control theory.

For more detailed information on DBNs, including an overview of the various model structures and inference techniques, refer to Murphy (2002).

#### C.4 Inference in Bayesian Networks

Bayesian networks provide a framework for reasoning under uncertainty (Korb and Nicholson 2004). In diagnostic reasoning, a particular symptom, or outcome is observed, and the analyst attempts to determine what may have caused the observation. The reverse is true in predictive reasoning – new information is received about a variable known to influence the system, and the analyst attempts to predict the outcome. Inter-causal reasoning involves reasoning about the mutual causes of a common effect, for example, using evidence to explain away an alternative cause for a given event.

In all cases, evidence is used to update beliefs about the system. This process is known alternatively as conditioning, probabilistic inference, belief updating, and probability propagation. As new evidence becomes available, it is used to calculate the posterior probability distribution for each node in the Bayesian network. The posterior probability distribution is simply the probability after incorporating evidence; the prior probability distribution can thus be thought of as one's belief before evidence is received.

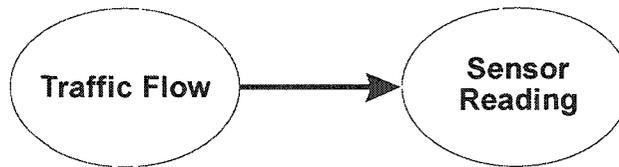
Calculation of the posterior probability distribution can be carried out using Bayes' rule:

$$P(\mathbf{a} | \mathbf{e}) = \frac{P(\mathbf{e} | \mathbf{a}) P(\mathbf{a})}{P(\mathbf{e})}$$

According to Bayes' rule, the probability of an event ' $\mathbf{a}$ ' conditioned on some evidence ' $\mathbf{e}$ ' is equal to the likelihood of the evidence,  $P(\mathbf{e} | \mathbf{a})$ , multiplied by the prior probability of ' $\mathbf{a}$ ', normalized by  $P(\mathbf{e})$ .

The application of Bayes' rule provides a basis for updating our beliefs given new evidence. After applying Bayes' rule to obtain  $P(\mathbf{a} | \mathbf{e})$ , the prior probability conditioned on the evidence, this probability is adopted as the posterior belief in ' $\mathbf{a}$ ':  $\mathbf{Bel}(\mathbf{a}) = P(\mathbf{a} | \mathbf{e})$ .

The following example demonstrates the application of Bayes' Rule for probabilistic inference in a simple Bayesian network with two nodes.

**EXAMPLE: Monitoring of Traffic Flow****Prior Probabilities:**

$P(\text{Flow} = \text{Congested}) = 75\%$   
 $P(\text{Flow} = \text{Not Congested}) = 25\%$

**Conditional Probabilities:**

$P(\text{Sensor} = \text{True} \mid \text{Flow} = \text{Congested}) = 90\%$   
 $P(\text{Sensor} = \text{False} \mid \text{Flow} = \text{Congested}) = 10\%$   
 $P(\text{Sensor} = \text{True} \mid \text{Flow} = \text{Not Congested}) = 20\%$   
 $P(\text{Sensor} = \text{False} \mid \text{Flow} = \text{Not Congested}) = 80\%$

**Given a sensor reading of “True”, what is the probability that traffic flow has broken down?**

**Using Bayes’ Rule:**

$P(\text{Flow} = \text{“Congested”} \mid \text{Sensor} = \text{“True”}) =$

$$\frac{P(\text{Sensor} = \text{“True”} \mid \text{Flow} = \text{“Congested”}) \times P(\text{Flow} = \text{“Congested”})}{P(\text{Sensor} = \text{“True”})}$$

$$= \frac{90\% \times 75\%}{75\% \times 90\% + 25\% \times 20\%} = 93\%$$

**Figure C-6 Example of Probabilistic Inference Using Bayes’ Rule**

In general, probabilistic inference in network chains is straightforward:

- **Predictive Inference:** If the evidence is in the root node, updating is performed in the direction of the arcs using the chain rule from probability theory and taking advantage of the independencies represented in the network.
- **Diagnostic Inference:** If the evidence is in the leaf node, updating is performed using Bayes’ rule in combination with the chain rule, again taking advantage of the conditional independence assumptions encoded in the network.

As the above example demonstrates, for very simple network structures (i.e. chains), it is possible to apply Bayes’ rule directly to compute the posterior probability distribution for a given node. However, more complicated network structures require more sophisticated algorithms. Appendix E provides an overview of various inference techniques. While exact inference often works well in small to medium sized networks (Korb and Nicholson 2004), approximate methods are generally needed as the size and complexity of the network increases.

In the case of the ramp control algorithm developed in this research, approximate inference is used for updating beliefs in the Bayesian network representing the freeway state. The use of approximate inference reflects the dynamic nature of the system, the complexity of the relationships involved, and the real-time constraints of the problem. The specific type of approximate inference used in the algorithm is known as particle filtering, since the posterior distribution is computed by generating a large number of samples or particles. A description of particle filtering can be found in Section 7.5.

### C.5 Decision-Making and Bayesian Networks

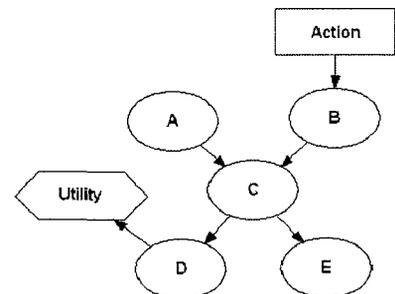
Since Bayesian networks are designed to model uncertainty, they are particularly well-suited for decision applications where impacts cannot be predicted precisely.

In decision theory, utility represents the level of satisfaction derived from the outcome of a particular course of action, taking risk and uncertainty into account. Where outcomes are uncertain, it is assumed that people will act to maximize their expected utility. Expected utility reflects the probability of each possible outcome,  $i$ , as well as its corresponding utility:

$$\text{Expected Utility} = \sum_i \text{Pr obability}_i \times \text{Utility}_i$$

Thus, for a given option, the expected utility is calculated by multiplying the probability of each potential outcome by the utility of the outcome, and summing the results.

Outcomes may be defined based on a single attribute, or may incorporate multiple attributes which reflect different objectives. In the case of the latter, the multi-attribute utility is a function of the utility of each of the individual attributes aggregated in a way that reflects their relative importance and degree of interaction.



In Bayesian networks, a multi-attribute utility function can be represented as a utility node (or a series of nodes, if the single-attribute utilities are shown separately). The parents of a utility node represent the system attributes on which the utility is based. Given a probability distribution for these attributes, the expected utility is easily computed.

To be used in decision problems, a Bayesian network must also include action (or decision) nodes. Such nodes represent specific actions or policies that can be carried out on or by the system, impacting the system state and the associated utility.

A Bayesian network containing utility and action nodes is called a decision network or influence diagram. Using such networks, it is possible to compute the optimal action (or sequence of actions) which will maximize the expected utility. In the ramp metering control problem, decision nodes represent the ramp metering rates to be applied at different on-ramps, while utility nodes represent preferences for various system outcomes as related to the control objectives.

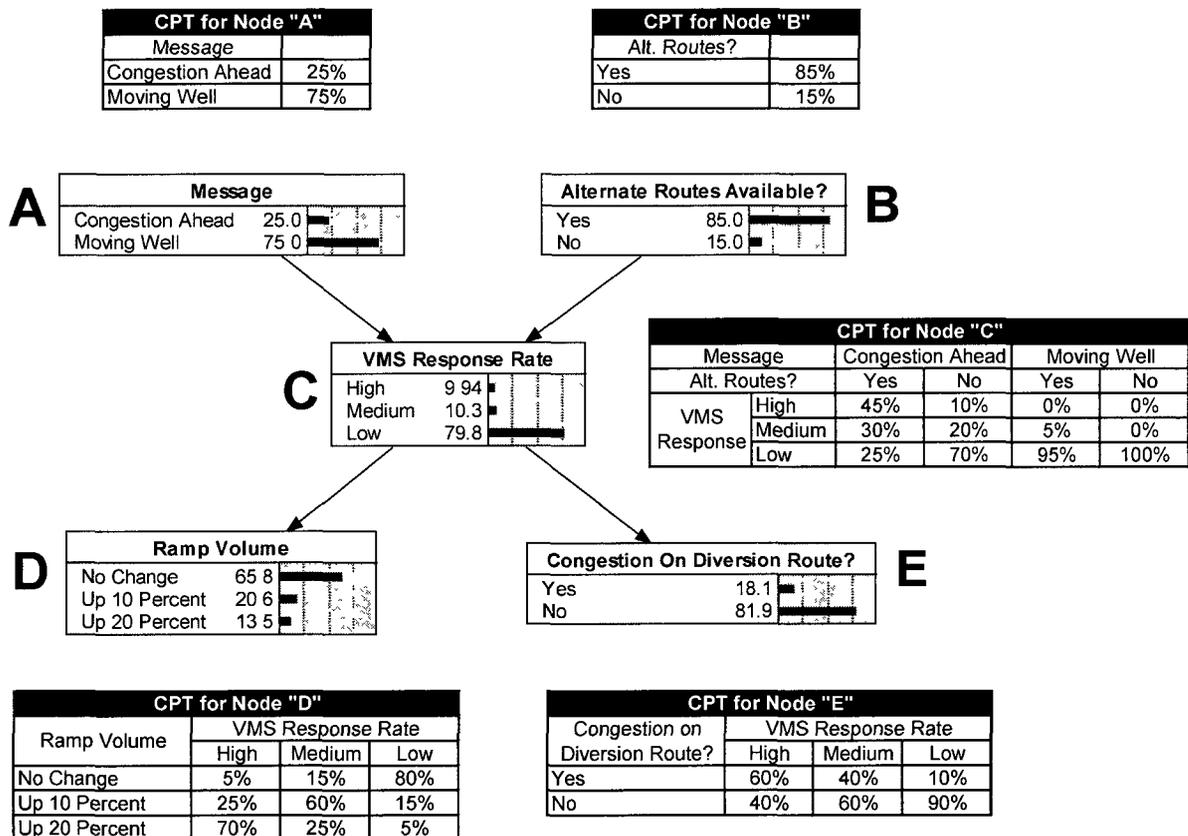
# **APPENDIX D**

## **BAYESIAN NETWORK EXAMPLE**

**D. DRIVER RESPONSE TO VARIABLE MESSAGE SIGNS – A SIMPLE BAYESIAN NETWORK EXAMPLE**

A Bayesian network consists of a directed acyclic graph and a set of local conditional probability distributions which together represent the joint probability distribution of the random variables encoded in the network. The graph identifies direct dependencies between variables in qualitative terms, while the probability distributions quantify the strength of those relationships.

Figure D-1 presents a simple Bayesian network (developed in Netica) for modelling the response of freeway drivers to variable message signs (VMS). In this illustrative example, the response rate is influenced by both the message on the sign, as well as the availability of alternative routes. In turn, the response rate influences the traffic volume observed on the exit ramp downstream from the sign, as well as the level of congestion along the diversion route.



**Figure D-1 Bayesian Network Example – Response to Variable Message Signs**

According to convention, Node E is denoted as a child of Node C reflecting the directionality of the arc connecting the two nodes. Node C is therefore considered a parent of Node E, while Nodes A and B are ancestors of Node E. A similar naming structure can be applied for the remaining nodes in the network.

For each node, a conditional probability table (CPT) is specified. These tables give the probability of each potential outcome given all possible combinations of values of the parent nodes. For example, the probability of congestion on the diversion route is estimated to be 10% given a low VMS response rate. This reflects the fact that congestion may sometimes occur even without significant diversion taking place. For nodes without parents, the prior probability is specified. From an examination of all variable message signs in the corridor, it is estimated that the message “moving well” is displayed roughly 75% of the time. Likewise, of all VMS locations, roughly 85% are in areas with viable alternative routes.

Once the conditional probability distributions have been specified, the prior probability distributions can be calculated for each node (i.e. the probabilities in the absence of evidence). The calculations are straightforward – working from top to bottom, simply multiply each entry in the node’s conditional probability table corresponding to a particular instantiation of the parent nodes by the probability that the parent nodes will take on those values, and sum over each possible outcome. For example, the probability of a high VMS response rate can be calculated as follows:

$$\begin{aligned}
 & P(\text{Message} = \text{M. Well}) \times P(\text{Alt. Route} = \text{yes}) \times P(\text{Response} = \text{high} \mid \text{Message} = \text{M. Well}, \text{Alt. Route} = \text{yes}) \\
 & + P(\text{Message} = \text{M. Well}) \times P(\text{Alt. Route} = \text{no}) \times P(\text{Response} = \text{high} \mid \text{Message} = \text{M. Well}, \text{Alt. Route} = \text{no}) \\
 & + P(\text{Message} = \text{Cong}) \times P(\text{Alt. Route} = \text{yes}) \times P(\text{Response} = \text{high} \mid \text{Message} = \text{Cong}, \text{Alt. Route} = \text{yes}) \\
 & + P(\text{Message} = \text{Cong}) \times P(\text{Alt. Route} = \text{no}) \times P(\text{Response} = \text{high} \mid \text{Message} = \text{Cong}, \text{Alt. Route} = \text{no}) \\
 & = (75\% \times 85\% \times 0\%) + (75\% \times 15\% \times 0\%) + (25\% \times 85\% \times 45\%) + (25\% \times 15\% \times 10\%) = 9.94\%
 \end{aligned}$$

The prior probabilities for each node computed using Netica are illustrated in the bar graphs in Figure D-1.

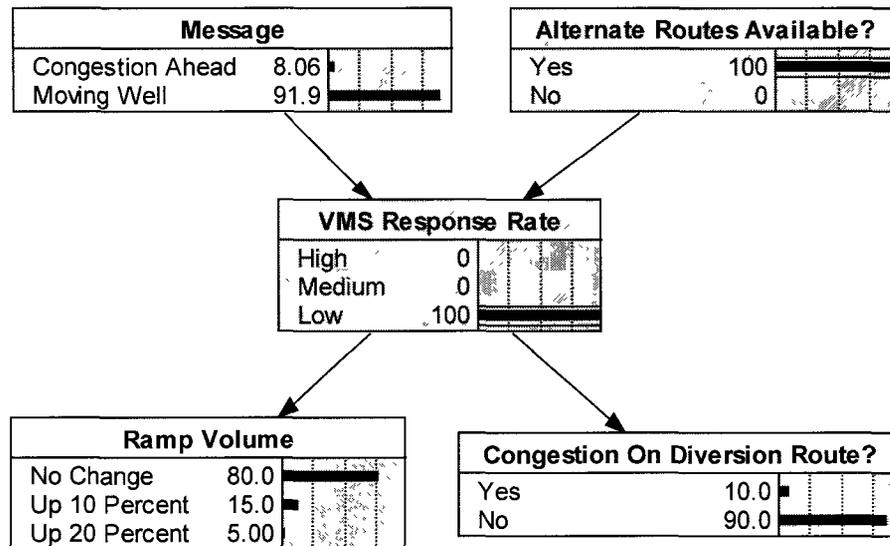
An examination of Figure D-1 also provides insight into the relationship between variables which are not connected by arcs. Based on the network structure, congestion on the diversion route is conditionally independent of ramp volume, given the VMS response rate. If there is no information on the VMS response rate, knowledge of the ramp volume will influence the probability distribution for the VMS response rate, which in turn will influence the probability distribution for the congestion variable. However, if the response rate is known, additional information on the ramp volume is superfluous, and will not tell us anything more about the probability of congestion on the diversion route.

Similarly, both the ramp volume and congestion variables are conditionally independent of the message shown on the VMS and the availability of alternative routes, given the VMS response rate. Since the VMS response rate is a common “cause” of the ramp volume and congestion variable, information about the parents (or ancestors) of the VMS response rate does nothing to influence our beliefs about the ramp volume or congestion level, given knowledge of the response rate itself.

In contrast, the message displayed on the VMS and availability of alternative routes are conditionally dependent on the value of the VMS response rate. If the response rate is

unknown, the two variables are independent. However, if the response rate is known, knowledge about the availability of alternative routes will influence our beliefs about the message displayed on the VMS, and vice versa. For example, if the VMS response rate is low, and we know that there are many alternative routes available, the probability of a message indicating congestion ahead declines, and we should update our beliefs accordingly (refer to Figure D-2).

The above example also illustrates the various types of inference that can be performed using Bayesian networks. For example, if the message displayed on the VMS is known, the impact on the VMS response rate can be estimated using predictive inference. In contrast, if evidence is available on the ramp volume downstream of the VMS, it is possible to work backwards and update beliefs about the message shown on the sign using diagnostic inference. A discussion of inference techniques can be found in Appendix E.



**Figure D-2 Example of Belief Updating Given Evidence**

# **APPENDIX E**

## **INFERENCE TECHNIQUES FOR BELIEF UPDATING IN BAYESIAN NETWORKS**

## E. INFERENCE IN BAYESIAN NETWORKS

### E.1 Exact Inference

A number of exact inference algorithms have been developed to compute the posterior probability distribution for a set of query nodes given the value of one or more evidence nodes.<sup>1</sup> Some algorithms have been incorporated into commercially available software packages, others have had only minimal application outside the research environment where they were developed (Korb and Nicholson 2004). In general, different algorithms are suited to different network structures. Simple chains can be solved using repeated application of Bayes' rule. Polytrees, or singly-connected networks, require more complex algorithms, while multiply-connected networks tend to be the most difficult to solve.

Polytrees have at most one path between any pair of nodes. The resulting "tree" structure provides a basis for probabilistic inference. In polytrees, local belief updating at a particular node must incorporate evidence from all other parts of the network, including evidence transmitted downward via parent nodes (predictive evidence), as well as evidence propagated upward via child nodes (diagnostic evidence). In such networks, a simple message passing algorithm based on local computations can be employed to propagate evidence to other parts of the network. In essence, at each iteration of the algorithm, local belief updating is carried out at node  $X$  based on messages arriving from parent and child nodes. New messages are then propagated to neighboring nodes so that they can perform updates. The algorithm relies on Bayes' rule, and takes advantage of the conditional independence assumptions encoded in the polytree structure.

In the more general case, Bayesian networks are comprised of directed acyclic graphs, not trees, forming multiply-connected networks. In such networks, pairs of nodes may be connected by multiple paths – a situation which can arise when one variable influences another through more than one causal mechanism. For example, weather can have a direct influence on the level of congestion on a particular roadway. Weather can also have an impact on collision occurrence, which in turn influences road congestion as well.

In multiply-connected networks, a two-step inference process is often followed. In the first step, clustering methods are used to convert the network to an equivalent polytree by merging nodes and restructuring the network so that multiple paths are eliminated. In the second step, belief updating is performed on the converted network using a message passing algorithm.

One of the most popular clustering techniques is the so-called junction tree algorithm which provides a systematic and efficient method of clustering (for a good explanation of the junction tree algorithm, refer to any introductory textbook, such as Korb and Nicholson 2004). In general, transforming the network into an equivalent polytree may be computationally slow depending on the conditional probability distributions to be

---

<sup>1</sup> Note that much of the discussion in this appendix is based on the work by Korb and Nicholson (2004) and Jensen (2001). The reader is urged to consult these references directly for additional information.

adjusted. Computer memory requirements may also be an issue if the original network is highly connected. Fortunately, the network conversion only needs to be completed once as long as the original model remains unchanged.

The above discussion provides a brief introduction to the most popular exact inference techniques. Variations of these and other algorithms are reported in the literature (see for example Guo and Hsu 2002). In general, the performance of a particular algorithm depends on the network structure, including its level of connectivity, the number of undirected loops, and the location of evidence and query nodes (Korb and Nicholson 2004). Exact inference generally works well for small to medium sized networks, particularly if the level of connectivity is low (Korb and Nicholson 2004). However, larger, more complex networks often require approximate inference techniques.

## E.2 Approximate Inference

As the size and complexity of a Bayesian network increases, exact inference becomes infeasible, and approximate methods must be used. Many approximate methods are based on stochastic sampling: by generating a large number of samples which reflect the probability distributions in the Bayesian network, an estimate of the posterior probability distribution for each variable can be obtained.

### E.2.1 Logic Sampling

In logic sampling, a number of simulations are conducted using the Bayesian network as a model of the process or system under investigation. In each simulation, the nodes in the network are sampled sequentially in the order of the arcs, so that parent nodes are sampled prior to sampling their children. At each node, a value is selected randomly based on the node's conditional probability distribution and the value assigned to its parents, thus ensuring that the sampled values are weighted by their probability of occurrence. If the node is a root node, and has no parents, sampling is carried out using the node's prior probability distribution.

Once the network has been fully sampled, the values generated during the simulation are compared with the observed evidence. If any of the values are inconsistent with the evidence, the simulation is discarded. This process is repeated until a sufficient number of 'valid' simulations have been generated.

The probability of a query node assuming a particular value  $\mathbf{x}_1$  given the observed evidence is calculated by counting the frequency with which  $\mathbf{x}_1$  is observed in the set of valid simulations consistent with the evidence. In other words, the posterior probability that  $\mathbf{X} = \mathbf{x}_1$  given evidence  $\mathbf{E} = \mathbf{e}$  can be calculated by taking the ratio of the number of simulations where both  $\mathbf{X}$  and  $\mathbf{E}$  are true to the number of simulations where just  $\mathbf{E}$  is true, as follows:

$$P(\mathbf{X} = \mathbf{x}_1 \mid \mathbf{E} = \mathbf{e}) = \frac{\text{Number of Simulations with } \mathbf{X} = \mathbf{x}_1 \text{ and } \mathbf{E} = \mathbf{e}}{\text{Number of Simulations with } \mathbf{E} = \mathbf{e}}$$

Since the network is sampled following the order of the arcs, logic sampling is considered a type of forward sampling. In general, logic sampling works well if no evidence has been observed. However, if evidence exists, samples inconsistent with the evidence must be discarded. If the evidence is unlikely, many of the samples generated will be wasted, and the algorithm performs poorly. This is particularly true for large networks with several evidence nodes, since the prior probability of the evidence is often quite small (Guo and Hsu 2002).

An algorithm for logic sampling can be found in Korb and Nicholson (2004). Cheng and Druzdzel (2000) describe how logic sampling fits within the more general framework of importance sampling in Bayesian networks.

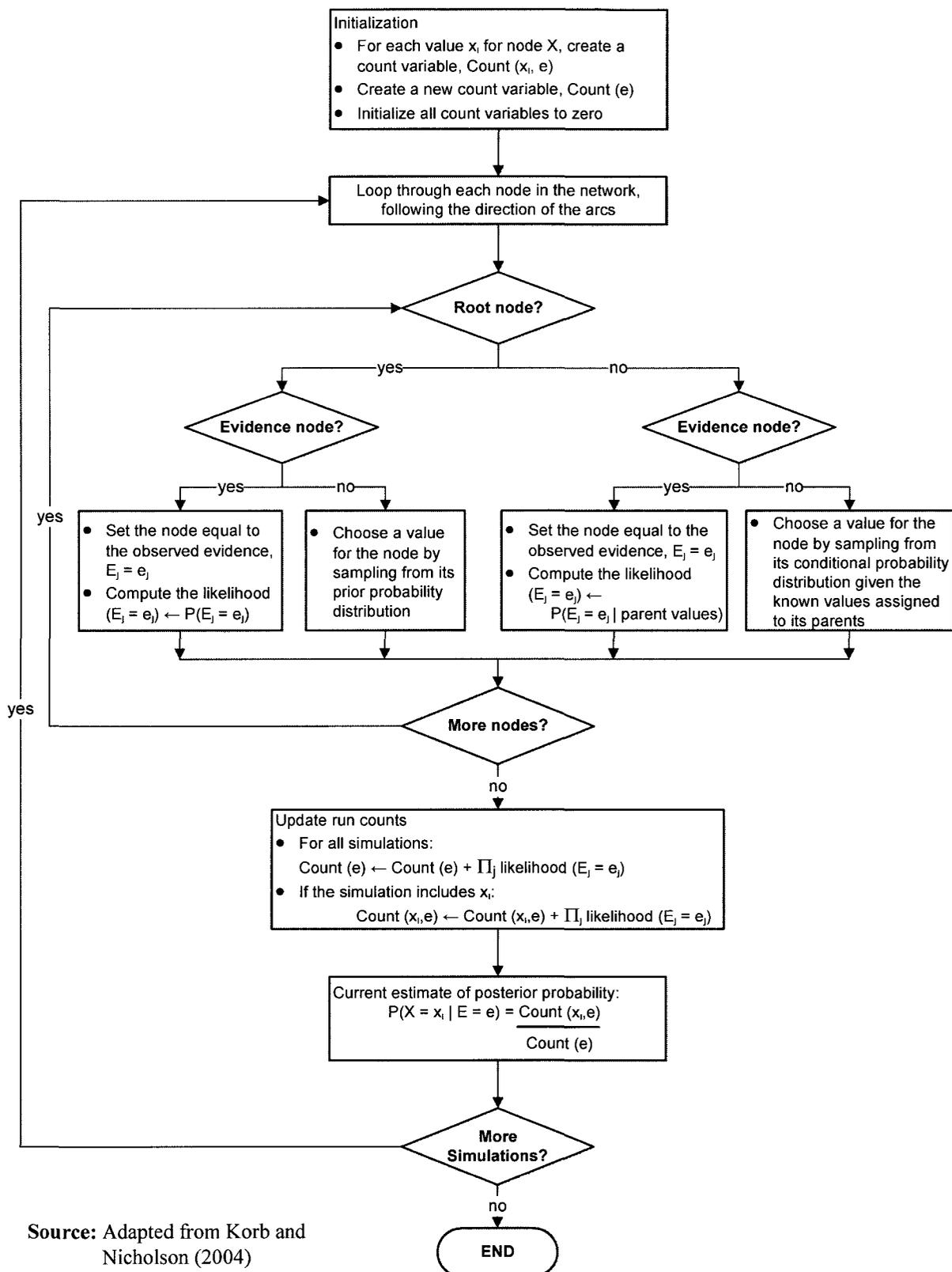
### E.2.2 Likelihood Weighting

Likelihood weighting overcomes the deficiencies of logic sampling by setting the value of each evidence node equal to the observed evidence. Each simulation is then weighted by the likelihood of the evidence conditional on the values assigned to the other nodes.

Unlike logic sampling, no simulations are discarded, and convergence is typically achieved much faster (Guo and Hsu 2002). According to Cheng and Druzdzel (2000), likelihood weighting has been the most widely used simulation method for Bayesian network inference, largely due to its simplicity. Not only is the algorithm easy to implement, but in many cases, it is able to match the performance of other more sophisticated algorithms since it can generate more samples in the same amount of time, increasing precision. Likelihood weighting can handle very large networks, however, as with logic sampling, convergence can be slow for unlikely events.

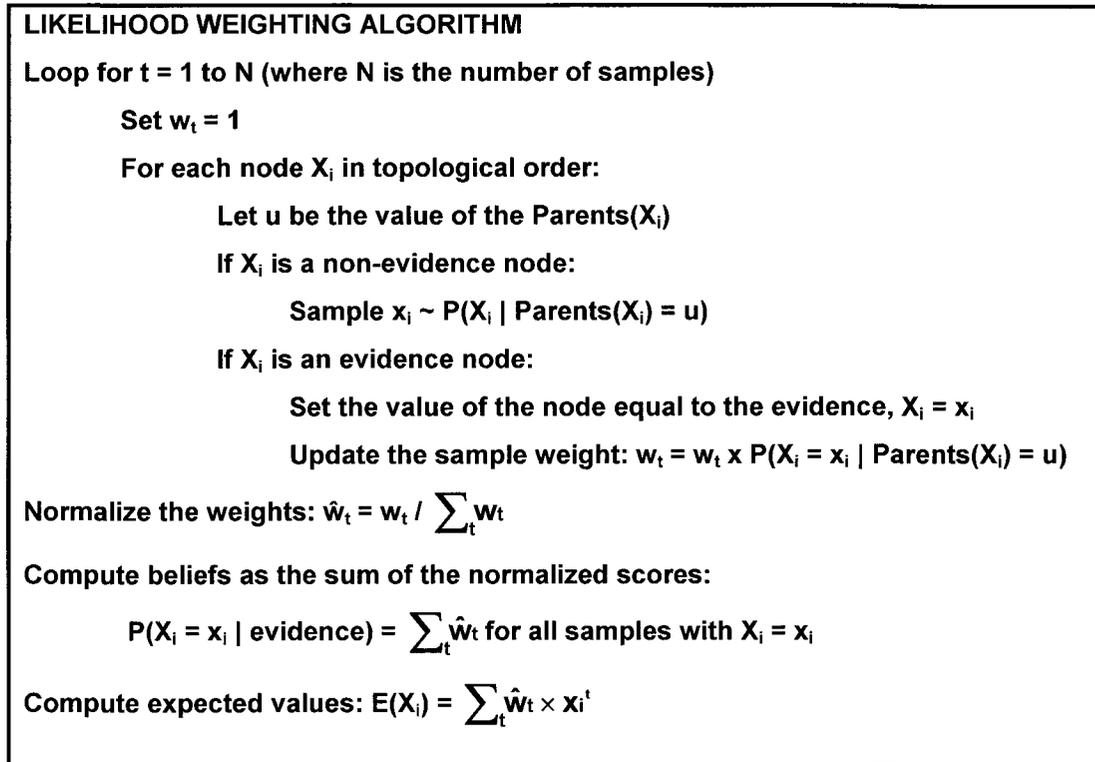
An algorithm for likelihood weighting is presented in Figure E-1. In this figure, the likelihood of the evidence for a particular node is calculated as  $P(E_j = e_j \mid \text{Parents}(E_j))$  – the probability of the evidence node taking on the observed value given the values assigned to its parents. Multiplying these individual likelihood values together provides an estimate of the overall likelihood of the complete set of evidence for the simulation in question.

The algorithm as presented in Figure E-1 is particularly suited for Bayesian networks containing discrete random variables, since it computes the posterior probability of a query node  $X$  taking on some particular value  $x_i$ . In the case of continuous variables, the usual practice is to compute the expected value of the variable given the observed evidence, rather than its posterior probability distribution. This can be done by weighting each trial by the likelihood of the observed evidence, and then normalizing the weights. A weighted average can then be computed for each variable of interest, representing its expected value. The process is shown in Figure E-2.



Source: Adapted from Korb and  
Nicholson (2004)

**Figure E-1 Likelihood Weighting Algorithm – Version 1**



Source: Adapted from Murphy (2002)

**Figure E-2 Likelihood Weighting Algorithm – Version 2**

The algorithm in Figure E-2 is essentially the same as the algorithm depicted graphically in Figure E-1, with one or two minor differences related primarily to how the likelihood scores are developed and applied. In Figure E-1, the likelihood of the evidence is calculated individually for each node and then combined, whereas in Figure E-2, the evidence is used to update the sample weight incrementally as each evidence node is considered in sequence. Once all the samples have been generated, the weights are normalized. These normalized weights replace the count variables used in Figure E-1. In essence, the weights represent the likelihood of the particular combination of evidence that was observed given the values of the query nodes in a particular sample.

Instead of calculating the posterior probability for a particular event as the ratio of the two count variables, the posterior probability simply equals the sum of the normalized weights for those samples where the event occurred (i.e.  $X_i = x_i$ ). This simplification is possible since all samples are consistent with the observed evidence, and by definition, the sum of the normalized weights for the entire sample set (i.e. the denominator of the count ratio) is 1.

### E.2.3 Markov Chain Monte Carlo

In both logic sampling and likelihood weighting, the samples are generated independently. Algorithms based on Markov Chain Monte Carlo (MCMC) techniques also fall within the domain of stochastic sampling, but instead of generating the samples

individually, the samples are dependent. To generate a new sample, a random change is made to the previous sample. The resulting sequence forms a Markov chain, since values from the current step depend only on values from the previous step. It can be shown that under certain conditions, the Markov chain has a stationary limiting distribution which approximates the target distribution. After a suitable burn-in period to ensure convergence, the Markov chain can be used to estimate the expected value of a given node by calculating the sample mean.

Two of the most popular MCMC algorithms are the Metropolis-Hastings algorithm and Gibbs sampling.

In the Metropolis-Hastings algorithm, we start with an arbitrary point  $\mathbf{j}$  in the state space  $\mathbf{S}$ . A random variable  $\mathbf{k}$  is then generated from an arbitrary but fixed proposal distribution  $\mathbf{P}(\mathbf{k} | \mathbf{j})$ , representing a proposed move from state  $\mathbf{j}$  to state  $\mathbf{k}$ . A decision is made to reject or accept the proposed change, and the process is repeated.

Gibbs sampling is a special case of the Metropolis-Hastings algorithm in which each non-evidence node in the Bayesian network is sampled sequentially conditioned on the current value of the nodes in its Markov blanket (i.e. its parents, children, and children's parents). In the simple case of two variables,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , Gibbs sampling proceeds by setting the variables equal to some initial value, and generating a sequence of  $(\mathbf{X}_1^{(k)}, \mathbf{X}_2^{(k)})$  by sampling from the conditional probability distributions as follows:

- $\mathbf{X}_1^{(k)} \sim \mathbf{P}(\mathbf{X}_1 | \mathbf{X}_2^{(k-1)})$
- $\mathbf{X}_2^{(k)} \sim \mathbf{P}(\mathbf{X}_2 | \mathbf{X}_1^{(k)})$

The resulting sequence from this “random walk” forms a Markov Chain which tends to the desired distribution. Note that Gibbs sampling requires sampling from a variable's full conditional probability distribution (i.e. its distribution conditional on all variables), which, in a Bayesian network, simplifies to the node's Markov blanket. By setting the proposal distribution equal to the full conditional distribution, all proposed changes are automatically accepted.

A more detailed explanation of MCMC methods can be found in Brooks (1998). Chib and Greenberg (1995) provide a good overview of the Metropolis-Hastings algorithm, while Casella and George (1992) describe the Gibbs Sampler. The use of Gibbs sampling in graphical models is described in Korb and Nicholson (2004) in the context of learning causal models, and more generally in Cowell et al. (1999). The BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling) software system is a popular tool for carrying out Bayesian analysis of complex statistical models and is available on the internet (BUGS 2008; Lunn et al. 2000).

## E.2.4 Particle Filters

Particle filters approximate the belief state using a set of weighted samples (or “particles”):

$$P(\mathbf{X}_t | \mathbf{y}_{1:t}) \approx \sum_{i=1}^N w_t^i \delta(\mathbf{X}_t, \mathbf{X}_t^i)$$

where  $\mathbf{X}_t^i$  is the  $i^{\text{th}}$  sample of state  $\mathbf{X}$  at time  $t$ ,  $\mathbf{y}$  is the evidence,  $\mathbf{w}$  is the sample weight and  $\delta$  is the Dirac delta function.

Particle filtering is essentially sequential importance sampling with re-sampling. With a prior distribution represented by a set of particles, the posterior distribution can be estimated by sampling from a proposal distribution, and weighting the samples appropriately. The most common proposal distribution is the transition prior:

$P(\mathbf{X}_t | \mathbf{X}_{t-1}^i)$ . In this case, the weights simplify to  $w_t^i = P(\mathbf{y}_t | \mathbf{X}_t^i)$ , where  $\mathbf{X}$  represents the state, and  $\mathbf{y}$  represents the evidence.

Additional information on particle filters can be found in the main report.

## E.2.5 Other Inference Techniques

The above discussion has touched on only a few of the many inference algorithms reported in the literature. Guo and Hsu (2002) present a more comprehensive survey of the various exact and approximate inference techniques that have been developed, with particular emphasis on algorithms that can be used for real-time inference. The major classes of inference algorithms identified by Guo and Hsu are presented in Table E-1. For each class of algorithm, several variants, refinements, hybrids, generalizations, and/or heuristic solutions have been developed – a fact reflected in Guo and Hsu’s extensive reference list, which includes over 110 citations.

**Table E-1 Major Classes of Exact and Approximate Inference Techniques**

Exact Inference	Approximate Inference
<ul style="list-style-type: none"> <li>• Message passing algorithm (for polytrees)</li> <li>• Conditioning</li> <li>• Clustering</li> <li>• Arc reversal/node reduction</li> <li>• Variable elimination</li> <li>• Symbolic</li> <li>• Differential method</li> </ul>	<ul style="list-style-type: none"> <li>• Stochastic Sampling (logic sampling, likelihood weighting, importance sampling, MCMC, etc.)</li> <li>• Model simplification</li> <li>• Search-based</li> <li>• Loopy belief propagation</li> </ul>

Adapted from: Guo and Hsu (2002)

Minka (2007) provides an overview of on-line filtering algorithms for dynamic Bayesian networks. Such algorithms are often categorized according to how the posterior probability distribution is represented over the hidden state. For example, models with Gaussian noise can be analyzed using the Kalman filter or one of its variants, while particle filters represent the state posterior by a set of samples. A more detailed discussion of inference techniques for DBNs is presented in Murphy (2002). Algorithms range from the forwards-backwards algorithm for HMMs to the Boyen-Koller algorithm which represents the belief state as a product of marginals over clusters of variables.

Clearly, one of the major challenges in working with Bayesian networks is identifying the most appropriate inference algorithm for carrying out belief updating as evidence is received. This is particularly true for real-time inference in dynamic networks, where algorithm performance is measured not only by accuracy, but also speed.

# **APPENDIX F**

## **MACROSCOPIC MODELS OF TRAFFIC FLOW**

## F. MACROSCOPIC MODELS OF TRAFFIC FLOW

### F.1 The METANET Model

The METANET model is a second-order non-linear model of traffic flow that is discrete in both time and space (Kotsialos et al. 2002). To compute the speed, density, and flow for each segment  $j$  at time step  $k$ , the following equations are applied (using the notation from Bellemans et al. 2006a):

$$\text{Equation 1: } \rho_j(\mathbf{k} + \mathbf{1}) = \rho_j(\mathbf{k}) + \frac{\Delta T}{n_j l_j} [\mathbf{q}_{j-1}(\mathbf{k}) - \mathbf{q}_j(\mathbf{k})]$$

Equation 2:

$$v_j(\mathbf{k} + \mathbf{1}) = v_j(\mathbf{k}) + \underbrace{\frac{\Delta T}{\tau} [v(\rho_j(\mathbf{k})) - v_j(\mathbf{k})]}_{\text{Relaxation term}} + \underbrace{\frac{\Delta T}{l_j} v_j(\mathbf{k}) [v_{j-1}(\mathbf{k}) - v_j(\mathbf{k})]}_{\text{Convection term}} - \underbrace{\frac{v \Delta T [\rho_{j+1}(\mathbf{k}) - \rho_j(\mathbf{k})]}{\tau l_j [\rho_j(\mathbf{k}) + \kappa]}}_{\text{Anticipation term}}$$

$$\text{Equation 3: } \mathbf{q}_j(\mathbf{k}) = \rho_j(\mathbf{k}) v_j(\mathbf{k}) n_j$$

$$\text{Equation 4: } v(\rho_j(\mathbf{k})) = v_f \exp\left(-\frac{1}{a_m} \left(\frac{\rho_j(\mathbf{k})}{\rho_{\text{crit},j}}\right)^{a_m}\right)$$

where  $\rho$ ,  $v$ , and  $\mathbf{q}$  refer to the density, speed, and flow respectively,  $n$  is the number of lanes on section  $j$ ,  $l$  is the length of section  $j$ ,  $\Delta T$  is the simulation time step,  $v_f$  is the free flow speed,  $\rho_{\text{crit}}$  is the critical density, and  $\tau$ ,  $\kappa$ ,  $v$ , and  $a_m$  are calibration parameters.

- **Equation 1** ensures the conservation of vehicles – the traffic density on a given segment equals the previous density, adjusted for the number of vehicles entering and leaving the segment during the time interval  $\Delta T$ .
- **Equation 2** provides an estimate of the average speed in section  $j$  as a function of three phenomena: relaxation, convection, and anticipation. Relaxation refers to the tendency of the average speed to evolve in accordance with an empirically derived relationship between density and speed (expressed in **Equation 4**). Convection accounts for the impact of the vehicles entering the segment, while anticipation captures any speed adjustments due to downstream conditions.
- **Equation 3** gives the basic relationship between speed, density, and flow.

Additional terms can be added to Equation 2 to model lane drops or merging behaviour near on-ramps.

## F.2 The Cell Transmission Model

The cell transmission model was first proposed by Daganzo (1994). As originally defined, the cell transmission model is a relatively simple first order model which captures the forward propagation of traffic flow and the backward propagation of congestion. The model is discrete in both space and time. To apply the model, the highway is divided into homogenous cells such that no vehicle can cross more than one cell boundary during a single time step. The model attempts to predict the number of vehicles  $n$  in each freeway cell  $j$  at time  $k$  by applying the conservation of vehicles principle at each time step:

$$n_j(k+1) = n_j(k) + q_j(k) - q_{j+1}(k)$$

The flow  $q$  into each cell is defined as the minimum of: the flow that can be supplied by the upstream cell, the flow that can be absorbed by the cell given its available storage, and the maximum observable flow (i.e. the flow capacity). Thus, the number of vehicles in a given cell  $j$  is equal to the previous number of vehicles in the cell, plus the number of vehicles arriving in the current time step, minus the number of vehicles departing.

Several research teams have modified the cell transmission model for specific applications. The asymmetric cell transmission model developed by Gomes and Horowitz (2006) includes a modified treatment for on-ramp merges. Sun et al. (2003) used the cell transmission model as the basis for a switching state-space model with two discrete modes of operation, congested and free-flow.

In another notable example, Mihaylova et al. (2007) expanded Daganzo's cell transmission model to create the 'compositional traffic model' (see also Boel and Mihaylova 2004, 2006). Whereas Daganzo modeled forward and backward traffic waves using deterministic relationships, Mihaylova et al. (2007) define stochastic sending and receiving functions, and also introduce a speed variable to model changes in the average speed over time. Both the sending and receiving functions are speed-dependent, the latter influenced by the state of downstream cells, the former influenced by the state of upstream cells. The equations defining the compositional traffic model are summarized in Figure F-1. The sending function  $S_i$  represents the number of vehicles which would leave cell  $i$  if not constrained by downstream conditions. The receiving function  $R_i$  represents the maximum number of vehicles allowed to exit cell  $i$  based on the state of cell  $i+1$ . The actual flow between the two cells is the lower of  $S_i$  and  $R_i$ .

The sending function is based on the number of vehicles in the upstream cell that are within a distance  $v\Delta t$  of the cell boundary. Assuming a uniform distribution of vehicles, the probability of crossing the cell boundary during the time interval  $\Delta t$  is  $v\Delta t / L$ . To prevent the flow from dropping to zero in the event of a severe traffic jam, a minimum outflow speed is defined. The receiving function is based on the maximum number of vehicles that can be simultaneously present in the downstream cell, calculated as the total available space,  $L$ , divided by the space taken up by vehicles,  $A_v + vt_d$  (the average vehicle length plus the safety distance required to achieve a minimum time separation  $t_d$ ).

The speed calculation includes two components. The intermediate speed assumes that vehicle speeds generally remain constant as vehicles move downstream. Thus, the average speed in a cell depends only on the speed of the vehicles entering the cell during the current time step, and the vehicles remaining in the cell from previous time steps. In contrast, the density-dependent speed reflects drivers' tendency to adjust their speed based on prevailing conditions. The density-dependent speed can be calculated using any empirically defined speed-density relationship, including the one used in the METANET model (see Equation 4 above). The overall speed is a weighted average of the intermediate speed and density-dependent speed, with the weighting factor  $\beta$  expressing how aggressively drivers adjust their speed in response to changing conditions. Note that should the receiving function dominate (i.e. congestion is moving upstream), the average speed is re-calculated to reflect the corresponding flow rate between cells.

A more detailed description of the compositional traffic model can be found in the references cited above.

**Equations:**

1. *Forward wave*: for  $i = 1, 2, \dots, n$ ,

$$S_{i,k} = \max \left( N_{i,k} \frac{v_{i,k} \cdot \Delta t_k}{L_i} + \eta_{S_{i,k}}, N_{i,k} \frac{v_{\min} \cdot \Delta t_k}{L_i} \right)$$

and set  $Q_{i,k} = S_{i,k}$ .

2. *Backward wave*: for  $i = n, n-1, \dots, 1$ ,

$$R_{i,k} = N_{i+1,k}^{\max} - N_{i+1,k} + Q_{i+1,k},$$

where  $N_{i+1,k}^{\max} = (L_{i+1} v_{i+1,k}) / (A_l + v_{i+1,k} t_d)$ ,

if  $S_{i,k} < R_{i,k}$ ,  $Q_{i,k} = S_{i,k}$ ,

else  $Q_{i,k} = R_{i,k}$ ,  $v_{i,k} = Q_{i,k} L_i / (N_{i,k} \Delta t_k)$ .

3. Update the number of vehicles inside segments.

for  $i = 1, 2, \dots, n$ ,

$$N_{i,k+1} = N_{i,k} + Q_{i-1,k} - Q_{i,k}.$$

4. Update the density, for  $i = 1, 2, \dots, n$ .

$$\rho_{i,k+1} = N_{i,k+1} / (L_i v_{i,k+1}),$$

$$\rho_{i,k+1}^{\text{antic}} = \alpha \rho_{i,k+1} + (1 - \alpha) \rho_{i+1,k+1}.$$

5. Update of the speed, for  $i = 1, 2, \dots, n$ .

$$v_{i,k+1}^{\text{interim}} = \begin{cases} \frac{v_{i-1,k} Q_{i-1,k} + v_{i,k} (N_{i,k} - Q_{i,k})}{N_{i,k+1}} & \text{for } N_{i,k+1} \neq 0, \\ v_f & \text{otherwise,} \end{cases}$$

$$v_{i,k+1}^{\text{interim}} = \max(v_{i,k+1}^{\text{interim}}, v_{\min}).$$

$$v_{i,k+1} = \beta_{k+1} v_{i,k+1}^{\text{interim}} + (1 - \beta_{k+1}) v^e(\rho_{i,k+1}^{\text{antic}}) + \eta_{v_{i,k+1}}.$$

where

$$\beta_{k+1} = \begin{cases} \beta^I & \text{if } |\rho_{i+1,k+1}^{\text{antic}} - \rho_{i,k+1}^{\text{antic}}| \geq \rho_{\text{threshold}}, \\ \beta^{II} & \text{otherwise.} \end{cases}$$

**Variables:**

$i$  = freeway cell number

$k$  = time step

$\Delta t$  = time step duration

$L$  = cell length

$l$  = number of lanes

$S$  = sending function

$R$  = receiving function

$Q$  = flow

$N$  = number of vehicles

$N^{\max}$  = maximum number of vehicles

$\rho$  = density

$\rho^{\text{antic}}$  = anticipated density as drivers look downstream

$v$  = average speed

$v_{\min}$  = minimum outflow speed from a queue

$v^{\text{interim}}$  = intermediate speed accounting for convection

$v^e(\rho)$  = average speed corresponding to density  $\rho$

$A_l$  = average vehicle length

$t_d$  = minimum time separation between vehicles

$\alpha$  = weighting parameter which reflects how far ahead drivers look

$\beta$  = weighting parameter which reflects how aggressively drivers adjust their speed in response to changing conditions

$\eta$  = random noise to account for unpredictable driver behaviour and model error

Source: Mihaylova et al. 2007, pg. 292

**Figure F-1 The Compositional Traffic Model**

# **APPENDIX G**

## **THE FREEWAY TRAFFIC MODEL**

## The Freeway Traffic Model

### Variables:

A, B, C	Parameters for flow breakdown estimation
$\beta$	Parameter for estimating the uncongested speed in the next time step
$\eta$	Used to refer to a random noise term
$\omega$	Used in conjunction with $K^{\text{Breakdown}}$ to establish a threshold for checking for freeway recovery. If the density is greater than $\omega K^{\text{Breakdown}}$ than the freeway is assumed to be congested
Capacity <sup>Congested</sup>	Maximum flow out of the queue (veh/hr/lane)
Capacity <sup>Ramp</sup>	Ramp capacity (veh/hr/lane):
Delay	Ramp delay (minutes)
dt	Duration of each model time step (assumed to be 10 seconds)
FlowType	Indicates type of flow: 0 = uncongested; 1 = breakdown at bottleneck; 2 = within queue
K	Segment density (veh/km)
$K^{\text{Antic}}$	Anticipated density used to develop the uncongested speed estimate for the next time step (veh/km/lane)
$K^{\text{Breakdown}}$	Mainline density when breakdown in the freeway segment was first initiated (veh/km/lane)
$K^{\text{Crit}}$	Critical density (veh/km/lane)
$K^{\text{Queue}}$	Average density within the queue (veh/km/lane)
i	Index referring to the freeway segment number
L	Length of freeway segment
Lanes	Number of through lanes in freeway segment
MaxMeterRate	The maximum ramp metering rate for single lane metering (i.e. 900 veh/hr)
MaxQ <sup>Main</sup>	Maximum potential flow into segment from upstream cell assuming no flow breakdown (excludes ramp flows) (veh/hr)
MeterRate	Ramp metering rate to be applied at the on-ramp (veh/hr)
NVeh	Number of vehicles in the segment
ObsOnFlow	Observed on-ramp flow entering the freeway (veh/hr)
ObsQ <sup>Main</sup>	Observed mainline flow (veh/hr)
ObsRampDem	Observed ramp demand (veh/hr)
ObsXPer	Observed proportion of traffic exiting the freeway (%)
Pr(BD)	Probability of flow breakdown
40SecQ <sup>Main</sup>	40 second mainline flow approaching the merge area (veh/hr)
Q <sup>Main</sup>	Flow into mainline segment (excludes ramp flows) (veh/hr)
Q <sup>Off</sup>	Flow exiting the freeway at an off-ramp (veh/hr)
40SecQ <sup>On</sup>	40 second ramp flow entering the merge area (veh/hr)
Q <sup>On</sup>	Flow entering the freeway at an on-ramp (veh/hr)
Queue	Ramp queue (vehicles)
RampDem	Ramp demand (veh/hr)
Storage	Mainline storage remaining after downstream flow has departed (vehicles)
t	Index referring to the model time step
UpQ <sup>Main</sup>	Flow that can be contributed by upstream segment and any associated on-ramps, ignoring downstream conditions (veh/hr)
V	Segment speed (km/hr)

$V^{Antic}$	Density-dependent speed (km/hr)
$V^{CellMix}$	Inertial speed as vehicles in the cell mix at the end of the current time step (km/hr)
$V^{Min}$	Minimum segment speed (km/hr)
$V^{Queue}$	Speed in the segment downstream of the queue (km/hr)
$V^{Recovery}$	Recovery speed once congestion has dissipated (km/hr)
$V^{Uncong}$	Average uncongested segment speed in the next time step (km/hr)
Violat	Ramp meter violation rate (%)
XPer	Percentage of mainline traffic exiting the freeway at a given off-ramp

Divide freeway into  $i$  segments and define

Length of each freeway segment **L**  
 Number of lanes in each freeway segment **Lanes**  
 Duration of each model time step **dt**

For each model update interval  $\Delta t$  ← *Corresponds to the data collection interval*

Set model parameters for each segment  $i$  by sampling from the appropriate distribution

Recovery speed once congestion has dissipated (km/hr)  $V^{Recovery}$   
 Speed in the segment downstream of the queue (km/hr)  $V^{Queue}$   
 Minimum segment speed (km/hr)  $V^{Min}$   
 Critical density (veh/km/lane)  $K^{Crit}$   
 Ramp capacity (veh/hr/lane) **Capacity<sup>Ramp</sup>**  
 Parameters for flow breakdown estimation **A, B, C** ← *Dependent on the length of the acceleration lane*

For each time step  $t$  within the model update interval

Set model parameters for each segment  $i$  by sampling from the appropriate distribution

Average density within the queue (veh/km/lane)  $K^{Queue}$   
 Maximum flow out of the queue (veh/hr/lane) **Capacity<sup>Congested</sup>**

Estimate demand entering study area

Flow into Segment 1  $Q_{1,t}^{Main} = f(ObsQ_{1,t}^{Main})$

Estimate on-ramp flows

For each segment  $i$  with an on-ramp

On-ramp Demand **RampDem<sub>i,t</sub> = f(ObsRampDem<sub>i,t</sub>)**

Check that the meter is "on" if a ramp queue exists

If **MeterRate<sub>i,t</sub> = "off"** AND **Queue<sub>i,t-1</sub> > 0**

**MeterRate<sub>i,t</sub> = MaxMeterRate**

End if

If working in tracking mode & on-ramp flow measurements are available

On-ramp Flow  $Q_{i,t}^{On} = f(\text{ObsOnFlow}_{i,t})$

*Note* The on-ramp flow can be adjusted for the proportion of vehicles stored in the speed change lane (not shown)

Else

If no ramp metering

On-ramp Flow  $Q_{i,t}^{On} = \min(\text{RampDem}_{i,t}, \text{Capacity}_{i,t}^{\text{Ramp}})$

Else

On-ramp Flow

$Q_{i,t}^{On} = \min(\text{MeterRate}_{i,t}(1 + \text{Violat}_{i,t}), \text{RampDem}_{i,t} + \text{Queue}_{i,t-1} \frac{3600}{dt})$

End if

End if

End for

Estimate off-ramp exit percentages

For each segment *i* with on-ramp

Exit Percentage  $XPer_{i,t} = f(\text{ObsXPer}_{i,t})$

End for

Estimate flow into each segment ignoring downstream conditions

For each segment *i*

Set speed equal to the uncongested travel speed computed in the previous time step

$V_{i,t} = V_{i,t-1}^{\text{Uncong}}$

Maximum Potential Flow into Segment ← *Must be greater than zero*

$\text{Max}Q_{i,t}^{\text{Main}} = \left( N\text{Veh}_{i-1,t-1} \times V_{i-1,t} \times \frac{1}{L_{i-1}} \right) (1 - XPer_{i,t}) + \eta_{i,t}^{\text{Max}Q^{\text{Main}}}$  ← *Noise term to account for uncertainty in the flow estimation*

Estimate probability of flow breakdown at on-ramp merges

If  $\text{FlowType}_{i,t-1} = 0$  (i.e. flow was not broken down in the previous time step)

OR

If  $\text{FlowType}_{i,t-1} \neq 0$  AND  $\frac{N\text{Veh}_{i-1,t-1}}{L_{i-1} \times \text{Lanes}_{i-1}} \leq \omega K_{i-1}^{\text{Breakdown}}$  (i.e. flow was previously congested, but the density is low enough that the freeway may have recovered)

$$40 \text{ Second Mainline Flow: } 40\text{Sec}Q_{i,t}^{\text{Main}} = \frac{\left( \text{Max}Q_{i,t}^{\text{Main}} + \sum_{k=t-3}^{k=t-1} Q_{i,k}^{\text{Main}} \right) \left( \frac{dt}{3600} \right)}{\text{Lanes}_i}$$

$$40 \text{ Second Ramp Flow: } 40\text{Sec}Q_{i,t}^{\text{On}} = \left( Q_{i,t}^{\text{On}} + \sum_{k=t-3}^{k=t-1} Q_{i,k}^{\text{On}} \right) \frac{dt}{3600}$$

$$U_{i,t} = A_i + B_i \times 40\text{Sec}Q_{i,t}^{\text{On}} + C_i \times 40\text{Sec}Q_{i,t}^{\text{Main}}$$

$$\text{Probability of Flow Breakdown: } \text{Pr}(\text{BD})_{i,t} = \frac{e^{U_{i,t}}}{1 + e^{U_{i,t}}}$$

Else if  $\text{FlowType}_{i,t-1} \neq 0$  AND  $\frac{N\text{Veh}_{i-1,t-1}}{L_{i-1} \times \text{Lanes}_{i-1}} > \omega K_{i-1}^{\text{Breakdown}}$  (i.e. flow was previously congested and the density is too high for the freeway to have recovered)

$$\text{Probability of Flow Breakdown: } \text{Pr}(\text{BD})_{i,t} = 1$$

End if

$z \sim U(0,1)$  (sample  $z$  from the uniform distribution)

If  $z \leq \text{Pr}(\text{BD})_{i,t}$  (i.e. breakdown is predicted)

If  $\text{FlowType}_{i,t-1} = 2$

← If segment was previously part of mainline queue & flow breakdown is predicted, assume flow type has not changed – the validity of this assumption is confirmed below

$\text{FlowType}_{i,t} = 2$

No modification to  $K_{i-1}^{\text{Breakdown}}$

Else

$\text{FlowType}_{i,t} = 1$  (i.e. flow has broken down at the start of a bottleneck)

If  $\text{FlowType}_{i,t-1} = 1$

(i.e. if the flow had previously broken down, there is no need to recalculate  $K^{\text{Breakdown}}$ )

No modification to  $K_{i-1}^{\text{Breakdown}}$

Else

$$K_{i-1}^{\text{Breakdown}} = \frac{N\text{Veh}_{i-1,t-1}}{L_{i-1} \times \text{Lanes}_{i-1}}$$

(i.e. if flow breakdown has only been initiated in the current time step, set the mainline density at breakdown for testing for recovery)

End if

End if

Else

$$\text{FlowType}_{i,t} = 0$$

$$K_{i-1}^{\text{Breakdown}} = 0$$

← Breakdown is not predicted & freeway is assumed to be uncongested

End If

Calculate mainline flow ignoring potential for queue spillback

If  $\text{FlowType}_{i,t} = 0$  (i.e. flow has not broken down)

$$\text{Flow Contributed by Upstream Cell } \text{UpQ}_{i,t}^{\text{Main}} = \text{MaxQ}_{i,t}^{\text{Main}} + Q_{i,t}^{\text{On}}$$

Else (i.e. flow has broken down)

Flow Contributed by Upstream Cell

$$\text{UpQ}_{i,t}^{\text{Main}} = \min(\text{Capacity}_{i,t}^{\text{Congested}} \times \text{Lanes}_i, \text{MaxQ}_{i,t}^{\text{Main}} + Q_{i,t}^{\text{On}})$$

← Note that is also possible to use evidence reversal to calculate the congested capacity directly from sensor measurements (not shown)

Update mainline speed under breakdown

$$V_{i,t} = V_i^{\text{Queue}}$$

$$V_{i-1,t} = \frac{\text{UpQ}_{i,t}^{\text{Main}} - Q_{i,t}^{\text{On}}}{N\text{Veh}_{i-1,t-1} / L_{i-1}}$$

End if

End for

Adjust flow based on downstream conditions & calculate segment density

For each segment  $i$  in reverse order (i.e. downstream to upstream)

$$\text{Available Storage } \text{Storage}_{i,t} = K_{i,t}^{\text{Queue}} \times L_i \times \text{Lanes}_i - N\text{Veh}_{i,t-1} + \frac{dt}{3600} (Q_{i+1,t}^{\text{Main}} + Q_{i+1,t}^{\text{Off}})$$

$$\text{If } \text{Storage}_{i,t} < \frac{dt}{3600} \text{UpQ}_{i,t}^{\text{Main}}$$

(i.e. available storage in downstream cell is less than flow that can be contributed by upstream cell)

**FlowType**<sub>i,t</sub> = 2 (flow is being constrained by downstream queues)

$$\text{Flow into Section: } Q_{i,t}^{\text{Main}} = \frac{3600}{dt} \text{Storage}_{i,t} - Q_{i,t}^{\text{On}}$$

$$\text{Adjusted Mainline Speed: } V_{i-1,t} = \frac{Q_{i,t}^{\text{Main}}}{\text{NVeh}_{i-1,t-1} / L_{i-1}}$$

If  $V_{i-1,t} < V_{i-1}^{\text{Min}}$

← Check that the speed is not less than the minimum cell speed. If it is, adjust the mainline flow

$$V_{i-1,t} = V_{i-1}^{\text{Min}}$$

$$Q_{i,t}^{\text{Main}} = \frac{\text{NVeh}_{i-1,t-1}}{L_{i-1}} \times V_{i-1,t}$$

End if

Else

$$\text{Flow into Section: } Q_{i,t}^{\text{Main}} = \text{Up}Q_{i,t}^{\text{Main}} - Q_{i,t}^{\text{On}}$$

Reset FlowType as mainline queue dissipates and a new bottleneck emerges

If **FlowType**<sub>i,t</sub> = 2 AND **FlowType**<sub>i+1</sub> = 0

**FlowType**<sub>i,t</sub> = 1

Set the density at flow breakdown if it has not previously been set

$$K_{i-1}^{\text{Breakdown}} = K_{i-1}^{\text{Crit}} \times L_{i-1} \times \text{Lanes}_{i-1}$$

End if

End if

$$\text{Number of Vehicles: } \text{NVeh}_{i,t} = \text{NVeh}_{i,t-1} + \frac{dt}{3600} (Q_{i,t}^{\text{Main}} - Q_{i+1,t}^{\text{Main}} + Q_{i,t}^{\text{On}} - Q_{i+1,t}^{\text{Off}})$$

$$\text{Density: } K_{i,t} = \frac{\text{NVeh}_{i,t}}{L_i}$$

$$\text{Off-ramp Flow: } Q_{i,t}^{\text{Off}} = \frac{X\text{Per}_{i,t} \times Q_{i,t}^{\text{Main}}}{1 - X\text{Per}_{i,t}}$$

End for

Calculate ramp queue and delay

For each segment i with an on-ramp

If no ramp metering

Ramp Queue (vehicles):

$$\text{Queue}_{i,t} = \max\left(\text{Queue}_{i,t-1} + \frac{dt}{3600} (\text{RampDem}_{i,t} - \text{Capacity}_{i,t}^{\text{Ramp}}), 0\right)$$

$$\text{Ramp Delay (minutes): } \text{Delay}_{i,t} = \frac{\text{Queue}_{i,t}}{\text{Capacity}_{i,t}^{\text{Ramp}}} * 60$$

Else

Ramp Queue (vehicles):

$$\text{Queue}_{i,t} = \max\left(\text{Queue}_{i,t-1} + \frac{dt}{3600} (\text{RampDem}_{i,t} - Q_{i,t}^{\text{On}}), 0\right)$$

$$\text{Ramp Delay (minutes): } \text{Delay}_{i,t} = \frac{\text{Queue}_{i,t}}{\text{MeterRate}_{i,t}} * 60$$

End if

End for

Update speed estimate for next time step, assuming no congestion

Estimate  $\beta$  by sampling from the appropriate distribution

For each segment  $i$

If upstream and downstream are free-flow (no congestion)

$$\text{Anticipated Density: } K_{i,t+1}^{\text{Antic}} = \frac{N\text{Veh}_{i,t} + N\text{Veh}_{i+1,t}}{(L_i \times \text{Lanes}_i) + (L_{i+1} \times \text{Lanes}_{i+1})}$$

$$\text{Density-Dependent Speed: } V_{i,t+1}^{\text{Antic}} = V_{\text{ff}} + \alpha \times K_{i,t+1}^{\text{Antic}} \quad \leftarrow V_{\text{ff}} \text{ is the free flow speed and } \alpha \text{ is a calibration parameter for the freeway corridor in question}$$

Inertial speed as vehicles within the cell mix:

If  $N\text{Veh}_{i,t} = 0$  (to prevent division by zero)

$$V_{i,t+1}^{\text{CellMix}} = V_{i,t+1}^{\text{Antic}}$$

Else

$$V_{i,t+1}^{\text{CellMix}} = \frac{V_{i-1,t} (Q_{i,t}^{\text{Main}} + Q_{i,t}^{\text{On}}) \left( \frac{dt}{3600} \right) + V_{i,t} (N\text{Veh}_{i,t-1} - (Q_{i+1,t}^{\text{Main}} + Q_{i+1,t}^{\text{Off}}) \left( \frac{dt}{3600} \right))}{N\text{Veh}_{i,t}}$$

End if

Average uncongested speed in the next time step:

$$V_{i,t+1}^{\text{Uncong}} = (\beta \times V_{i,t+1}^{\text{CellMix}} + (1 - \beta) \times V_{i,t+1}^{\text{Antic}}) \times \eta_{li,t+1}^V \quad \leftarrow \text{Noise term to account for uncertainty in the speed estimation}$$

Else if upstream and downstream are congested (in middle of queue)

$$\text{Average uncongested speed in the next time step: } V_{i,t+1}^{\text{Uncong}} = V_i^{\text{Recovery}}$$

Else if downstream is congested but upstream is not (approaching end of queue)

$$\text{Anticipated Density: } K_{i,t+1}^{\text{Antic}} = \frac{N\text{Veh}_{i,t}}{(L_i \times \text{Lanes}_i)}$$

$$\text{Density-Dependent Speed: } V_{i,t+1}^{\text{Antic}} = V_{\text{ff}} + \alpha \times K_{i,t+1}^{\text{Antic}} \quad \leftarrow V_{\text{ff}} \text{ is the free flow speed and } \alpha \text{ is a calibration parameter for the freeway corridor in question}$$

Inertial speed as vehicles within the cell mix:

If  $N\text{Veh}_{i,t} = 0$  (to prevent division by zero)

$$V_{i,t+1}^{\text{CellMix}} = V_{i,t+1}^{\text{Antic}}$$

Else

$$V_{i,t+1}^{\text{CellMix}} = \frac{V_{i-1,t} (Q_{i,t}^{\text{Main}} + Q_{i,t}^{\text{On}}) \left( \frac{dt}{3600} \right) + V_{i,t} (N\text{Veh}_{i,t-1} - (Q_{i+1,t}^{\text{Main}} + Q_{i+1,t}^{\text{Off}}) \left( \frac{dt}{3600} \right))}{N\text{Veh}_{i,t}}$$

End if

Average uncongested speed in the next time step:

$$V_{i,t+1}^{\text{Uncong}} = (\beta \times V_{i,t+1}^{\text{CellMix}} + (1 - \beta) \times V_{i,t+1}^{\text{Antic}}) \times \eta_{li,t+1}^V \quad \leftarrow \text{Noise term to account for uncertainty in the speed estimation}$$

Else if upstream is congested but downstream is not (flow departing queue at start of bottleneck)

$$\text{Average uncongested speed in the next time step: } V_{i,t+1}^{\text{Uncong}} = V_i^{\text{Recovery}}$$

End if

End for

**Note:** Equations must be adjusted to account for boundary conditions

# **APPENDIX H**

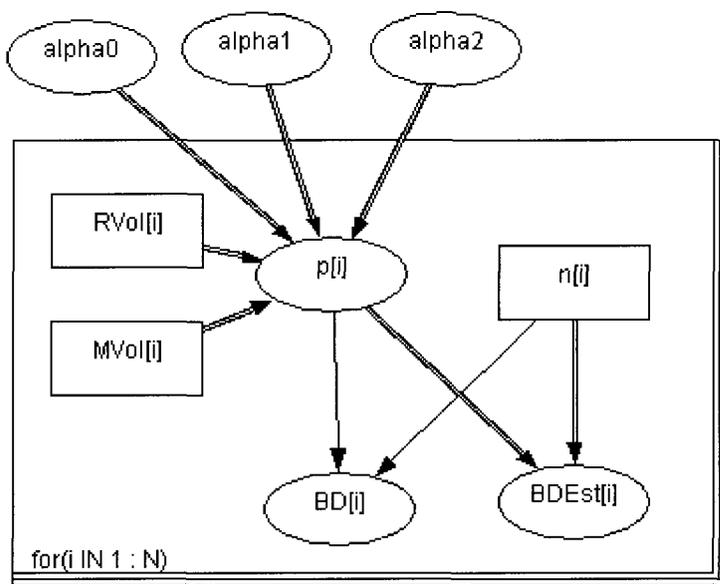
## **DEVELOPMENT OF FLOW BREAKDOWN MODEL**

## WinBUGS Analysis

### Variable Definition:

RVol - Ramp flow (veh/40 sec)  
 MVol - Mainline flow (veh/lane/40 sec)  
 i - Index referring to a particular set of ramp & mainline flows  
 n - Total number of observations for group i  
 p - Probability of breakdown for any given observation  
 BD - Number of breakdown occurrences actually observed  
 BDEst - Number of breakdown occurrences estimated by the model  
 alpha0, alpha1, alpha2 - Model parameters to be determined

### Model Specification:



**Figure H-1 Graphical Specification of Breakdown Model**

```

model;
{
  for( i in 1 : N ) {
    logit(p[i]) <- alpha0 + alpha1 * RVol[i] + alpha2 * MVol[i]
    BD[i] ~ dbin(p[i],n[i])
    BDEst[i] <- p[i] * n[i]
  }
  alpha0 ~ dnorm( 0.0,1.0E-6 )
  alpha1 ~ dnorm( 0.0,1.0E-6 )
  alpha2 ~ dnorm( 0.0,1.0E-6 )
}

```

**Figure H-2 WinBUGS Model Code**

## Final Parameter Estimates

Table H-1 Logit Model Parameters as Estimated Using WinBUGS

Geometric Configuration	Parameter	Parameter Estimation Results						
		Mean	Std. Dev.	MC Error	2.50%	Median	97.50%	Sample
100 m Speed Change Lane	alpha0	-8.742	0.1274	0.004295	-8.983	-8.744	-8.482	75,000
	alpha1	0.3936	0.005751	7.44E-05	0.3823	0.3936	0.4048	75,000
	alpha2	0.2258	0.005064	1.70E-04	0.2155	0.2259	0.2354	75,000
150 m Speed Change Lane	alpha0	-11.27	0.151	0.005377	-11.57	-11.27	-10.98	75,000
	alpha1	0.2648	0.00447	5.20E-05	0.256	0.2648	0.2736	75,000
	alpha2	0.2922	0.005767	2.04E-04	0.281	0.2921	0.3037	75,000
175 m Speed Change Lane	alpha0	-12.03	0.1612	0.005795	-12.35	-12.04	-11.72	75,000
	alpha1	0.2809	0.004512	5.73E-05	0.272	0.2809	0.2897	75,000
	alpha2	0.3098	0.006058	2.17E-04	0.2979	0.3099	0.3217	75,000
200 m Speed Change Lane	alpha0	-12.47	0.1678	0.006058	-12.8	-12.47	-12.13	75,000
	alpha1	0.2986	0.004488	6.12E-05	0.2897	0.2986	0.3074	75,000
	alpha2	0.3143	0.006238	2.24E-04	0.302	0.3143	0.3268	75,000
250 m Speed Change Lane	alpha0	-14.93	0.197	0.007477	-15.34	-14.92	-14.55	75,000
	alpha1	0.3002	0.004553	7.43E-05	0.2913	0.3002	0.3092	75,000
	alpha2	0.3936	0.007092	2.68E-04	0.3801	0.3935	0.4082	75,000
300 m Speed Change Lane	alpha0	-15.29	0.1937	0.007308	-15.69	-15.29	-14.93	75,000
	alpha1	0.2953	0.004352	7.51E-05	0.2868	0.2953	0.3039	75,000
	alpha2	0.4024	0.006894	2.59E-04	0.3894	0.4022	0.4166	75,000
350 m Speed Change Lane	alpha0	-14.6	0.1851	0.00699	-14.99	-14.6	-14.25	75,000
	alpha1	0.2649	0.004062	6.76E-05	0.2569	0.2648	0.2729	75,000
	alpha2	0.3789	0.006614	2.48E-04	0.3665	0.3788	0.3929	75,000

**Note:**

Parameters correspond to the following logit equation

$$\text{logit}(p[i]) = \alpha_0 + \alpha_1 * \text{rvo}[i] + \alpha_2 * \text{mvo}[i]$$

(where all variables are as defined previously)

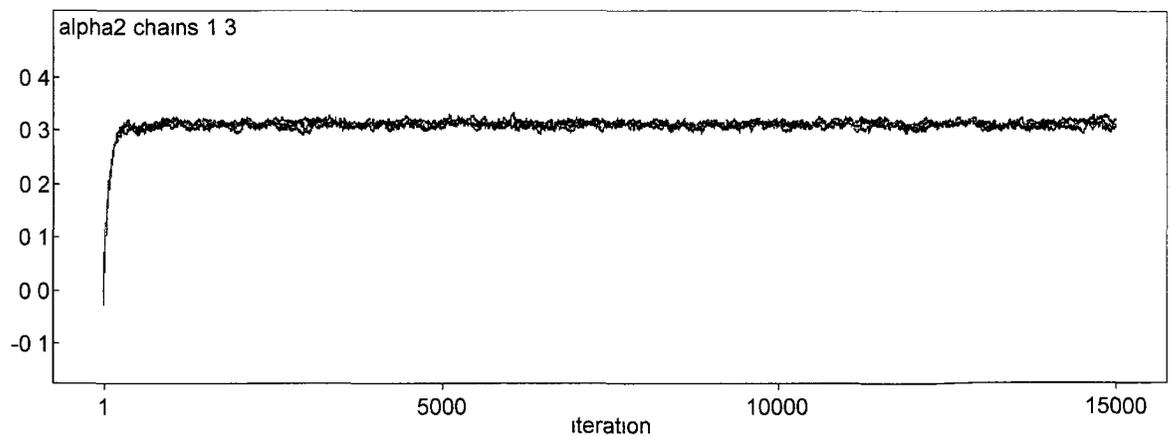
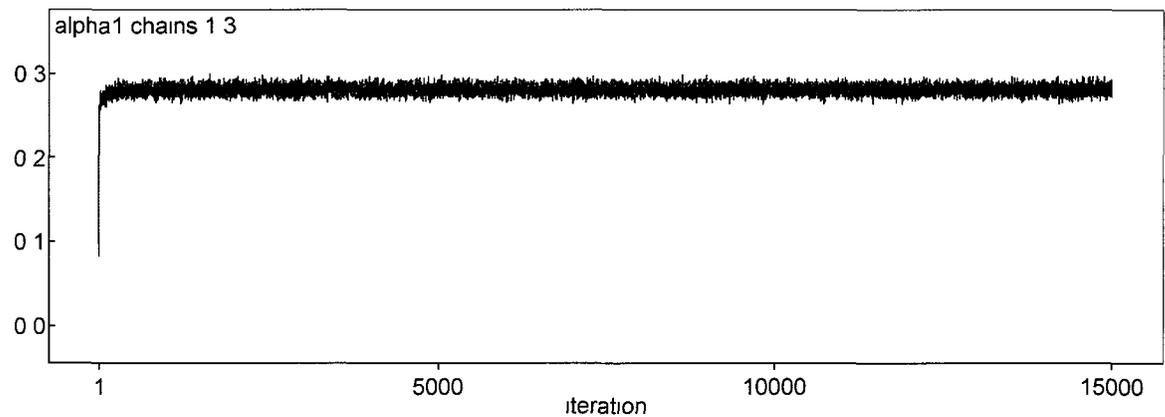
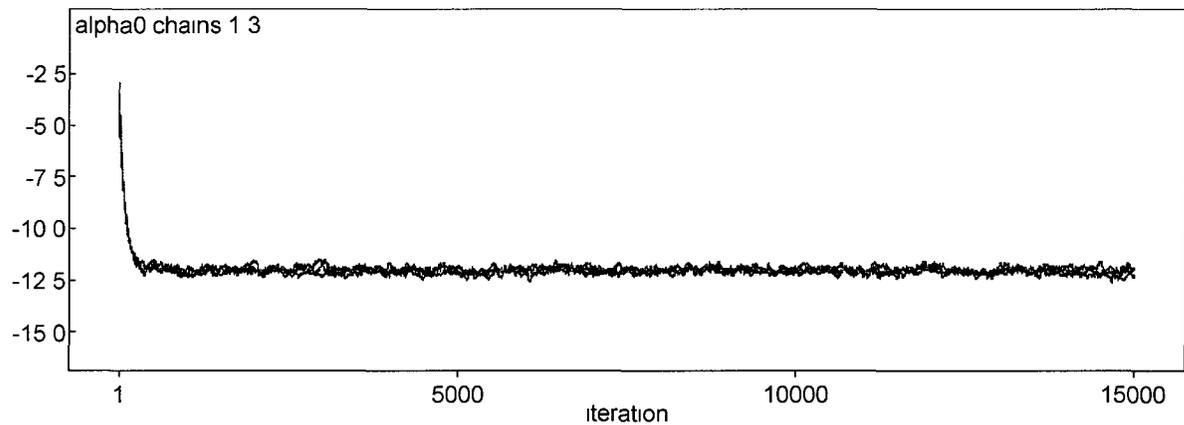
## Typical WinBUGS Outputs: Development of Model for 175 m Speed Change Lane

*Checking for Convergence (3 chains, 15,000 updates):*

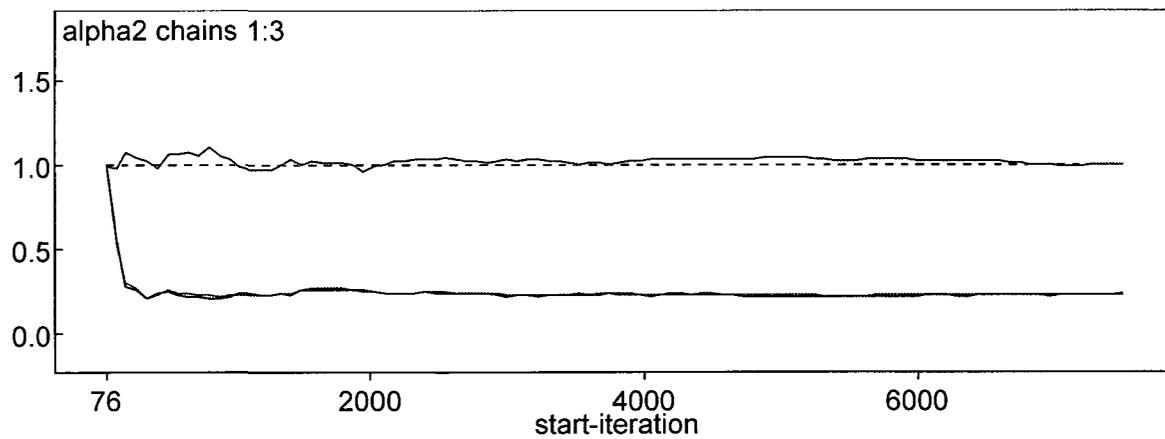
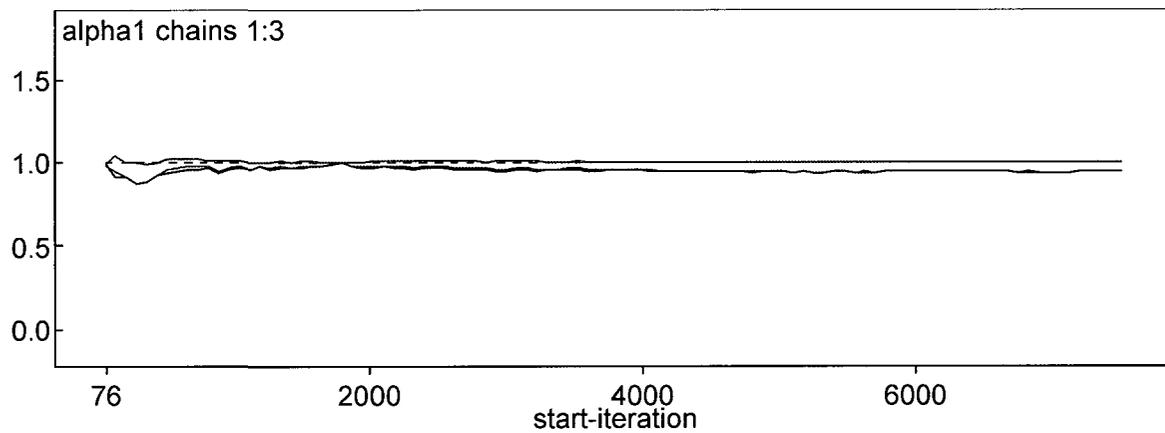
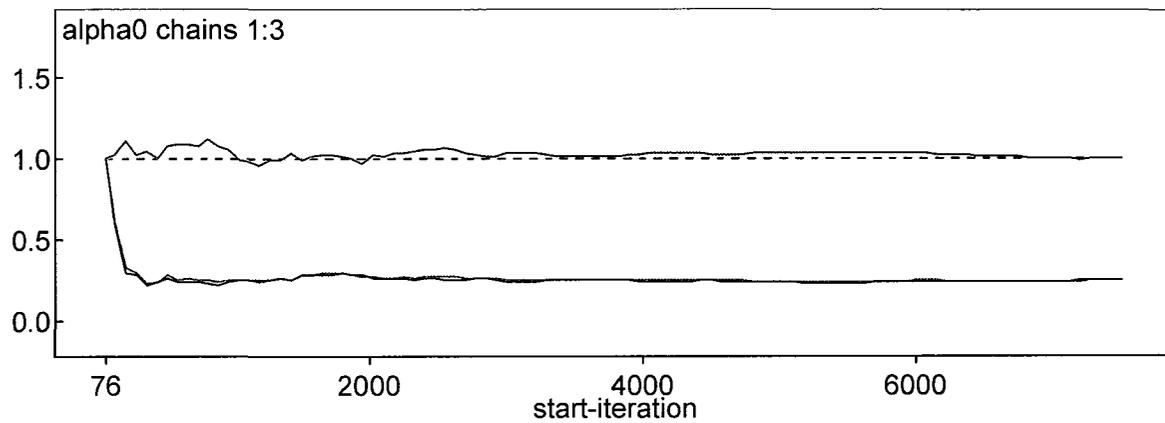
### Node Statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha0	-12.0	0.4617	0.02191	-12.33	-12.04	-11.66	1	45000
alpha1	0.2808	0.005159	9.565E-5	0.2716	0.2809	0.2898	1	45000
alpha2	0.3086	0.01798	8.536E-4	0.2957	0.3102	0.3211	1	45000

### Time Series



## Gelman Rubin Statistic

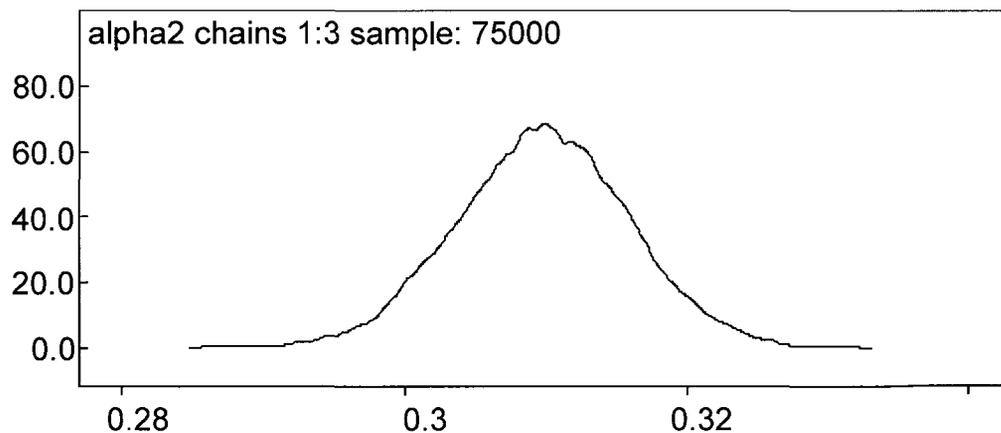
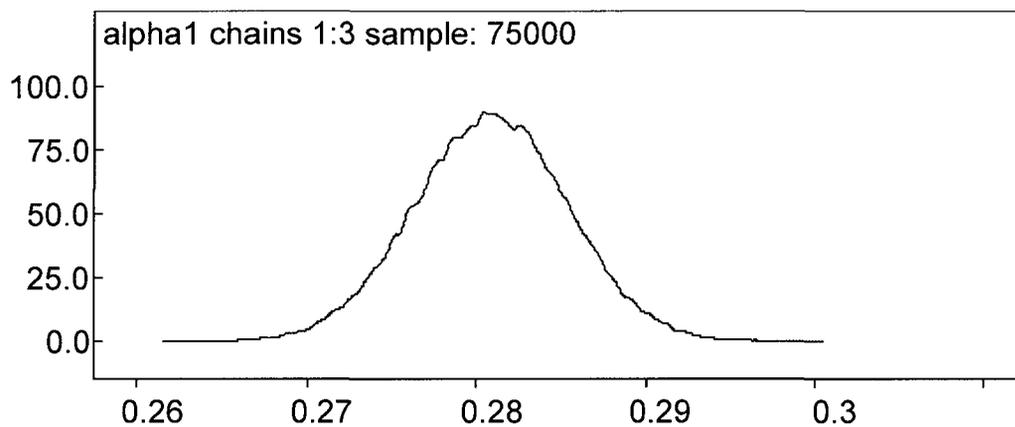
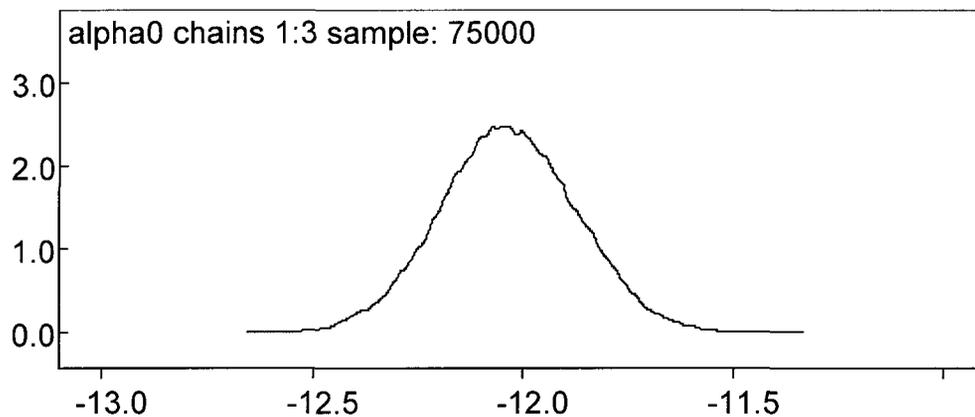


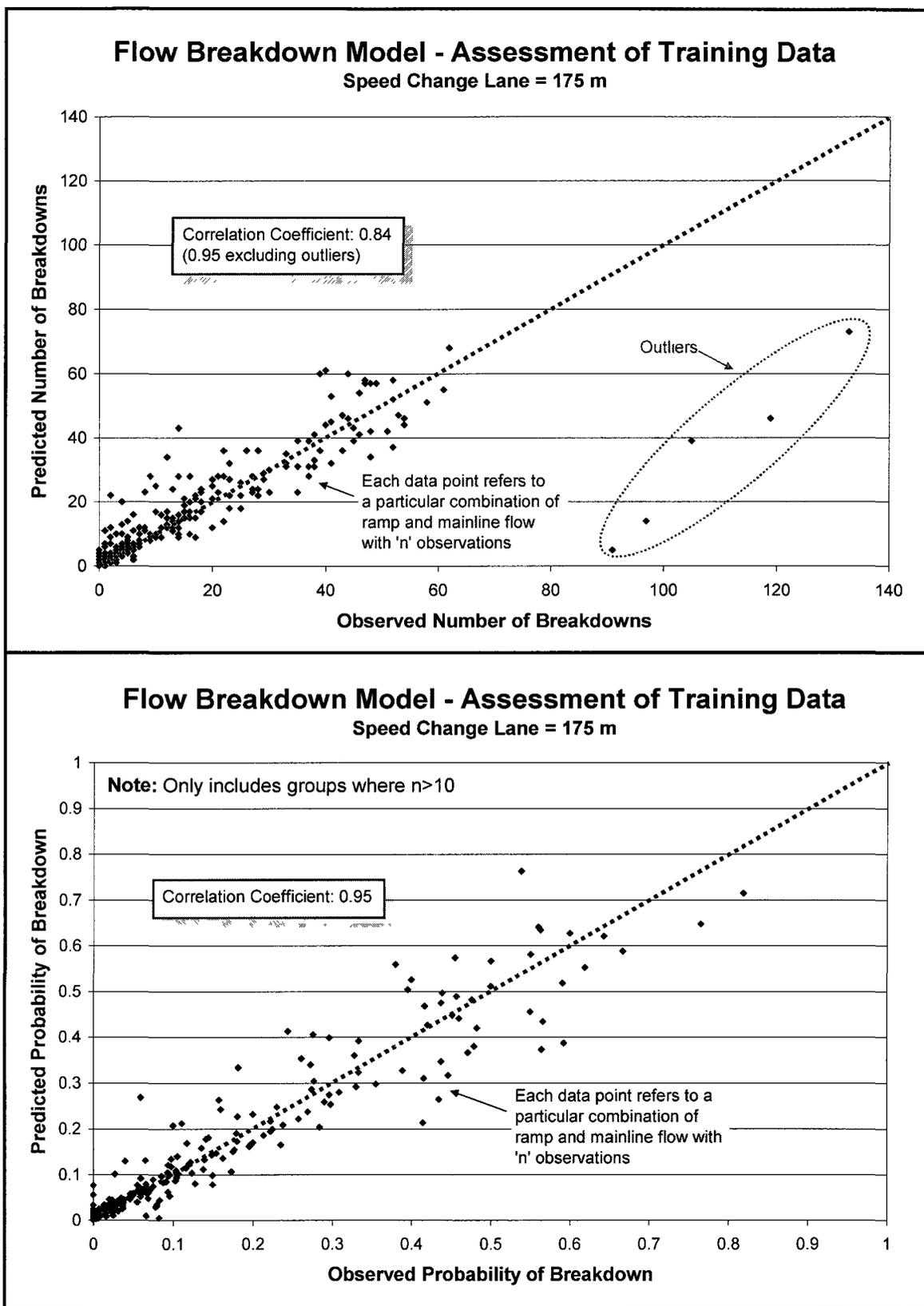
*Sampling from the Posterior Distribution (3 chains, 25,000 additional updates):*

Node Statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha0	-12.03	0.1612	0.005795	-12.35	-12.04	-11.72	15001	75000
alpha1	0.2809	0.004512	5.729E-5	0.272	0.2809	0.2897	15001	75000
alpha2	0.3098	0.006058	2.168E-4	0.2979	0.3099	0.3217	15001	75000
deviance	2407.0	2.437	0.04369	2404.0	2406.0	2413.0	15001	75000

Kernel Density





**Figure H-3 Assessment of Logit Model Performance Using Training Dataset**

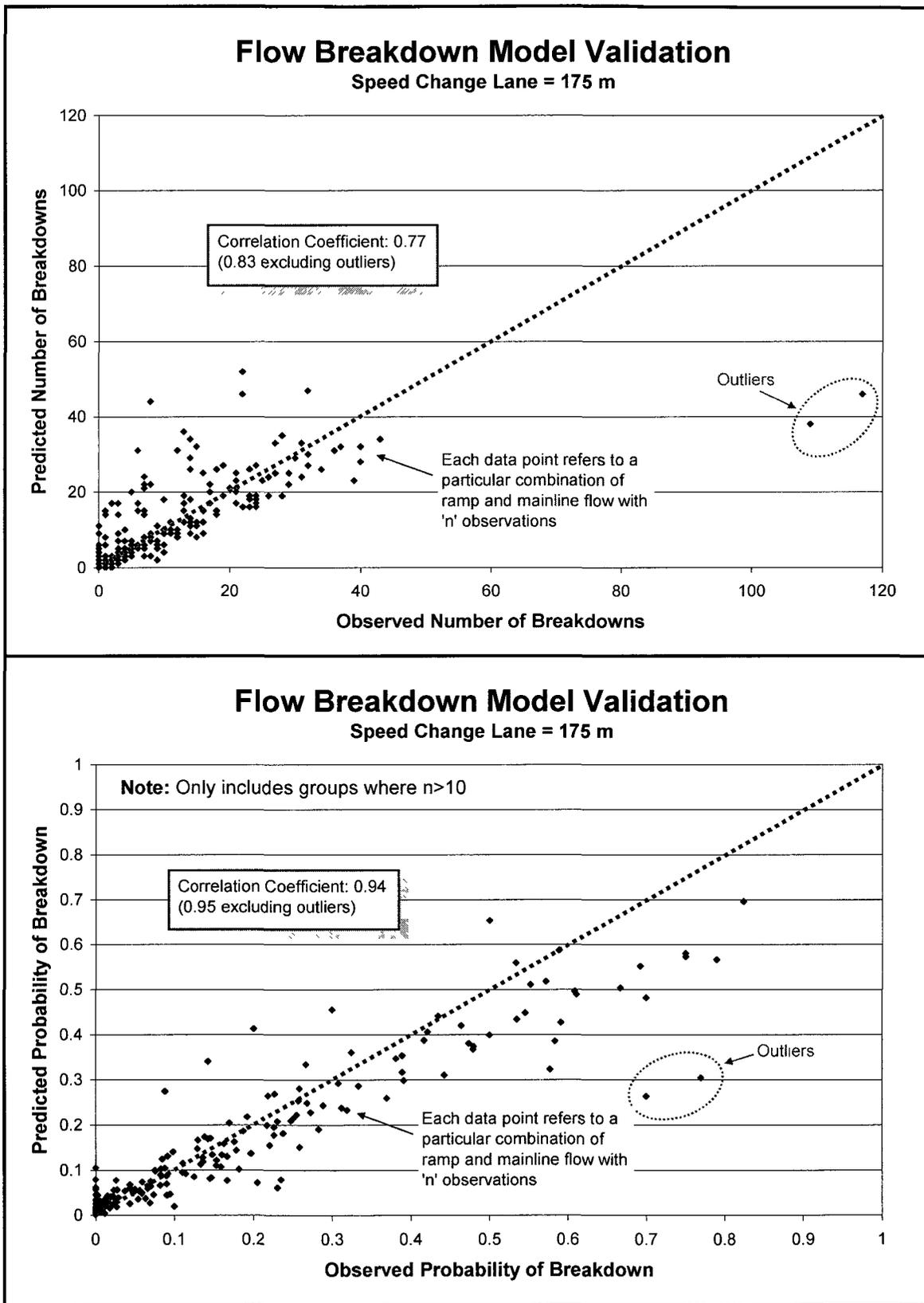


Figure H-4 Assessment of Logit Model Performance Using Validation Dataset

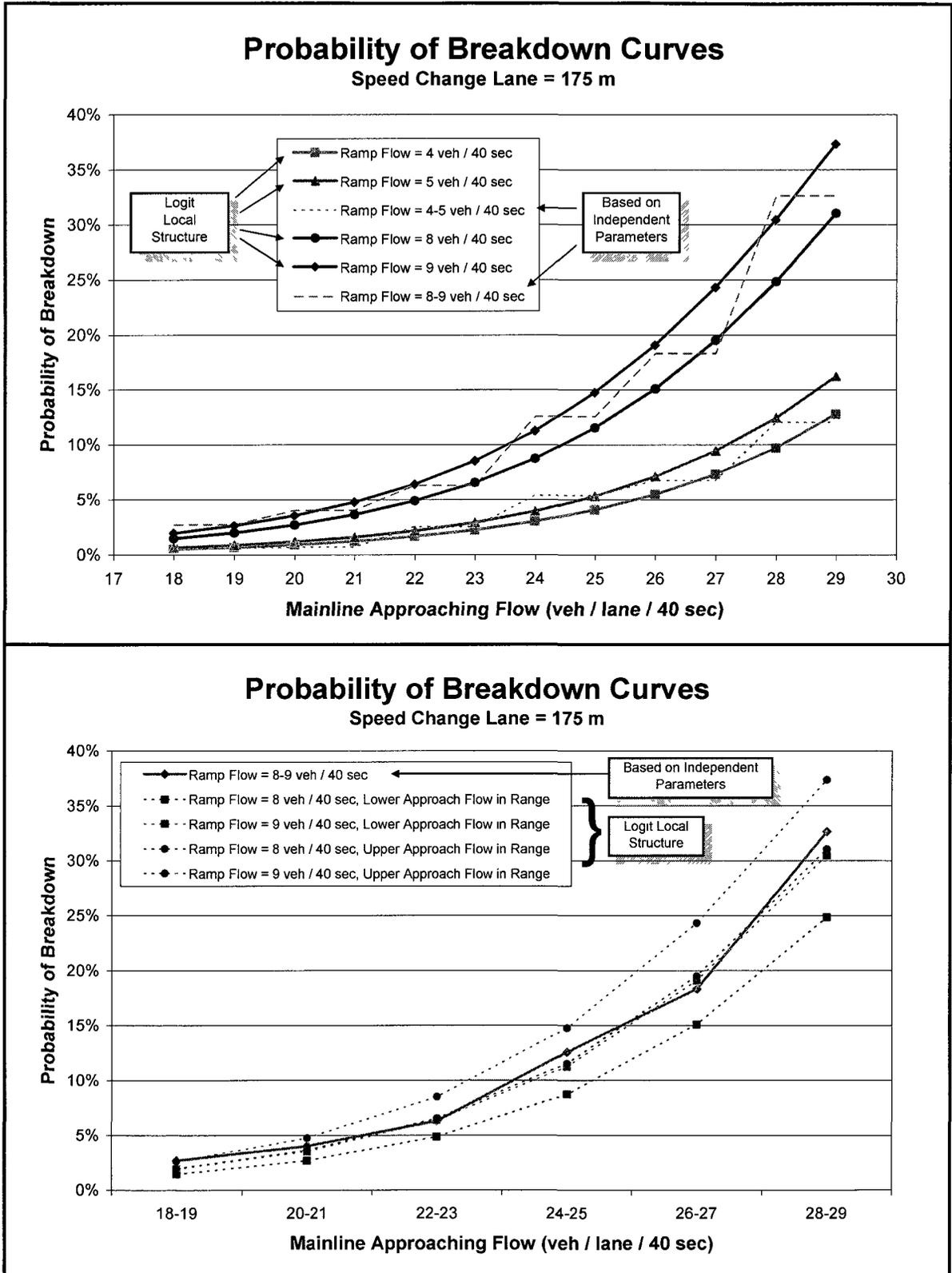
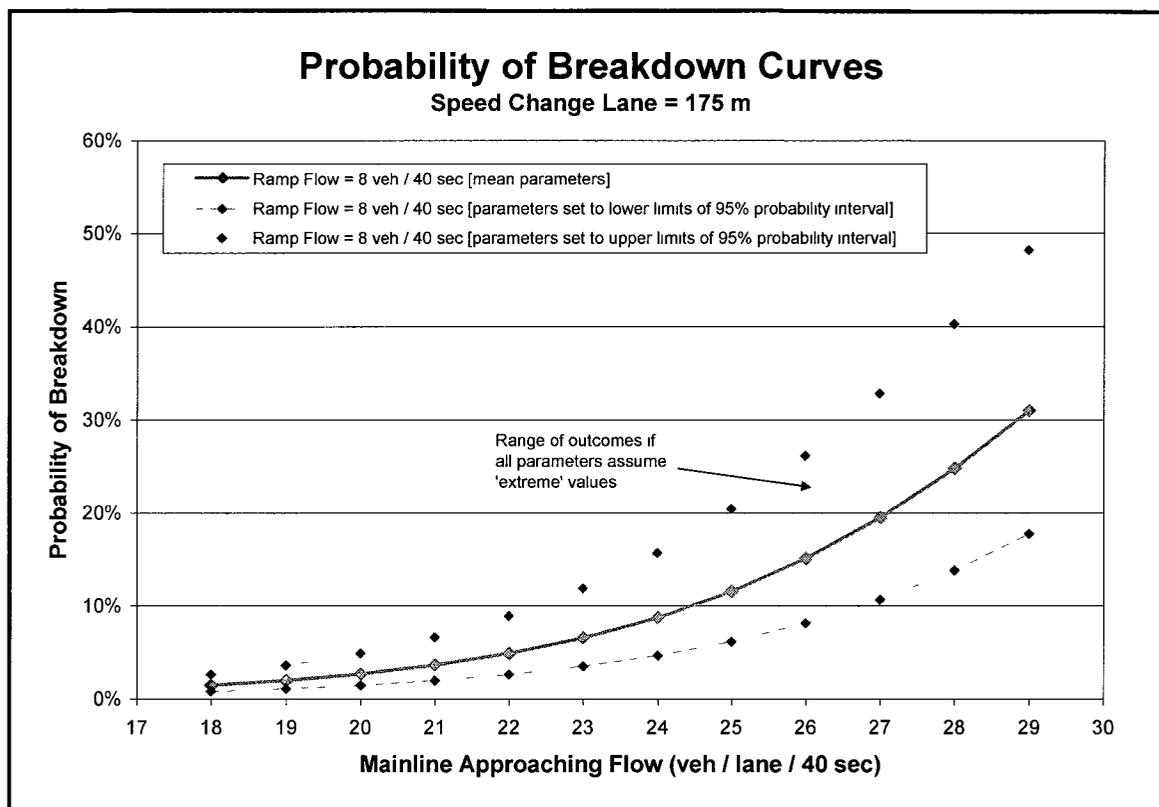


Figure H-5 Comparison of Modelling Approaches



**Figure H-6 Assessment of Logit Model Sensitivity to Parameter Distributions**

**Comparison of WinBUGS Results with Maximum Likelihood Estimation using SAS  
(Speed Change Lane = 175 m)**

*SAS Model Outputs:*

The LOGISTIC Procedure

Model Information

Data Set	PROB_BD.DATA40SEC_R175
Response Variable	Mode
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	83283
Number of Observations Used	83283

Response Profile

Ordered Value	Mode	Total Frequency
1	1	4352
2	0	78931

Probability modeled is Mode='1'.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	1488.7393	368	4.0455	<.0001
Pearson	2847.5188	368	7.7378	<.0001

Number of unique profiles: 371

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	34165.320	27284.897
SC	34174.650	27312.887
-2 Log L	34163.320	27278.897

R-Square	0.0793	Max-rescaled R-Square	0.2358
----------	--------	-----------------------	--------

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6884.4225	2	<.0001
Score	7694.0472	2	<.0001
Wald	5618.1540	2	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-12.0511	0.1628	5476.9001	<.0001		0.000
RVeh	1	0.2809	0.00457	3784.6737	<.0001	0.4541	1.324
AppVeh	1	0.3105	0.00613	2569.3090	<.0001	0.5051	1.364

## Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
RVeh	1.324	1.313	1.336
AppVeh	1.364	1.348	1.381

## Association of Predicted Probabilities and Observed Responses

Percent Concordant	81.8	Somers' D	0.648
Percent Discordant	17.0	Gamma	0.656
Percent Tied Pairs	1.2	Tau-a	0.064
	343507712	c	0.824

## Profile Likelihood Confidence Interval for Parameters

Parameter	Estimate	95% Confidence Limits	
Intercept	-12.0511	-12.3718	-11.7335
RVeh	0.2809	0.2720	0.2899
AppVeh	0.3105	0.2985	0.3226

Parameter	Estimate	95% Confidence Limits	
Intercept	-12.0511	-12.3703	-11.7320
RVeh	0.2809	0.2720	0.2899
AppVeh	0.3105	0.2985	0.3225

## Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
RVeh	1.0000	1.324	1.313	1.336
AppVeh	1.0000	1.364	1.348	1.381

Effect	Unit	Estimate	95% Confidence Limits	
RVeh	1.0000	1.324	1.313	1.336
AppVeh	1.0000	1.364	1.348	1.381

Parameter	Estimated Covariance Matrix		
	Intercept	RVeh	AppVeh
Intercept	0.026517	-0.00024	-0.00097
RVeh	-0.00024	0.000021	3.754E-6
AppVeh	-0.00097	3.754E-6	0.000038

Parameter	Estimated Correlation Matrix		
	Intercept	RVeh	AppVeh
Intercept	1.0000	-0.3190	-0.9764
RVeh	-0.3190	1.0000	0.1342
AppVeh	-0.9764	0.1342	1.0000

## Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.020	4038	36525	42406	314	48.7	92.8	46.3	91.3	0.9
0.100	2368	70508	8423	1984	87.5	54.4	89.3	78.1	2.7
0.200	1408	76322	2609	2944	93.3	32.4	96.7	64.9	3.7
0.300	886	77852	1079	3466	94.5	20.4	98.6	54.9	4.3
0.400	538	78416	515	3814	94.8	12.4	99.3	48.9	4.6
0.500	322	78702	229	4030	94.9	7.4	99.7	41.6	4.9
0.600	186	78845	86	4166	94.9	4.3	99.9	31.6	5.0
0.700	99	78897	34	4253	94.9	2.3	100.0	25.6	5.1
0.800	43	78926	5	4309	94.8	1.0	100.0	10.4	5.2
0.900	10	78931	0	4342	94.8	0.2	100.0	0.0	5.2

# **APPENDIX I**

## **INTRODUCTION TO UTILITY THEORY**

## I. INTRODUCTION TO UTILITY THEORY

### I.1 Overview

This appendix provides a general introduction to utility theory. The discussion draws on material from several references, including Keeney and Raiffa (1993), Edwards et al. (2007), Lindley (1985), Marshall and Oliver (1995) and Russell and Norvig (1995). For a detailed review of utility and its use in decision problems, the reader is referred to the work of Keeney and Raiffa. Edwards et al. contains a compilation of articles describing both the foundations of decision analysis, as well as recent advances. The remaining references provide a good introduction to decision-making under uncertainty, with somewhat less emphasis on theory.

### I.2 Utility Defined

Utility theory provides a basis for making decisions under uncertainty. Where outcomes are subject to a random process, it is assumed that the best course of action is to select the alternative with the highest **expected utility**. Expected utility (**EU**) reflects the probability **P** of each possible outcome **x** associated with alternative **X**, as well as each outcome's corresponding utility **U**:

$$EU(X) = \sum_{x \in X} P(x)U(x) \quad (1)$$

Thus, the expected utility for a given alternative can be calculated by simply multiplying the probability of each potential outcome by the utility of that outcome and summing the results.

Expected utility theory was first formalized by von Neumann and Morgenstern in the 1940's. Von Neumann and Morgenstern asserted that if certain axioms hold, there exists a real-valued function **U**, such that for any two lotteries **P** and **Q**, **P** is preferred to **Q** if and only if:

$$\sum_{x \in X} P(x)U(x) \geq \sum_{x \in X} Q(x)U(x) \quad (2)$$

In this formulation, **U** represents the utility associated with each potential outcome, and is unique up to a positive linear transformation. Given the form of the inequality, **P** is preferred to **Q** if and only if it has the highest expected utility.

An important feature of the von Neumann / Morgenstern theory of expected utility is that preferences are expressed over lotteries and not directly over outcomes. A simple example illustrates the idea:

*Consider two alternatives, each with three possible outcomes, **A**, **B**, and **C**, where **A** is preferred to **B**, which in turn is preferred to **C**.*

- *Alternative 1 has a 50% probability of Outcome A occurring and a 50% probability of Outcome C occurring.*
- *Alternative 2 has a 100% probability of Outcome B occurring.*

*In utility theory, it is not the preference for the individual outcomes that is important, but rather the preference for the lottery (probability distribution) associated with each alternative as reflected in the alternative's expected utility. In this example, Alternative 1 is preferred to Alternative 2 if the lottery associated with Alternative 1 is preferred to the lottery associated with Alternative 2, or on the basis of expected utility, if:*

$$0.5 \times U(A) + 0 \times U(B) + 0.5 \times U(C) > 0 \times U(A) + 1.0 \times U(B) + 0 \times U(C)$$

Different sets of axioms have been put forward which imply the existence of utility functions. For utility theory to apply, it is assumed that decision-makers will act in a rational manner in accordance with these axioms. The following provides a summary of the axioms of utility theory as commonly denoted in the literature.<sup>1</sup>

1. **Orderability:** For every pair of potential outcomes, the decision-maker must either prefer one outcome to the other, or be indifferent between the two:

**A > B** if **A** is preferred to **B**

**B > A** if **B** is preferred to **A**

**A ~ B** if the decision-maker is indifferent between **A** and **B**

2. **Transitivity:** If the decision-maker prefers **A** to **B** and **B** to **C**, the decision-maker must also prefer **A** to **C**.
3. **Continuity:** Assume the decision-maker has the following preference ordering for any three possible outcomes:

**A > B > C**

Given this ordering, there is some probability **p** for which the decision-maker is indifferent between getting **B** for certain, and the lottery which yields **A** with probability **p** and **C** with probability **(1 - p)**.

4. **Substitutability (Independence):** If a decision-maker is indifferent between two lotteries, **A** and **B**, then the decision-maker will also be indifferent between any two more complicated lotteries involving **A** and **B** as long as the only difference between the two lotteries is that **A** has been substituted for **B**.
5. **Monotonicity:** When confronted with two lotteries with the same outcomes, **A** and **B**, the decision-maker will select the lottery which has the highest probability for the preferred outcome.

---

<sup>1</sup> Note that some references omit axioms 5 and 6.

6. **Decomposability:** If a risky venture involving compound lotteries can be simplified using the rules of probability, a decision-maker should be indifferent between the more complex form of the lottery and the simplified form.

To determine the expected utility associated with each alternative, a utility function, **U**, must be defined which assigns a real number to every possible outcome in such a way that reflects the decision-maker's preferences. For two outcomes **A** and **B**:

- **U(A) > U(B)** if and only if **A** is preferred to **B**
- **U(A) = U(B)** if and only if the decision-maker is indifferent between **A** and **B**

In general, utility is measured on a scale of 0 to 1, with the worst possible outcome given a utility of 0, and the best possible outcome given a utility of 1. Utilities are therefore dependent on the range of outcomes considered in the decision problem.

To develop the utility function for a particular attribute, indifference methods are often used. These methods rely on two concepts which follow from the axiom of continuity: indifference probabilities and certainty equivalents.

Assume the decision-maker is faced with two alternatives, each with three possible outcomes, **A**, **B**, and **C**, where **A > B > C**. The first alternative is considered a risky venture, with probability **p** that the best outcome (**A**) occurs, and probability **(1 - p)** that the worst outcome (**C**) occurs. The second alternative has no risk involved; if this alternative is selected, Outcome **B** is obtained with 100% certainty. The **indifference probability** is the value of **p** that would make the decision-maker indifferent between the two alternatives. The **certainty equivalent** is the smallest value that the decision maker would have to obtain for certain to be indifferent to the risky venture if the probability **p** of achieving the favourable outcome is known.

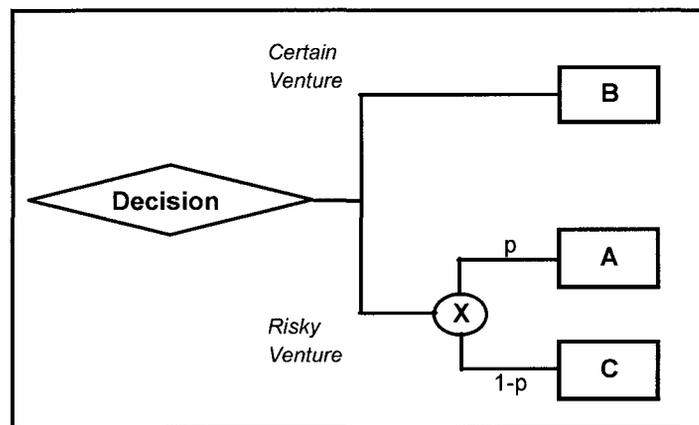
Both indifference probabilities and certainty equivalents can be used to determine the utility of a particular outcome. Consider the decision problem shown in Figure I-1. In this problem, the decision-maker can choose either the certain venture in which Outcome **B** is guaranteed, or the risky venture, which will yield Outcome **A** with probability **p** and Outcome **C** with probability **(1 - p)**. If Outcome **A** is the best outcome from the set of all possible results, and Outcome **C** is the worst, then set the utility of Outcome **A** to one, and the utility of Outcome **C** to zero. From the theory of expected utility, if the decision-maker is indifferent between the two ventures:

$$EU(\text{Certain Venture}) = 1.0 \times U(B) = EU(\text{Risky Venture}) = p \times U(A) + (1 - p) \times U(C)$$

Since **U(A) = 1** and **U(C) = 0**, the above equation simplifies to: **U(B) = p**. Thus, the utility of a particular outcome equals the indifference probability for the choice between obtaining that outcome with 100% certainty, and a risky venture involving the best and worst outcomes.

In practice, utility functions can be derived using either indifference probabilities or certainty equivalents. To determine a point on the utility curve, the decision-maker is

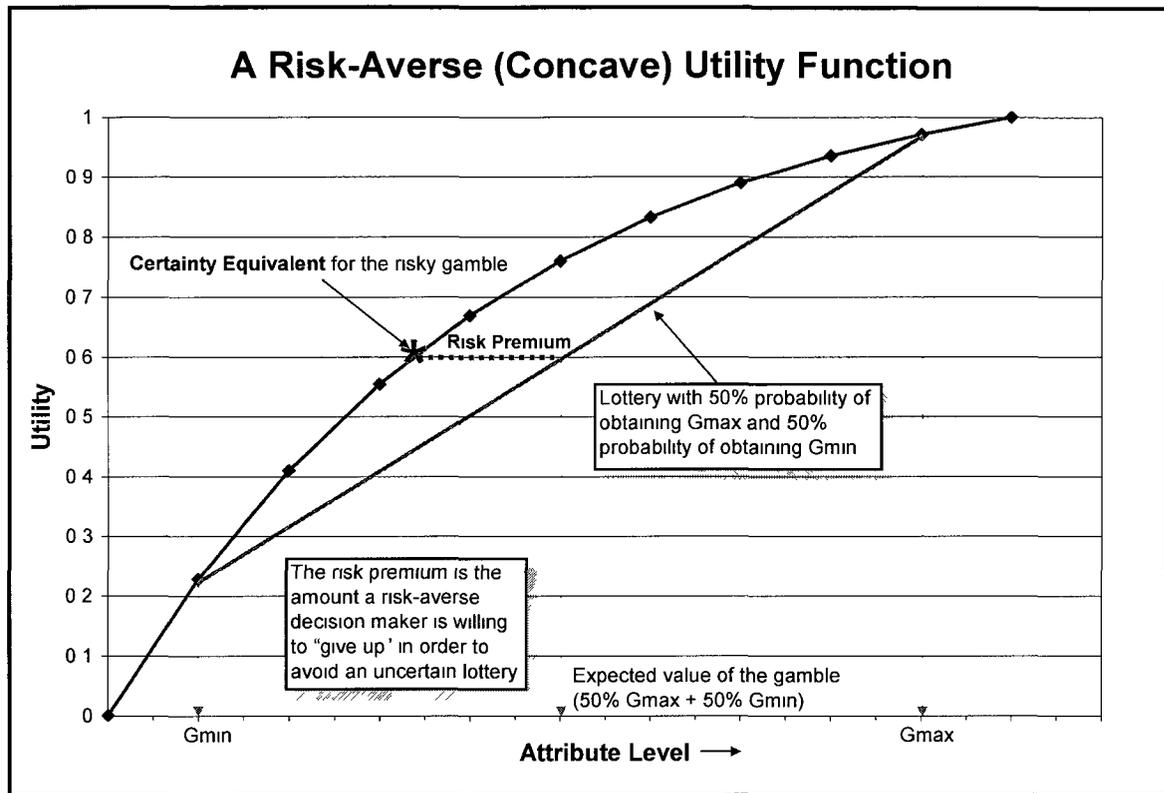
presented with a decision problem of the type described above, and must either specify a value for  $p$  (the indifference probability), given a pre-defined value of  $B$ , or alternatively, specify a value for  $B$  (the certainty equivalent), given a pre-defined value of  $p$  (usually 50%). The process typically involves asking a series of questions involving different risky ventures, with the outcomes used to define the risky venture being modified as needed. Initially, the best and worst outcomes are used, since the utility for these outcomes is known (i.e. 1 and 0, respectively). However, as information on the utility of intermediate outcomes is obtained, lotteries can be defined for different intervals. The process is repeated until enough utilities are obtained to determine an appropriate function to represent the data. Since decision-makers may provide responses which conflict, particular care is needed to ensure any inconsistencies in the derived utility values are resolved.



**Figure I-1 A Simple Decision Tree**

In general, the shape of the utility function describes the decision maker's attitude towards risk. For decision-makers who are risk-averse, the utility curve takes a concave shape, while a convex curve describes someone who is risk-prone. Formally, a decision-maker is **risk-averse** if their certainty equivalent for a given risky gamble is less than the expected value of the gamble. In other words, to avoid the risk associated with the uncertain venture, the decision-maker is prepared to accept a lower return which is guaranteed. In contrast, someone who is **risk-prone** will tend to accept the risky gamble, and will only choose the risk-free alternative if its value is some magnitude greater than the expected payoff of the gamble.

The **risk premium** is defined as the difference between the expected value of the gamble and decision-maker's certainty equivalent. When the risk premium is positive, the decision-maker is risk-averse; when the risk premium is negative, the decision-maker is risk-prone. The concept is illustrated in Figure I-2 for the risk-averse case.



**Figure I-2 Utility Functions & Risk Aversion**

It has been found that many utility functions have the same general shape, and can be expressed using common functional forms. For example, the utility function for someone who is risk-averse is often modelled using an exponential or logarithmic relationship. Such relationships are characterized by decreasing marginal utility; as outcomes improve, the slope of the utility function decreases, giving the curve its concave shape. Since the slope is steepest near outcomes with low utility, a given improvement will have a much greater impact on utility when outcomes are relatively poor, and a much smaller impact as outcomes approach optimal.

The utility curve in Figure I-2 is based on the exponential form of the utility function presented in Marshall and Oliver (1995), which is defined in terms of a unique parameter  $\beta$ . In the Marshall and Oliver formulation,  $\beta$  represents the relative rate of change in the utility function near the best and worst outcomes as measured by the ratio of the slopes at the two extremes (i.e. the slope at the worst outcome divided by the slope at the best outcome). As  $\beta$  increases, the curve becomes increasingly concave, indicating a greater tendency to avoid risk.

Other commonly used exponential utility functions incorporate a term which reflects the level of risk aversion. In this context, risk aversion is defined as the negative ratio of the second derivative of the utility function to the first derivative. If the utility function is of the form  $U(x) = a + be^{-cx}$ , the risk aversion is constant, and equal to the value of parameter  $c$ . This form of the utility function is used in the decision software Logical

Decisions (2009). The software ASSESS uses a similar formulation, but instead of risk aversion, defines a parameter called the risk tolerance, which is simply the inverse of the risk aversion (Delquié 2008).

The property of constant risk aversion is common to all exponential utility functions. For a given risky venture involving a potential gain of  $x$  units, and a potential loss of  $y$  units, the decision-maker exhibits the same aversion to risk at every point on the utility curve. The logarithmic utility function, in contrast, models decreasing risk aversion; the more of something someone has, the more risks they are willing to take.

Since each decision-maker is unique, it follows that, for a given attribute, each decision-maker will have a unique utility function which captures the decision-maker's attitude towards risk. Such attitudes are encoded in the shape of the utility curve; the more concave the shape, the greater the aversion to risk when confronted with a given lottery at a given attribute level. The shape of the curve also reflects how the decision-maker reacts to risk as attribute levels change.

If it is assumed that the utility function for a given attribute has a certain functional form, the process of estimating the function can often be simplified. In theory, it is possible to determine individual points on the utility curve by posing a series of hypothetical questions aimed at soliciting the decision-maker's indifference probability or certainty equivalent for various risky gambles. In practice, such an approach is not without challenges: decision-makers may have difficulty relating to the hypothetical questions; inconsistencies may arise in the decision-maker's responses; there may be time constraints which limit how much information the decision-maker can provide. By assuming that the utility function has a certain mathematical form, it may be possible to use that knowledge to inform the interview process, so that fewer data points are needed to fit the curve, and any inconsistencies in the data are more easily resolved.<sup>2</sup>

### I.3 Multi-Attribute Utility Functions

One of the main strengths of utility theory is the ability to compare alternatives on a multi-criteria basis. In real-world problems, different courses of action produce different outcomes. If each outcome is defined in terms of multiple attributes, the preferred outcome may not be readily apparent, particularly if certain attributes interact or conflict. The theory of expected utility provides a mechanism for selecting the best option, however to calculate the expected utility, a multi-attribute utility function is needed which captures the relative preference for each potential outcome. Such a function is ideally expressed as a weighted combination of the utilities of the individual attributes.

The multiplicative form of the multi-attribute utility function is provided below, where  $U(\mathbf{x})$  is the multi-attribute utility for the attributes contained in the vector  $\mathbf{x}$ , and  $u_j(x_j)$  is the single attribute utility for attribute  $x_j$ :

---

<sup>2</sup> If the utility function is assumed to have an exponential or logarithmic form as defined by Marshall and Oliver (1995), it may be possible to completely define the function by simply estimating the value of  $\beta$ . This requires examining the decision-maker's preferences near the extremes of the utility curve.

$$U(\mathbf{x}) = \frac{\prod_{j=1}^n [1 + K k_j u_j(\mathbf{x}_j)] - 1}{K} \quad (3)$$

As with the single attribute case, the multi-attribute utility is scaled between 0 and 1. Setting  $U(\mathbf{x}) = 0$  for the worst outcome (so that all  $u_j(\mathbf{x}_j) = 0$ ), and  $U(\mathbf{x}) = 1$  for the best outcome (so that all  $u_j(\mathbf{x}_j) = 1$ ), it can be shown that  $K$  must satisfy:

$$1 + K = \prod_{j=1}^n (1 + K k_j) \quad (4)$$

For the multiplicative form of the multi-attribute utility function to apply, the attributes must satisfy the property of mutual utility independence. This concept, and the related concept of mutual preferential independence, is defined below:

- **Mutual Preferential Independence (MPI)** – Consider two attributes,  $\mathbf{j}$  and  $\mathbf{k}$ . If the preference structure for  $\mathbf{j}$  is independent of the value of  $\mathbf{k}$ ,  $\mathbf{j}$  is said to be preferentially independent of  $\mathbf{k}$ . For example, if  $\mathbf{j}_1$  is preferred to  $\mathbf{j}_2$ , this must be true for every possible value of  $\mathbf{k}$  for preferential independence to hold. If  $\mathbf{j}$  is preferentially independent of  $\mathbf{k}$  and  $\mathbf{k}$  is preferentially independent of  $\mathbf{j}$ , then  $\mathbf{j}$  and  $\mathbf{k}$  are said to be mutually preferentially independent. The concept extends to subsets of attributes. In general, a set of attributes has the MPI property if every subset and its complement are MPI.
- **Mutual Utility Independence (MUI)** – An attribute  $\mathbf{j}$  is considered to be utility independent of an attribute  $\mathbf{k}$  if preferences between lotteries on  $\mathbf{j}$  are independent of the value of  $\mathbf{k}$ . Utility independence also applies to sets of attributes: A set of attributes  $\mathbf{J}$  is utility independent of a set of attributes  $\mathbf{K}$  if preferences between lotteries involving the attributes in  $\mathbf{J}$  remain the same regardless of the specific values assumed by the attributes in  $\mathbf{K}$ . For mutual utility independence to hold, each subset of attributes must be utility independent of its complement.

In general, utility independence is a much stronger assumption than preferential independence. To confirm that the MUI condition holds for the set of attributes under consideration, it would normally be necessary to check each possible subset of attributes to ensure the independence requirements between the subset and its complement are met. With  $n$  attributes, such an undertaking would require  $2^n - 2$  checks. Fortunately, the level of effort can be reduced substantially by carrying out the following simplified process.

1. Find an attribute  $\mathbf{j}$  that is utility independent of its complement.
2. Pair attribute  $\mathbf{j}$  with each of the remaining attributes, creating  $n - 1$  subsets.
3. Verify that each pair of attributes is preferentially independent of its complement.

These steps (provided without proof) are sufficient to confirm that a set of attributes has the property of mutual utility independence. In total, only  $n$  checks are required, the majority involving the much simpler check for preferential independence.

As noted earlier, the multiplicative form of the multi-attribute utility function is valid if the attributes exhibit mutual utility independence. Although the multi-attribute utility function can be written in various forms, in general, it is comprised of  $n$  single-attribute utilities, and  $n$  constants (denoted by  $\mathbf{k}$  in Equation 3).<sup>3</sup> Since the single-attribute utilities can be developed independent of each other using the procedures discussed previously, the main challenge remaining is to determine appropriate values for the constants such that the resulting function captures the overall preferences of the decision-maker.

One method used to develop the value of the constants is to set all the attributes to their worst value, except for some arbitrary attribute  $\mathbf{j}$ , which is set to its best value. If the vector for these attribute values is denoted by  $\dot{\mathbf{x}}_j$ , then from Equation 3, we have:

$$U(\dot{\mathbf{x}}_j) = \mathbf{k}_j \quad (5)$$

Similar to the single-attribute case,  $U(\dot{\mathbf{x}}_j)$  represents the indifference probability of the choice between obtaining  $\dot{\mathbf{x}}_j$  for certain, and a lottery involving the best and worst possible outcomes. Once this indifference probability has been identified, it can be set equal to  $\mathbf{k}_j$  in accordance with Equation 5, thus providing a means to specify each of the constants of the multi-attribute utility function. The constant  $\mathbf{K}$  in Equation 3 is simply the root of Equation 4. If the sum of the  $\mathbf{k}_j$ 's is greater than one, it can be shown that  $\mathbf{K}$  lies in the interval  $(-1, 0)$ . If the sum of the  $\mathbf{k}_j$ 's is less than one,  $\mathbf{K}$  lies in the interval  $(0, \infty)$ . A special case occurs if the sum of the  $\mathbf{k}_j$ 's is equal to one. In this situation,  $\mathbf{K}$  is zero, and Equation 3 simplifies to:

$$U(\mathbf{x}) = \sum_{j=1}^n \mathbf{k}_j u_j(\mathbf{x}) \quad (6)$$

For additive utility to apply, the attributes must be additive independent, that is, preferences over lotteries must depend only on the marginal probability distribution of the attributes, and not their joint probability distribution.<sup>4</sup> According to Keeney and von Winterfeldt, if the objectives used to develop the utility model meet the criteria for fundamental objectives (i.e. objectives which define the basic reasons for being interested in the decision – the *ends* that are trying to be achieved rather than the *means*), then a strong case can be made that an additive model is appropriate.<sup>5</sup>

<sup>3</sup> Note that the constant  $\mathbf{K}$  in Equation 3 can be disregarded since it is uniquely defined as a function of the other constants (refer to Equation 4).

<sup>4</sup> In practice, additive independence can be assessed by examining two lotteries,  $\mathbf{A}$  and  $\mathbf{B}$ , involving two arbitrary attributes,  $\mathbf{X}$  and  $\mathbf{Y}$ , with levels  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , and  $\mathbf{y}_1$  and  $\mathbf{y}_2$  respectively (assuming all other attributes are fixed). In lottery  $\mathbf{A}$ , there is a 50-50 chance of obtaining  $\mathbf{x}_1, \mathbf{y}_1$ , and  $\mathbf{x}_2, \mathbf{y}_2$ . In lottery  $\mathbf{B}$ , there is a 50-50 chance of obtaining  $\mathbf{x}_1, \mathbf{y}_2$ , and  $\mathbf{x}_2, \mathbf{y}_1$ . If the decision-maker is indifferent between these two lotteries, then  $\mathbf{X}$  and  $\mathbf{Y}$  are additive independent.

<sup>5</sup> See Chapter 13, "Practical Value Models", in Edwards et al. (2007).

Where additive utility does not apply, interaction must be considered. In implementing the multiplicative form of the multi-attribute utility function, the software Logical Decisions (2009) distinguishes between constructive and destructive interaction, which are dependent on the value of  $K$  in the multiplicative equation. If  $K$  is greater than zero, the individual attributes interact destructively. In this situation, a low utility for any one attribute tends to result in a low utility overall. As  $K$  approaches infinity, the multi-attribute utility is simply the product of the individual utilities, and will therefore equal zero if any of the individual utilities are zero. In contrast, when  $K$  is less than zero, the attributes interact constructively; a high utility for one attribute tends to result in a high utility overall. In the extreme case where  $K$  equals -1, the multi-attribute utility will equal one if any of the individual utilities are one.

# **APPENDIX J**

## **ASSESSMENT OF UTILITY FOR RAMP CONTROL**

**Table J.1 Assessment of Potential Measures for Evaluating Freeway Performance in Relation to Ramp Metering Objectives**

MEASURE	ISSUES / COMMENTS
<b>OBJECTIVE: Minimize freeway congestion</b>	
Occurrence of flow breakdown	<ul style="list-style-type: none"> <li>• Works well if ramp metering is sufficient to prevent flow breakdown</li> <li>• Once flow breaks down, there is no way to distinguish between different levels of queuing and delay</li> <li>• Will need to aggregate the utilities for each freeway section to account for conditions along the entire corridor</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Freeway Level of Service <ul style="list-style-type: none"> <li>• Defined using the density criteria in the Highway Capacity Manual (TRB 2000)</li> </ul>	<ul style="list-style-type: none"> <li>• Little value in distinguishing between any LOS greater than “E” given the stated objective (presumably, any LOS greater than “E” is considered acceptable)</li> <li>• Is not well-suited to describing traffic operations once flow has broken down (LOS F covers a wide range of failure conditions)</li> <li>• Threshold defining LOS “F” may not correspond to the onset of congestion due to the probabilistic nature of flow breakdown</li> <li>• Will need to aggregate the utilities for each freeway section to account for conditions along the entire corridor</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Mainline density	<ul style="list-style-type: none"> <li>• The average density of a freeway segment may be misleading if it is only partially congested (i.e. it includes a transition from free flow to congested flow)</li> <li>• More difficult for drivers to relate to (very few drivers would be able to estimate the traffic density corresponding to a particular set of operating conditions)</li> <li>• Will need to aggregate the utilities for each freeway section to account for conditions along the entire corridor</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Number of freeway segments with mainline queues	<ul style="list-style-type: none"> <li>• Requires that freeway segments be similar in size</li> <li>• Reasonably easy for people to understand</li> <li>• Easy to compute</li> <li>• No need to aggregate results for different freeway sections</li> <li>• The length of the queue does not necessarily provide an indication of how long drivers will be delayed (particularly under non-recurrent congestion where queue discharge rates may vary substantially depending on the number of lanes affected)</li> </ul> <p><b>CONCLUSION: Do not use</b></p>

MEASURE	ISSUES / COMMENTS
Total time spent in the freeway system over the evaluation interval	<ul style="list-style-type: none"> <li>• Often used in the objective function for optimizing ramp metering performance</li> <li>• Would require ramp delay to be included in the travel time total, resulting in a possible overlap of objectives</li> <li>• Less meaningful to drivers</li> <li>• May be difficult to assign a utility value to different travel time outcomes (maximum value is not known a priori)</li> <li>• No need to aggregate results for different freeway sections</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Number of drivers exiting the freeway system over the evaluation interval	<ul style="list-style-type: none"> <li>• Has also been used in algorithms to optimize ramp metering performance</li> <li>• Less meaningful to drivers</li> <li>• May be difficult to assign a utility value to different outcomes (maximum value is not known a priori)</li> <li>• No need to aggregate results for different freeway sections</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Mainline delay (as measured between the start and end of the metered corridor)	<ul style="list-style-type: none"> <li>• Easier for people to understand (relates directly to how drivers typically perceive freeway performance)</li> <li>• Not estimated directly by the freeway traffic model <ul style="list-style-type: none"> <li>– Since delay is calculated as the difference between the uncongested travel time and the actual travel time, two sets of travel time calculations are needed</li> <li>– Requires knowledge of freeway speed which can be difficult to estimate accurately</li> </ul> </li> <li>• No need to aggregate results for different freeway sections</li> <li>• Ideally, travel time (and delay) would be computed by averaging the results for individual drivers. However, this is not possible using macroscopic models. Instead, it is recommended that the total travel time be estimated as the sum of the time required to travel each freeway cell based on the speed in each cell at the end of the evaluation interval. It is important to note that these speeds are not necessarily reflective of what a given driver will experience: <ul style="list-style-type: none"> <li>– Some drivers will exit the freeway before or mid-way through the congestion</li> <li>– The speed in each freeway cell is calculated as a function of the vehicles passing through the cell near the end of the evaluation interval. However, this speed is not necessarily the same as the speed experienced near the start or middle of the evaluation interval, nor is it necessarily the same as the speed experienced in subsequent intervals. As a result, by the time a driver reaches a particular cell, the speed may have changed, particularly under congested conditions</li> <li>– The more time required to travel the entire freeway corridor, the less likely that the estimated travel time will be reflective of actual vehicle</li> </ul> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<p>trips completed near the end of the evaluation interval. Not only is the travel speed more likely to change before the driver reaches the end of the corridor, but the number of drivers traveling the full length of the corridor within the evaluation interval is likely to be smaller, resulting in speed estimates for each cell that are based on an increasingly different subset of trips</p> <p>Since the calculated travel time (delay) may not reflect the experience of individual drivers, this measure should be considered a reflection of system performance</p> <ul style="list-style-type: none"> <li>• Will account for delay at all active bottlenecks in the system</li> <li>• May be difficult to structure the utility function to capture the risk/importance of flow breakdown</li> <li>• May be difficult to estimate the maximum potential delay for a given corridor</li> <li>• May be more appropriate to use travel time in the formulation of the utility function, rather than delay (see discussion below)</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
<p>Mainline travel time (as measured between the start and end of the metered corridor)</p>	<ul style="list-style-type: none"> <li>• Very similar to ‘Mainline delay’ (see comments above)</li> <li>• Somewhat easier to estimate than delay (no need to determine the travel time under free-flow conditions). However, without an appropriate benchmark, the travel time has little meaning, and is difficult to relate to actual operating conditions (congested or free-flow)</li> <li>• May be difficult to estimate the maximum potential travel time for a given corridor</li> <li>• Provides an indication of system performance, and is not necessarily representative of what an individual driver would experience</li> <li>• May be more appropriate to use travel time in the formulation of the utility function, rather than delay. Typically, utility is measured in terms of total assets (i.e. wealth), whereas delay essentially represents a change in assets. When delay is used to assess utility, preferences (and attitudes towards risk) are assumed to be independent of the trip duration <ul style="list-style-type: none"> <li>– In general, it is unclear how trip length may influence people’s preferences for varying levels of delay – it could be argued that a given savings in travel time provides a similar level of benefit regardless of the trip duration, and that people should make similar choices under a wide range of conditions</li> <li>– On the other hand, people may respond to delay differently depending on the situation. For long trips, a five minute increase in travel time may not be noticeable; for short trips, five minutes of delay could be substantial, influencing travel choices. Some people may be more willing to take risks when the trip is short since the overall impact of even the worst outcome is likely to be low. For others, there may be a hesitancy to select gambles that would make a long trip even longer. Given the uncertainty involved in how people value travel options, it seems prudent to provide context by framing the problem in terms of travel time, rather than delay</li> </ul> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<ul style="list-style-type: none"> <li>- Regardless of whether travel time or delay is used to elicit preferences, it is anticipated that people may be influenced by the size of the gamble. When the choices involve relatively small differences in travel time (delay), people may be more likely to gamble on the outcome. However, as the gamble becomes larger, there may be a greater tendency to avoid risk</li> <li>• When developing a utility function based on travel time, the travel time options should be defined in terms of a specific corridor, with any differences in travel time attributed to traffic congestion, not the use of alternative routes. Such an approach ensures that the range of travel times used in the utility function is reasonable, and provides a context for expressing preferences. As a result, the utility function for a given corridor is not necessarily transferable to other locations</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
<p>Travel time between Origin-Destination (OD) pairs (assuming freeway ramps represent trip origins and destinations)</p>	<ul style="list-style-type: none"> <li>• Travel time fits well into the utility framework (see discussion above), however, it may be difficult to establish an appropriate upper limit for travel time corresponding to zero utility, since this depends on the maximum possible travel time between the various OD pairs under consideration</li> <li>• More difficult to calculate (not a direct output of the freeway traffic model)</li> <li>• Requires knowledge of freeway speed which can be difficult to estimate accurately</li> <li>• Better captures drivers' individual experiences <ul style="list-style-type: none"> <li>- Measures such as mainline delay are based on conditions over the entire freeway corridor which a given driver may or may not experience depending on the trip origin and destination</li> <li>- However, the travel time between OD pairs is based on the speed observed in each freeway cell at the end of the evaluation interval. These speeds may not necessarily be reflective of what an actual vehicle will experience due to the time lag involved in traveling between different cells. Indeed, depending on the physical separation of the origin and destination ramp and the level of traffic congestion, there may be only a few vehicles who complete the full trip between the two ramps within the evaluation interval</li> </ul> </li> <li>• Better captures how drivers evaluate freeway performance (i.e. in terms of the specific trips they make). Measures that capture system-wide performance tend to be more difficult to relate to, making the formulation of a utility function potentially more challenging</li> <li>• Flow breakdown is considered indirectly. As a result, it may be difficult to structure the utility function to reflect the importance of preventing the onset of congestion</li> <li>• Will result in a separate utility for each OD pair, which must then be aggregated together using a multi-attribute utility function. <ul style="list-style-type: none"> <li>- Ideally, would weight the individual utilities by the demand between each OD pair, but this is typically unknown</li> </ul> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<ul style="list-style-type: none"> <li>- Given the need to aggregate results, it may be preferable to assess equity on an OD level as well</li> <li>- Note that combining the travel times and then computing the resultant utility is also an option, but this would have an entirely different meaning (utility of the (weighted) average travel time between OD pairs). Since utility functions are typically non-linear, this approach would result in a different value for utility.</li> <li>• Potential to include ramp delay in the travel time estimate, which would essentially combine the objectives for minimizing freeway congestion and minimizing ramp delay. Note that doing so would obscure differences in how people perceive delay on different facility types unless ramp and mainline travel times are somehow weighted differently</li> <li>• Provides no real advantage to the 'Mainline travel time' measure <ul style="list-style-type: none"> <li>- More complicated to calculate</li> <li>- Since the travel time utilities are not weighted by demand, this measure will essentially give more weight to freeway sections that are used by multiple OD pairs, which is not necessarily that meaningful.</li> <li>- Better captures drivers' actual experiences, but not perfectly (see discussion above)</li> </ul> </li> <li>• Fits well with how drivers evaluate freeway performance (in terms of individual trips). However, measures applied at a system level also provide a suitable basis for estimating utility as long as the individuals involved have some knowledge of freeway operations</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Average mainline speed	<ul style="list-style-type: none"> <li>• Requires knowledge of freeway speed which can be difficult to estimate accurately</li> <li>• For cells that are congested, the speed can vary substantially due to shock waves. The speed that is reported in the Freeway Traffic Model refers to the average cell speed</li> <li>• Different ways to calculate average speed: <ol style="list-style-type: none"> <li>a) Compute a simple average of the speed in each cell</li> <li>b) Compute a weighted average speed by weighting the speed in each cell by the cell length</li> <li>c) Compute a weighted average speed by weighting the speed in each cell by the time required to travel each cell. This is equivalent to dividing the length of the corridor by the time required to travel from one end of the corridor to the other. The resulting value can thus be interpreted as the average speed for a driver to travel the entire corridor at a given instant in time, if conditions do not change</li> <li>d) Weight the speed in each cell by the number of vehicles in the cell, or alternatively, by the vehicle-kilometers of travel (which also accounts for the cell length). When the speed is weighted by the number of vehicles, the resulting value represents the average speed being experienced by all</li> </ol> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<p>drivers in the network at the end of the evaluation interval</p> <p>Of the various ways to aggregate speed, Options C and D are preferred, since they tend to give more weight to sections that are experiencing traffic congestion</p> <ul style="list-style-type: none"> <li>• People tend to relate well to measures that involve speed, however, they may not understand that it is an average travel speed, and that the actual speed in different sections of the freeway will be higher or lower. In general, it is anticipated that people may prefer working with travel time to evaluate freeway performance – most people know what a five-minute increase in travel time feels like, but may not be able to relate to the change in operating conditions corresponding to a 5 km/hr decrease in speed. For this reason, it may be preferable to elicit preferences using travel time and then convert to speed</li> <li>• Provides a measure of system-wide performance; what drivers actually experience as they travel the corridor may be significantly different <ul style="list-style-type: none"> <li>– Since there is only one number to work with (the average speed), it is easy to compare freeway performance under different scenarios</li> <li>– However, the average speed is not necessarily an intuitive value to work with, and is highly dependent on the corridor length / number of cells. Moreover, an average speed may correspond to different operating conditions (i.e. an average speed of 70 km/hr could mean that every freeway section has speeds of 70 km/hr, or it could mean that one section is congested and the rest are not)</li> </ul> </li> <li>• Captures conditions at one instant in time</li> <li>• Unlike travel time (or delay), the minimum and maximum limits for this measure are easy to define, and do not depend on the corridor characteristics or anticipated level of congestion. Values will range from 0 km/hr to the free-flow speed, which is generally well-defined. In comparison, travel time can increase indefinitely depending on the situation</li> <li>• Since the utility function is based on speed, it may be possible to transfer the results to other locations</li> </ul> <p><b>CONCLUSION: Consider for inclusion in the utility function</b></p>
Speed within each freeway cell	<ul style="list-style-type: none"> <li>• Under this option, the utility is computed for each freeway cell based on the estimated speed at the end of the evaluation interval</li> <li>• Need to aggregate the utility values to provide a measure of performance for the entire corridor <ul style="list-style-type: none"> <li>– May be appropriate to weight the utilities by the number of vehicles in each freeway cell. Such an approach will tend to give more weight to cells with traffic congestion (and queues)</li> <li>– Could also weight the utilities by the vehicle-kilometers of travel (VKT) in each cell, which would have the added advantage of taking the cell length into account. The simplest way to calculate the VKT is to take the number of vehicles in the cell at the end of the evaluation interval and multiply by the length of the cell. However, this approach assumes that all drivers included in the VKT calculation will travel the full length of</li> </ul> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<p>the cell at the estimated cell speed, which is not strictly correct</p> <ul style="list-style-type: none"> <li>• Since utility functions are generally non-linear, the utility under this outcome will not be the same as the utility based on the average mainline speed (the latter involves aggregating the speed data and then computing a utility value, whereas the former involves computing utilities and then carrying out the aggregation)</li> <li>• Reflects the speed that drivers are actually experiencing in each freeway cell, and may therefore provide a better basis for calculating utility <ul style="list-style-type: none"> <li>– Tends to be easier for people to understand. Once the speed data has been aggregated for the entire corridor, it reflects average conditions, which tend to be more difficult to relate to</li> <li>– On the other hand, it is often useful to have a single measure of freeway performance, rather than trying to understand what is happening based on data for individual freeway segments. This is particularly true when comparing scenarios. Obviously, once the corresponding utility has been calculated, this is no longer an issue, however, it is often difficult to relate the utility to what is actually happening on the freeway</li> </ul> </li> <li>• Requires knowledge of freeway speed which can be difficult to estimate accurately</li> <li>• For cells that are congested, the speed can vary substantially due to shock waves. The speed that is reported in the Freeway Traffic Model refers to the average cell speed</li> <li>• Unlike travel time (or delay), the minimum and maximum limits for this measure are easy to define, and do not depend on the corridor characteristics or anticipated level of congestion. Values will range from 0 km/hr to the free-flow speed, which is generally well-defined. In comparison, travel time can increase indefinitely depending on the situation</li> <li>• Does not depend on the length of the corridor or the travel time under free-flow conditions. As a result, there is no need to frame options in terms of a specific corridor when eliciting preferences. The resulting utility function can be applied to any location</li> </ul> <p><b>CONCLUSION: Consider for inclusion in the utility function</b></p>
<b>OBJECTIVE: Minimize ramp delay</b>	
Ramp delay	<ul style="list-style-type: none"> <li>• Relatively easy to calculate from the freeway traffic model</li> <li>• Assumes that the only delay on the ramp is due to ramp metering</li> <li>• Could be somehow combined with mainline delay / travel time, but this would obscure differences in how people perceive delay on different facility types (more appropriate to combine ramp and mainline delay at an OD level)</li> <li>• The actual ramp delay experienced by a driver will depend in part on when the driver enters the queue. In general, the ramp delay at the end of the evaluation interval is considered to be an appropriate measure for assessing preferences</li> </ul> <p><b>CONCLUSION: Consider for inclusion in the utility function</b></p>

MEASURE	ISSUES / COMMENTS
Ramp queue length	<ul style="list-style-type: none"> <li>• Relatively easy to calculate from the freeway traffic model</li> <li>• Provides an indirect measure of the delay, but cannot be used to compare operating conditions at different ramps, due to differences in demand and metering rates</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
<b>OBJECTIVE: Minimize impacts to the arterial road network</b>	
Available ramp storage length	<ul style="list-style-type: none"> <li>• Could be included in the multi-attribute utility function, or could be used to over-ride the proposed ramp metering rates if queue lengths become excessive</li> <li>• If a queue over-ride feature is employed: <ul style="list-style-type: none"> <li>– Would need to set appropriate thresholds for what queue lengths are considered acceptable. If the queue length is exceeded, the ramp metering scenario is discarded <ul style="list-style-type: none"> <li>▪ Have the option of either determining an acceptable set of ramp metering rates for each ramp which will meet the queue storage length requirements and then performing the optimization, or performing the optimization on the full set of all possible ramp metering rates and imposing appropriate constraints. The former approach provides an opportunity to limit the size of the optimization problem, which may be beneficial from an algorithm implementation perspective (with a smaller solution space, the optimal set of ramp metering rates can be found faster, improving performance for real-time applications)</li> </ul> </li> <li>– If no ramp metering options result in an acceptable queue length, the maximum allowable ramp metering rate should be employed</li> </ul> </li> <li>• If ramp storage is included in the multi-attribute utility function: <ul style="list-style-type: none"> <li>– There is no mechanism to prevent queue spillback onto the arterial network (arterial impacts are traded off against other objectives in accordance with the preference structure encoded in the multi-attributed utility function)</li> <li>– Would need a method for dealing with negative storage (indicating that the storage length has been exceeded)</li> <li>– Utility function would likely show a high degree of risk-aversion</li> <li>– May result in smoother transitions in the ramp metering rate as the available queue storage approaches zero</li> <li>– May not be meaningful to have the utility vary between 0 and 1. From an operational perspective, it really doesn't matter if there are 2 or 12 spaces of ramp storage remaining; as long as there is no spillback, the storage is acceptable, as soon as there is spillback, the storage is not acceptable.</li> </ul> </li> </ul> <p><b>CONCLUSION: Consider for inclusion in the utility function</b></p>
Adequacy of ramp storage	<ul style="list-style-type: none"> <li>• Qualitative measure with two possible values: <ul style="list-style-type: none"> <li>– Yes (if adequate storage is still available)</li> <li>– No (if the ramp queue exceeds the available storage length)</li> </ul> </li> </ul>

MEASURE	ISSUES / COMMENTS
	<ul style="list-style-type: none"> <li>• No opportunity to respond as the queue length begins to approach critical levels, potentially resulting in sudden changes to the ramp metering rate (and less efficient performance)</li> <li>• Provides no real advantage over more quantitative measures, since it still requires calculation of the ramp queue length</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
<b>OBJECTIVE: Minimize Inequity</b>	
Ramp delay	<ul style="list-style-type: none"> <li>• Easy to calculate</li> <li>• Does not consider how equity may be influenced by mainline conditions</li> <li>• Fits well with how drivers perceive equity at ramp meters</li> <li>• Requires aggregation to be meaningful. Possible options include the maximum ramp delay, or the standard deviation of the ramp delay</li> </ul> <p><b>CONCLUSION: Consider for inclusion in the utility function</b></p>
Travel speed/time/delay between OD pairs	<ul style="list-style-type: none"> <li>• Much more complicated to calculate</li> <li>• Provides a better indication of true equity, but may not reflect how people view equity (i.e. strictly in terms of ramp delay)</li> <li>• Most meaningful if weighted by demand, but this information is typically unavailable in real-time</li> <li>• Could be incorporated into a Gini co-efficient as a measure of inequity</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Delay at each ramp relative to the maximum ramp delay	<ul style="list-style-type: none"> <li>• Easy to calculate</li> <li>• Does not consider how equity may be influenced by mainline conditions</li> <li>• Fits well with how drivers perceive equity at ramp meters</li> <li>• Less clear how this could be aggregated. Would likely require focusing on the minimum ratio, but this is essentially the equity index described below</li> </ul> <p><b>CONCLUSION: Do not use</b></p>
Equity index as defined by Meng and Khoo (2010)	<ul style="list-style-type: none"> <li>• Calculated as the ratio of the minimum and maximum average ramp delay for a given group of ramps</li> <li>• Focuses on extremes (doesn't consider improvements at ramps with intermediate levels of delay)</li> <li>• Average ramp delay may be difficult to calculate since the vehicles stopped at the ramp meter may experience ramp delay outside the evaluation interval. An alternative would be to calculate the ratio based on the maximum delay on each ramp at the end of the evaluation interval</li> <li>• Aggregation is taken care of implicitly</li> <li>• May be difficult to relate the index to a utility value</li> </ul> <p><b>CONCLUSION: Do not use</b></p>

Table J.2 Recommended Measures &amp; Proposed Method of Aggregation

Objective	Recommended Measure	Proposed Method of Aggregation	Issues / Comments
Minimize freeway congestion	Average mainline speed, weighted by the cell volume	<ul style="list-style-type: none"> <li>Not necessary due to the way the measure is defined</li> </ul>	<ul style="list-style-type: none"> <li>N/A</li> </ul>
Minimize ramp delay	Ramp delay	<ul style="list-style-type: none"> <li>Maximum ramp delay observed over all ramps within study area</li> <li>Assume all ramp delays correspond to the last vehicle in the queue at the end of the evaluation interval</li> </ul>	<ul style="list-style-type: none"> <li>This will tend to reduce the maximum ramp delay, which is more of an equity consideration</li> <li>For this measure, we are more interested in minimizing the total ramp delay (or the average ramp delay per vehicle)</li> <li><b>Conclusion: Do not use</b></li> </ul>
		<ul style="list-style-type: none"> <li>Average ramp delay experienced at each ramp within the study area, as calculated above</li> </ul>	<ul style="list-style-type: none"> <li>Has the advantage of including data for all ramps in the utility calculation</li> <li>Ramps with higher delays per vehicle do not necessarily incur the greatest total delay, since the ramp demand also comes into play</li> <li>As a result, this measure is more appropriate for measuring equity than efficiency, since it provides no indication of the total amount of ramp delay being incurred in the network</li> <li>Calculation of a weighted average would provide a better measure of efficiency, however, the ramp demand used to develop the weights would not necessarily correspond to the vehicles actually experiencing the delay.</li> <li><b>Conclusion: Do not use</b></li> </ul>
		<ul style="list-style-type: none"> <li>Total ramp delay experienced within the study area over the evaluation interval</li> </ul>	<ul style="list-style-type: none"> <li>Provides a good measure of efficiency</li> <li>Relatively easy to calculate by multiplying the ramp queue at each time step by the duration of the time step</li> <li>To compute a utility function, the maximum expected delay must be known in advance corresponding to a utility value of zero. In practice, this may be difficult to estimate</li> <li>Even if a maximum threshold can be developed, total delay is an abstract</li> </ul>

Objective	Recommended Measure	Proposed Method of Aggregation	Issues / Comments
			<p>concept, and it may be difficult to establish a meaningful utility function</p> <ul style="list-style-type: none"> <li>• Rather than working with the total ramp delay, another option is to calculate the average ramp delay per vehicle by dividing the total delay by the ramp demand. The average ramp delay tends to be easier to relate to, and maximum limits are easier to establish. However, there is no way to calculate a value for the ramp demand which perfectly corresponds to the delay incurred over the evaluation interval (some of the vehicles using the ramp will incur delay prior to the evaluation interval, and some will incur delay after the evaluation interval finishes)</li> <li>• <b>Conclusion: Do not use</b></li> </ul>
		<ul style="list-style-type: none"> <li>• Average ramp delay experienced by any vehicle stopped at the ramp meter during the evaluation interval</li> </ul>	<ul style="list-style-type: none"> <li>• This approach is identical to the one described above, but ensures that all delay incurred by vehicles using the ramp during the evaluation interval is properly accounted for</li> <li>• Includes any delay experienced prior to the evaluation interval by vehicles already queued at the meter, and any delay experienced after the evaluation interval while the queue dissipates</li> <li>• To calculate the delay incurred to clear the remaining queue, an assumption is needed for the ramp metering rate. Although it is reasonable to assume that the prevailing ramp metering rate will apply into the future, this may not be the case in reality</li> <li>• Provides a delay value which is easy to understand and relate to. It is also relatively straightforward to define <i>minimum and maximum</i> delay values for computing the utility</li> <li>• The calculation of total delay is somewhat more complex, but can be readily accomplished using outputs from the Freeway Traffic Model</li> <li>• <b>Conclusion: Carry forward</b></li> </ul>

Objective	Recommended Measure	Proposed Method of Aggregation	Issues / Comments
Minimize arterial impacts	Occurrence of queue spillback	<ul style="list-style-type: none"> <li>Total number of vehicles in the network who have spilled back onto the arterial network</li> </ul>	<ul style="list-style-type: none"> <li>Allows the algorithm to search for a solution which minimizes the total amount of spillback. If spillback cannot be avoided at one ramp even when the maximum metering rate is applied, the algorithm will still try to reduce spillback at all other ramps</li> <li><b>Conclusion: Carry forward</b></li> </ul>
		<ul style="list-style-type: none"> <li>Maximum number of vehicles spilled back at any one ramp</li> </ul>	<ul style="list-style-type: none"> <li>Provides no incentive to address spillback at other ramps if the “worst” ramp has a higher level of spillback which cannot be avoided</li> <li><b>Conclusion: Do not use</b></li> </ul>
Minimize inequity	Ramp delay	<ul style="list-style-type: none"> <li>Maximum ramp delay along the corridor</li> </ul>	<ul style="list-style-type: none"> <li>Only considers the worst ramp, and therefore will not have any impact at ramps with intermediate levels of delay. This may be a particular issue if several ramps have similar (high) levels of delay</li> <li><b>Conclusion: Do not use</b></li> </ul>
		<ul style="list-style-type: none"> <li>Standard deviation of ramp delays</li> </ul>	<ul style="list-style-type: none"> <li>Difficult for people to understand (less intuitive)</li> <li>Includes data for all ramps in the utility calculation</li> <li>Easy to calculate</li> <li><b>Conclusion: Carry forward</b></li> </ul>
		<ul style="list-style-type: none"> <li>Difference between minimum and maximum ramp delay</li> </ul>	<ul style="list-style-type: none"> <li>Easy for people to understand</li> <li>Ignores ramps with intermediate levels of delay (does not capture the full pattern of variation)</li> <li><b>Conclusion: Do not use</b></li> </ul>
		<ul style="list-style-type: none"> <li>Gini co-efficient</li> </ul>	<ul style="list-style-type: none"> <li>Difficult for people to understand (less intuitive)</li> <li>Requires definition of an appropriate baseline</li> <li>More suited to measures expressed on an origin-destination basis</li> <li><b>Conclusion: Do not use</b></li> </ul>

# **APPENDIX K**

## **THE VISSIM TEST NETWORK**

## **K. OVERVIEW OF THE VISSIM TEST NETWORK**

This appendix provides a more detailed description of the VISSIM test network which was developed to evaluate the new ramp metering algorithm.

Figure K-1 provides an illustration of the general on-ramp arrangement assumed in the VISSIM model, while Figure K-2 deals with the off-ramp case. The position of the ramp meter is shown, as well as all traffic sensors, some of which are required to provide input to the ramp metering algorithm, and others which are used solely by the VISSIM controller in implementing the signal timing. Figure K-2 also shows the assumed speed distribution for the freeway mainline and the ramps.

Other key parameters and assumptions are summarized in Table K-1, while Table K-2 provides additional information on the emergency stop and lane change position for link connectors. These latter parameters influence vehicle behaviour for mandatory lane changes, and must therefore be adjusted at major junctions (such as on- and off-ramps) to prevent unrealistic lane manoeuvres and associated congestion.

A summary of the travel demand assumptions used in the test network can be found in Figure K-3.

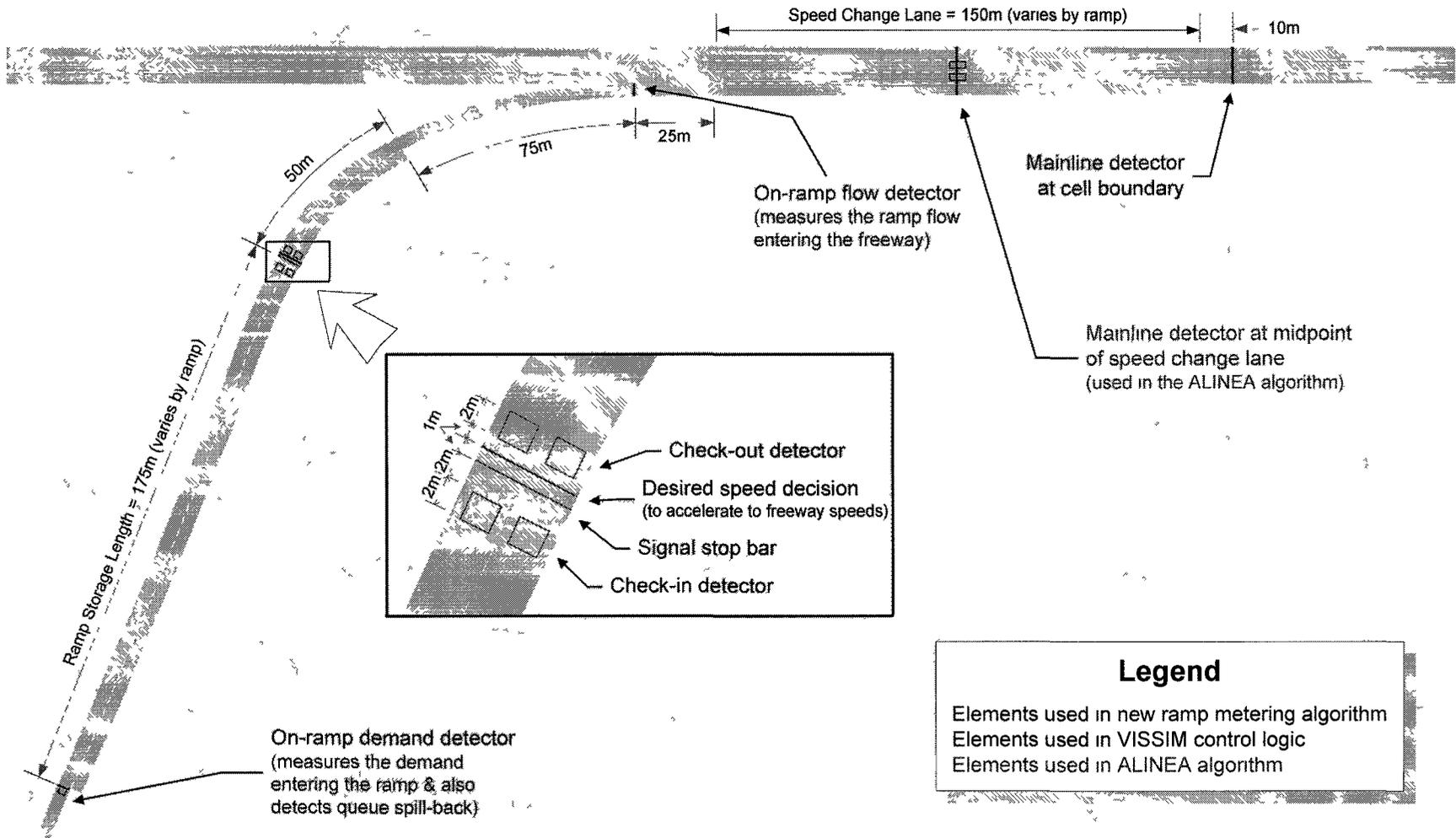


Figure K-1 Typical On-Ramp Configuration

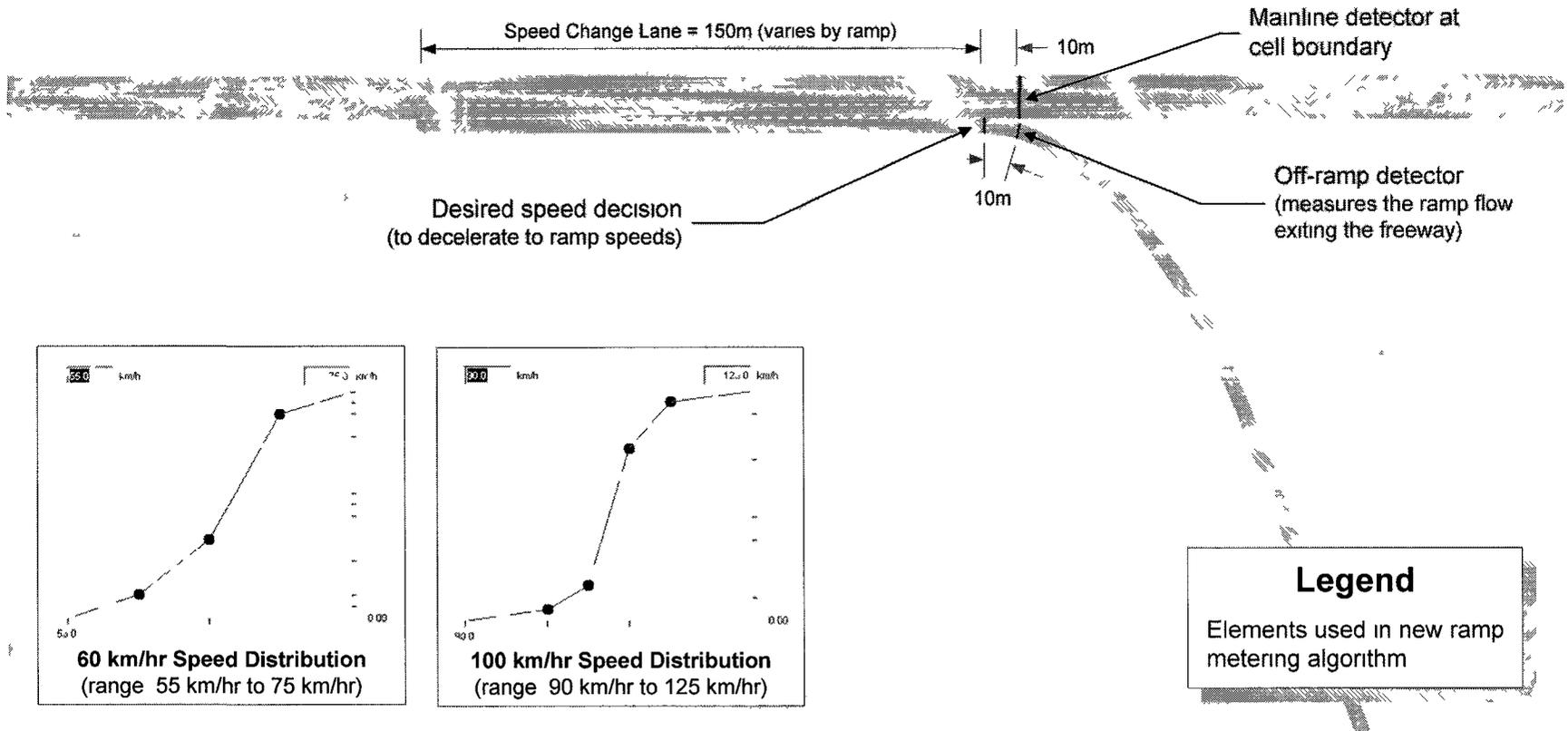


Figure K-2 Typical Off-Ramp Configuration & Assumed Speed Distributions

**Table K-1 The VISSIM Test Network: Key Parameters & Assumptions**

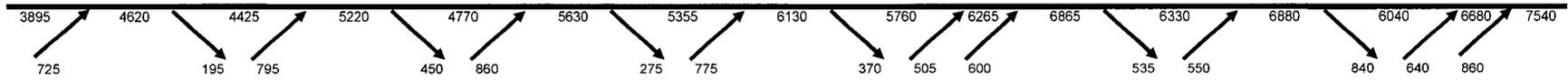
<b>Parameter</b>	<b>Assumed Value</b>
Simulation duration	<ul style="list-style-type: none"> <li>• 2.5 hours</li> <li>• Includes a 500 second interval for “loading” the network which is excluded from all network performance statistics</li> </ul>
Vehicle characteristics & traffic composition	<ul style="list-style-type: none"> <li>• VISSIM default values were used for all vehicle characteristics</li> <li>• The traffic stream was assumed to include 2% heavy vehicles</li> </ul>
Speed limit	<ul style="list-style-type: none"> <li>• The speed limit for the freeway was set at 100 km/hr, consistent with practice in Ontario</li> <li>• Ramp speeds were set at 60 km/hr</li> <li>• For vehicles entering the freeway, ‘desired speed decisions’ were placed roughly 150 m in advance of the speed change lane to allow vehicles to accelerate to freeway speeds prior to merging with mainline flow</li> <li>• For vehicles exiting the freeway, ‘desired speed decisions’ were placed at the ramp gore to model deceleration behaviour</li> <li>• For each speed ‘zone’, speed distributions were defined to capture variation among drivers (refer to Figure K-2)</li> </ul>
Traffic sensors	<ul style="list-style-type: none"> <li>• Traffic sensors were placed strategically throughout the freeway network to capture real-time data for use in the new ramp metering algorithm, as well as the VISSIM traffic controller</li> <li>• In general, sensors were placed in locations where they would typically already be present on the freeway (i.e. near on- and off-ramps). Sensors were also provided at other locations as needed to ensure reasonable algorithm performance, keeping real-world cost constraints in mind. While additional sensors could certainly be added to the test network to improve the accuracy of the traffic estimates, in reality, cost considerations will limit the extent of sensor deployment. As a result, an effort was made to adopt realistic sensor assumptions; the number of sensors included in the model is believed to represent a reasonable compromise between algorithm performance and cost of sensor installation</li> <li>• All VISSIM detectors were assumed to have a smoothing factor of 1 (rather than the default value of 0.25) for adjusting the detector occupancy rate, since any required adjustments are undertaken directly by the ramp meter controller</li> </ul>
Network geometry	<ul style="list-style-type: none"> <li>• The speed change lanes in the VISSIM test network were assumed to vary in length between 150 m and 250 m for on-ramps, and 150 m and 225 m for off-ramps</li> <li>• The ramp storage length was assumed to range from 250 m to 450 m, assuming 2 storage lanes per ramp</li> <li>• The above assumptions are loosely based on the characteristics of the existing freeway network in Ottawa, Ontario. The assumed geometry is therefore considered to be representative of a real-world network, and as such, may not meet current design standards</li> </ul>

Parameter	Assumed Value
Driver behaviour	<ul style="list-style-type: none"> <li>• Freeway links were modelled based on the ‘Freeway (free lane selection)’ driver behaviour parameter set, which uses the Wiedemann 99 car following model</li> <li>• Ramp links were modelled according to the ‘urban (motorized)’ parameter set, which uses the Wiedemann 74 car following model</li> <li>• For the most part, the default VISSIM parameters were adopted in the analysis (as defined in VISSIM version 5.10-11). However, certain adjustments were made to address on-ramp merging issues (refer to Appendix L for additional details)</li> <li>• In addition, the CC4 (negative following threshold) and CC5 (positive following threshold) parameters in the Wiedemann 99 model were modified from <math>\pm 0.35</math> to <math>\pm 0.5</math></li> </ul>
Reaction to amber signal	<ul style="list-style-type: none"> <li>• Reaction to the amber signal was modelled using the ‘one decision’ model. Under this type of model, it is assumed that drivers view the amber light and then make a single decision on whether or not to stop</li> <li>• Three parameters are used to calculate the probability of the driver stopping. In the VISSIM test network, the parameters were assigned to yield a probability of close to 100% in all cases</li> <li>• The above assumption is considered to be reasonable when the ramp meter is in operation, since drivers know that only one vehicle is allowed to proceed at a time and should be therefore be prepared to stop</li> </ul>
Lane change behaviour	<ul style="list-style-type: none"> <li>• Lane change behaviour in VISSIM is influenced by the driver behaviour parameters described above</li> <li>• Lane change behaviour is also influenced by the emergency stop and lane change position for link connectors which are described in Table K-2</li> <li>• On occasion, unrealistic lane change behaviour was observed in the auxiliary ramp storage lane: <ul style="list-style-type: none"> <li>– For the scenario with no ramp metering, the auxiliary storage lane was closed in VISSIM so that no vehicles could use it</li> <li>– In the scenarios involving ramp metering, there is no way to close the second storage lane if the meter is turned off. As a result, vehicles were sometimes observed to move into the auxiliary lane to pass a slower moving vehicle. Although this move would be illegal in reality, it should have minimal impact on the simulation results</li> <li>– It was also observed that under ramp metering operation, some vehicles in the continuous ramp lane would briefly merge into the auxiliary lane after clearing the meter and then exit the lane before the lane drop. Since vehicles may only proceed one at a time, there are no conflicts to impact the lane change manoeuvres, and the behaviour, though unrealistic, should again have minimal impact on the simulation results</li> </ul> </li> </ul>

**Table K-2 Emergency Stop & Lane Change Position for Link Connectors:  
Assumptions Adopted at Key Locations**

<b>Connector Location</b>	<b>Assumption</b>	<b>Rationale</b>
Off-ramp gore	Lane change distance set so that drivers would become aware of a freeway exit roughly 1500 m before reaching it	Provides ample opportunity for exiting drivers to position themselves in the correct lane in advance of their desired exit, thus eliminating the potential for aggressive last-minute lane changes
	Emergency stop position assumed to be located 100 m downstream from the start of the deceleration lane	Chosen to achieve a balance so that a) any vehicle deceleration occurs primarily in the speed change lane, and b) under congested off-ramp flow, exiting drivers join the end of any queue in the deceleration lane rather than queuing in the mainline lanes near the ramp gore
End of acceleration/ deceleration lane	Lane change distance set to a value greater than the length of the auxiliary lane	Ensures that mainline vehicles travelling through the ramp area are aware the auxiliary lane is ending and do not attempt to use it
End of 2-lane on-ramp section	Lane change distance set to coincide with the location of the ramp meter	Ensures that vehicles begin to respond to the downstream lane drop immediately after passing the ramp meter (but not before, otherwise, vehicles would not queue evenly in both storage lanes)

Hourly Flow During Peak 15-Minutes



Peak Hour Volume

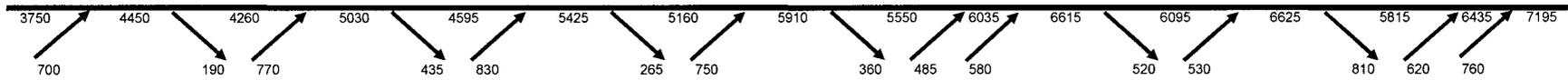


Figure K-3 Travel Demand Assumptions

# **APPENDIX L**

## **OPTIONS TO IMPROVE VISSIM ON-RAMP MERGE BEHAVIOUR**

**Table L-1 Summary of VISSIM Parameter Changes to Improve On-Ramp Merge Behaviour**

Parameter	Description	Default Value	Suggestion from PTV America	Value Applied	Rationale for Decision
Observed Vehicles	Affects how well vehicles in the network can predict other vehicles' movements and react accordingly.	2	Increasing the number of observed vehicles may help, but this is really to provide more guidance in the decision of whether or not to decelerate based on what vehicles directly in front of you may be doing.	10	No obvious downside to increasing this value other than simulation run-time.
CC0 – Standstill distance	Defines the desired distance between stopped cars.	1.5 m	Increase these values to have mainline vehicles drive with more space between them, thus providing more gaps for merging vehicles from the ramp.	Default	While adjusting the car-following parameters may 'trick' VISSIM into modeling more freeway merges under conditions of heavy demand, the effects will be felt throughout the entire freeway system in the form of lower capacities. Even if the changes were restricted to links in the vicinity of on-ramps, the reduction in capacity would be evident even under lower traffic loads where merging is not an issue. Moreover, even though the number of successful merges may increase due to the availability of larger gaps, the resultant behaviour is not reflective of actual behaviour in merge areas (i.e. where only certain vehicles behave cooperatively, instead of all vehicles arbitrarily leaving larger gaps).
CC1 – Headway Time	The time (in seconds) that a driver wants to keep between himself & the vehicle in front. The higher the value, the more cautious the driver is. Thus, at a given speed $v$ [m/s], the safety distance $dx\_safe$ is computed as: $dx\_safe = CC0 + CC1 * v$	0.9 sec		Default	

Parameter	Description	Default Value	Suggestion from PTV America	Value Applied	Rationale for Decision
Deceleration Thresholds for Necessary Lane Changes	The lane change aggressiveness can be adjusted by defining Accepted and Maximum Deceleration thresholds for both the lane changer ('own') and the vehicle that he is moving ahead of ('trailing'). The deceleration rate actually applied varies between the Accepted and Maximum value depending on the distance from the emergency stop position.	Maximum: -4.0 m/s <sup>2</sup> (own) -3.0 m/s <sup>2</sup> (trailing)  Accepted: -1.0 m/s <sup>2</sup> (own) -0.5 m/s <sup>2</sup> (trailing)	Increase the Accepted Deceleration for the 'own' & 'trailing' vehicles as this allows vehicles to adjust their speeds even more to allow for merging manoeuvres.	Default	While increasing the Accepted Deceleration may facilitate merging, it was found that reducing the 'aggressiveness' assumptions for the trailing vehicle may reduce unrealistic collisions at lower speeds.  Given the length of the acceleration lanes in the test network, vehicles will already be relatively close to the emergency stop position when they first attempt to change lanes, implying that the deceleration rate applied will be largely unaffected by the Accepted Deceleration threshold, even if somewhat higher values are applied.
Minimum Headway	Defines the minimum distance to the vehicle in front that must be available for a lane change in standstill condition.	0.5 m	Increasing these values (either solely or in combination) will reduce collisions (and capacity). Changes will also affect merging in free flow conditions.	Default	Increasing the safety distance reduction factor / minimum headway was found to make the problem worse, potentially because it reduces the number of acceptable gaps, resulting in more vehicles stored in the acceleration lane. Since collisions were most frequently associated with vehicles accelerating from a stopped position near the end of the acceleration lane, increasing the number of queued vehicles results in even more (unrealistic) collisions.
Safety Distance Reduction Factor	Reduces the safety distance during lane changes.	0.60		Default	
Maximum Deceleration for Cooperative Braking	Determines when mainline traffic will brake cooperatively. If the mainline vehicle would need to brake harder than this value for the merging vehicle to safely change lanes, the mainline vehicle will not begin braking. Thus, the higher the value, the more cooperative braking occurs.	-3.0 m/s <sup>2</sup>	0.7 to 1.0 times the local acceleration due to gravity  (-6.9 to -9.81 m/s <sup>2</sup> )	-7.5 m/s <sup>2</sup>	Substantially reduces unrealistic ramp queues by creating gaps to accommodate merging traffic.

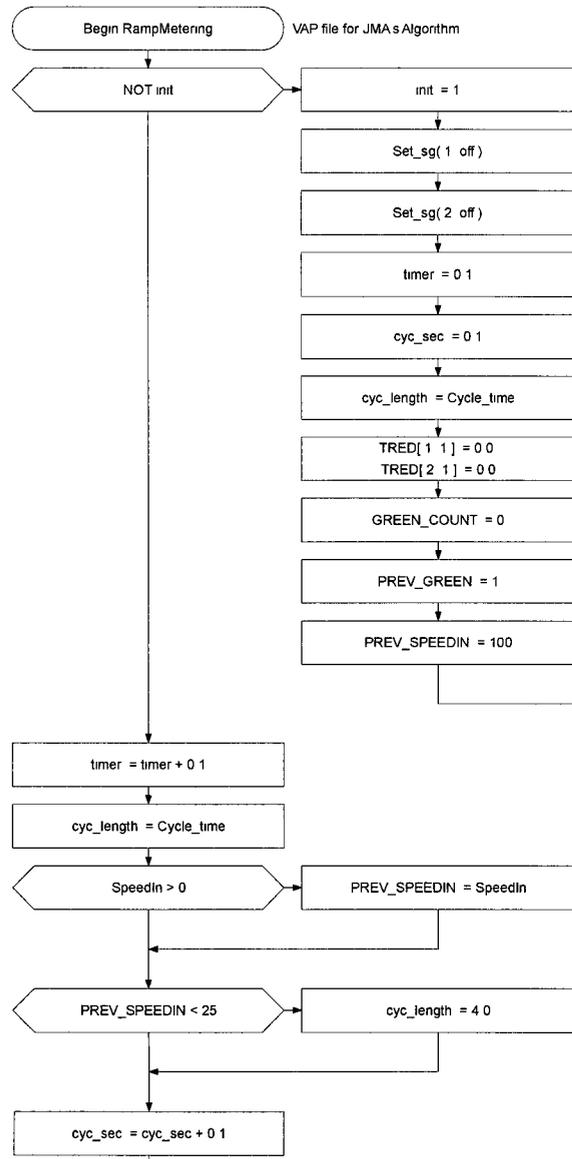
Parameter	Description	Default Value	Suggestion from PTV America	Value Applied	Rationale for Decision
Speed Change Rules in Vicinity of Merge	Can be implemented at the start and end of the merge area, causing mainline vehicles to slow down as they approach the on-ramp, then return to normal operating speeds further downstream.	Speed change rules must be added manually to the model (no defaults)	Suggested the use of speed change rules as a viable method for improving merge behaviour.	Speed change rules not used	Use of speed change rules is somewhat arbitrary. No good basis for determining what the speed reduction should be. <sup>1</sup> May cause unrealistic behaviour under light demand. May impact capacity in unexpected ways. Forces all drivers to react, when in reality, it is most likely to be drivers in the shoulder lane who slow down. <sup>2</sup> Assumes that it is easier to merge into slower moving traffic, which is not necessarily the case (since the safety distance is reduced at lower speeds, drivers may accept smaller gaps, but this also means that mainline drivers are travelling more closely together).
Priority Rules in Vicinity of Merge	Can be used to cause drivers in the shoulder lane to yield if a vehicle on the ramp reaches the end of the acceleration lane.	Priority rules must be added manually to the model (no defaults)	Do not consider priority rules to be an appropriate method for adjusting driving behaviour in merge areas.	Priority rules not used	While the use of priority rules reduces ramp queues, the resultant behaviour is not entirely realistic since mainline drivers are sometimes forced to come to a complete stop, rather than simply slowing down to let someone in. Once a driver on the mainline stops, the gap created is large enough that several vehicles on the ramp merge simultaneously, which again is unrealistic.

<sup>1</sup> The speed rules in VISSIM are static, and cannot be varied as a function of operating conditions. However, in reality, the speed reduction in merge areas depends on several factors, including the ramp flow (with no vehicles on the ramp, mainline drivers are unlikely to reduce their speed). Since ramp metering effects ramp flow, it too may influence the extent of speed reduction. Where ramp vehicles are present, the speed reduction can be precautionary in nature (as drivers slow down either consciously or subconsciously in response to potential conflicts which have not yet materialized), or mandatory (in the sense that action is necessary either to avoid a collision or act cooperatively towards a particular merging vehicle). Thus, speed reduction is an emergent behaviour as drivers respond to actual conditions, making it difficult to apply a set reduction factor which would be applicable for all situations.

<sup>2</sup> Applying the speed change rule to only the shoulder lane would partially address this concern, but vehicles moving into the shoulder lane after the rule would not be affected by the speed reduction.

# **APPENDIX M**

## **THE RAMP SIGNAL CONTROLLER IN VISSIM**



VAP file for JMA's Algorithm

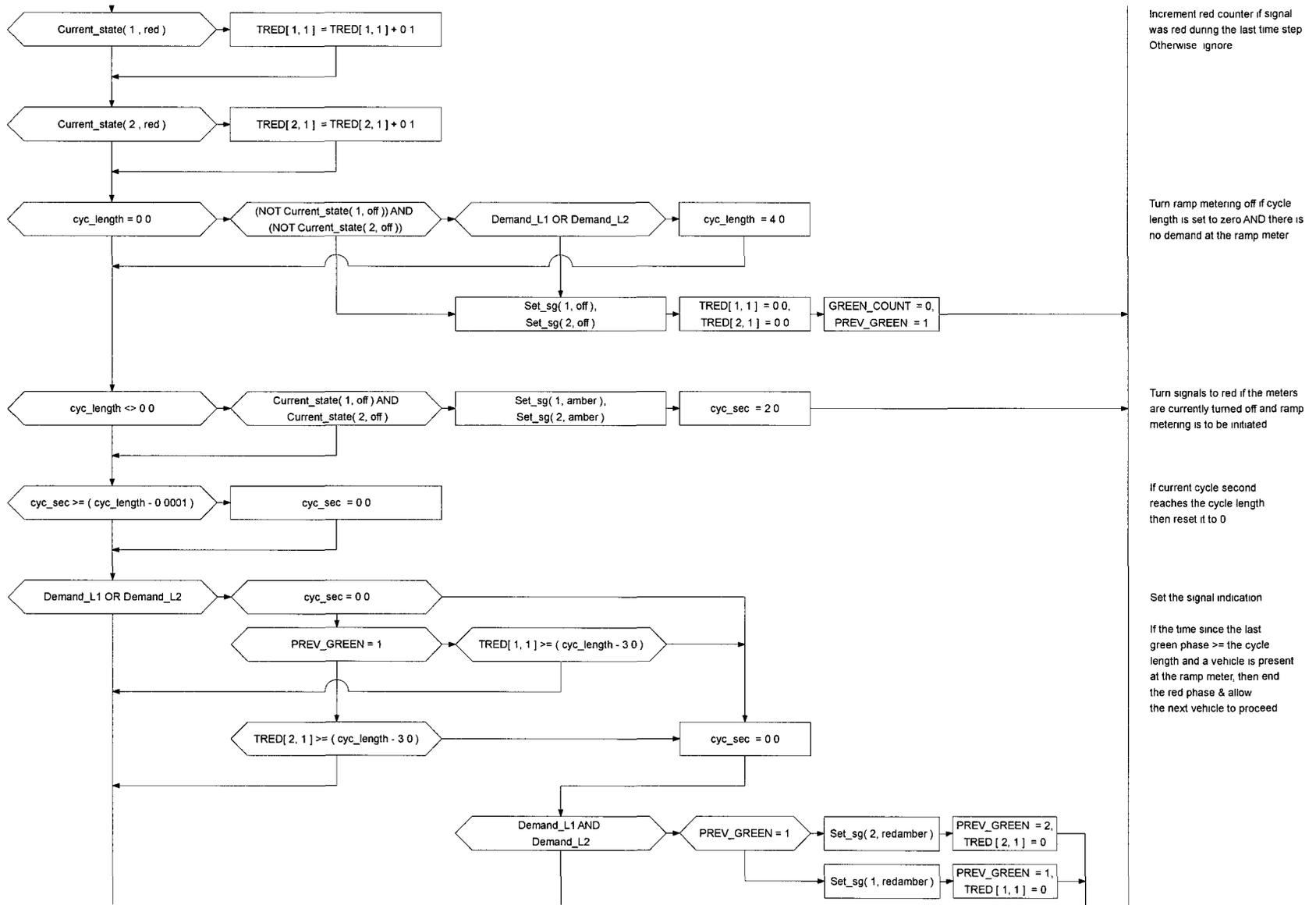
Program Comments

Initialize variables

**VisVAP Diagram for the Ramp Signal Controller in VISSIM: New Ramp Control Algorithm**

Adjust cycle length if queue spill back is detected at the ramp entrance

Increment cycle second



Increment red counter if signal was red during the last time step  
Otherwise ignore

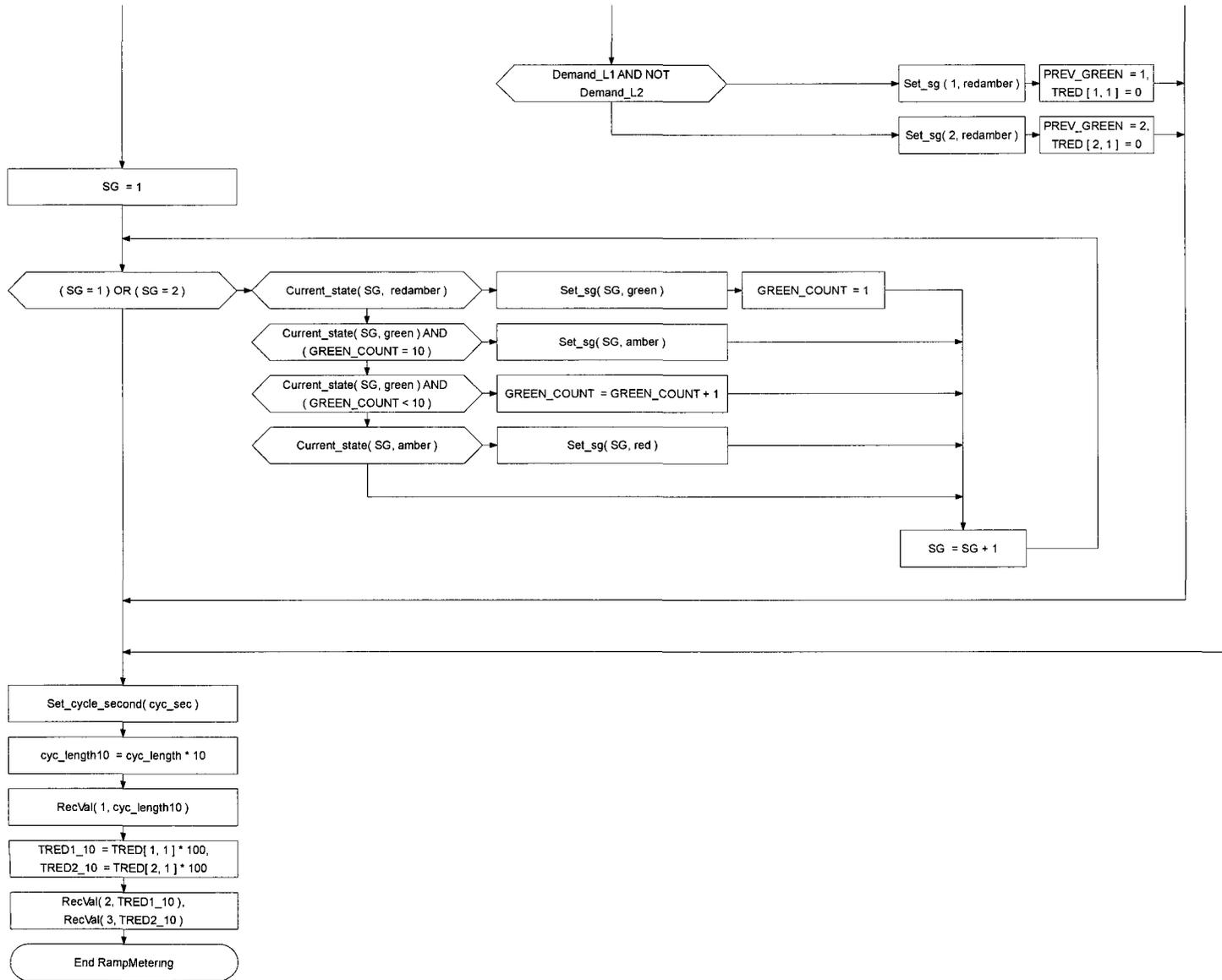
Turn ramp metering off if cycle length is set to zero AND there is no demand at the ramp meter

Turn signals to red if the meters are currently turned off and ramp metering is to be initiated

If current cycle second reaches the cycle length then reset it to 0

Set the signal indication

If the time since the last green phase >= the cycle length and a vehicle is present at the ramp meter, then end the red phase & allow the next vehicle to proceed



Switch signal states from red-amber to green to amber to red as appropriate  
 Note  
 Amber and red-amber times are set to 1 second by default

Send cycle second to VISSIM in order to be displayed in the signal times window  
 Send other variables to VISSIM

**PARAMETERS, ARRAYS, and EXPRESSIONS used in the VisVAP Diagram**

PARAMETERS	Gen	Prog 1	Prog 2	Prog 3	Prog 4	Prog 5	Prog 6	Prog 7	Prog 8	Comment
RDETAPP_L1	101									Ramp detector, approaching stop bar, lane 1
RDETAPP_L2	102									Ramp detector, approaching stop bar, lane 2
QDET	300									Queue spill-back detector

ARRAYS	Dim2	Dim1	[ 1 ]	Comment
TRED	2	1	0	Time since signal turned red
TRED[2]			0	

EXPRESSIONS	Contents	Comment
Demand_L1	Detection( RDETAPP_L1 )	Determines if someone is waiting at the stop bar in Lane 1
Demand_L2	Detection( RDETAPP_L2 )	Determines if someone is waiting at the stop bar in Lane 2
SpeedIn	Velocity ( QDET ) * 3.6	Gives the speed (in km/hr) of the last vehicle detected since the previous check

### VAP Code for Implementing the New Ramp Control Algorithm in VISSIM

```

PROGRAM RampMeteringLogic_WSplBckDet; /* RMLogic_Jan2011Ver_WithSpillBckDetection.vv */

VAP_FREQUENCY 10;

CONST
    RDETAPP_L1 = 101,
    RDETAPP_L2 = 102,
    QDET = 300;

/* ARRAYS */
ARRAY
    TRED[ 2, 1 ] = [[0], [0]];

/* SUBROUTINES */

/* PARAMETERS DEPENDENT ON SCJ-PROGRAM */

/* EXPRESSIONS */
    Demand_L1 := Detection( RDETAPP_L1 );
    Demand_L2 := Detection( RDETAPP_L2 );
    SpeedIn := Velocity ( QDET ) * 3.6;

/* MAIN PROGRAM */

S00Z001: IF NOT init THEN
S01Z001:   init := 1;
S01Z002:   Set_sg( 1, off );
S01Z003:   Set_sg( 2, off );
S01Z004:   timer := 0.1;
S01Z005:   cyc_sec := 0.1;
S01Z006:   cyc_length := Cycle_time;
S01Z007:   TRED[ 1, 1 ] := 0.0;
S01Z008:   TRED[ 2, 1 ] := 0.0;
S01Z009:   GREEN_COUNT := 0;
S01Z010:   PREV_GREEN := 1;
S01Z011:   PREV_SPEEDIN := 100
    ELSE
S00Z013:   timer := timer + 0.1;
S00Z014:   cyc_length := Cycle_time;
S00Z015:   IF SpeedIn > 0 THEN
S01Z015:     PREV_SPEEDIN := SpeedIn;
S00Z017:   IF PREV_SPEEDIN < 25 THEN
S01Z017:     cyc_length := 4.0;
S00Z019:     cyc_sec := cyc_sec + 0.1;
S00Z020:     IF Current_state( 1, red ) THEN
S01Z020:       TRED[ 1, 1 ] := TRED[ 1, 1 ] + 0.1;
S00Z022:     IF Current_state( 2, red ) THEN

```

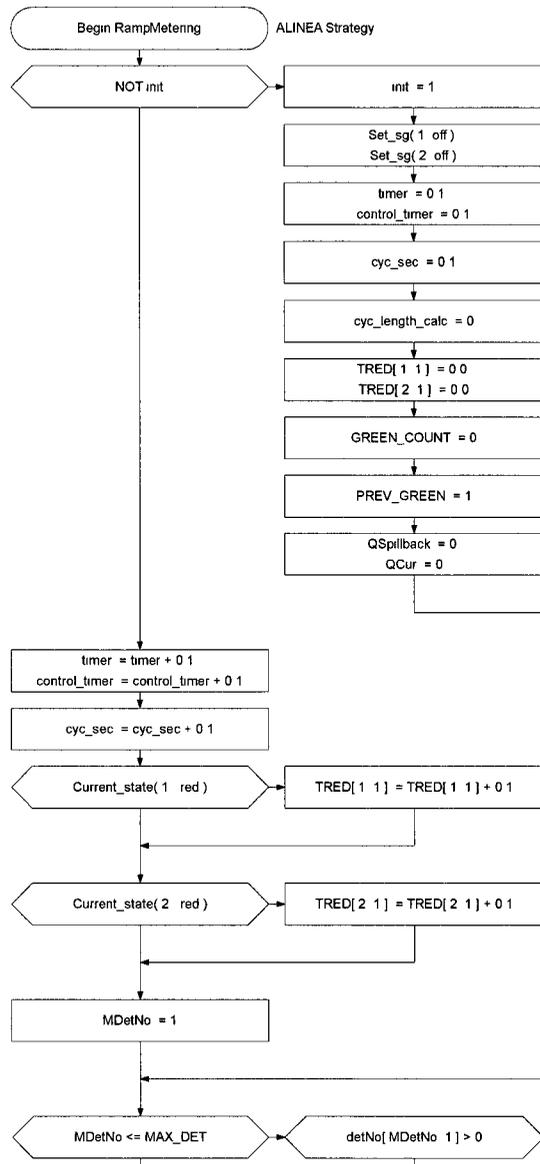
```

S01Z022: TRED[ 2, 1 ] := TRED[ 2, 1 ] + 0.1;
S00Z024: IF cyc_length = 0.0 THEN
S01Z024:   IF (NOT Current_state( 1, off )) AND (NOT Current_state( 2, off )) THEN
S02Z024:     IF Demand_L1 OR Demand_L2 THEN
S03Z024:       cyc_length := 4.0;
S00Z028:       IF cyc_length <> 0.0 THEN
S01Z028:         IF Current_state( 1, off ) AND Current_state( 2, off ) THEN
S02Z028:           Set_sg( 1, amber ); Set_sg( 2, amber );
S03Z028:           cyc_sec := 2.0
ELSE
S00Z030:   IF cyc_sec >= ( cyc_length - 0.0001 ) THEN
S01Z030:     cyc_sec := 0.0;
S00Z032:   IF Demand_L1 OR Demand_L2 THEN
S01Z032:     IF cyc_sec = 0.0 THEN
S03Z035:       cyc_sec := 0.0;
S02Z037:       IF Demand_L1 AND Demand_L2 THEN
S03Z037:         IF PREV_GREEN = 1 THEN
S04Z037:           Set_sg( 2, redamber );
S05Z037:           PREV_GREEN := 2; TRED [ 2, 1 ] := 0
ELSE
S04Z038:           Set_sg( 1, redamber );
S05Z038:           PREV_GREEN := 1; TRED [ 1, 1 ] := 0
END
ELSE
S02Z040:   IF Demand_L1 AND NOT Demand_L2 THEN
S04Z040:     Set_sg ( 1, redamber );
S05Z040:     PREV_GREEN := 1; TRED [ 1, 1 ] := 0
ELSE
S04Z041:     Set_sg( 2, redamber );
S05Z041:     PREV_GREEN := 2; TRED [ 2, 1 ] := 0
END
END
ELSE
S01Z033:   IF PREV_GREEN = 1 THEN
S02Z033:     IF TRED[ 1, 1 ] >= ( cyc_length - 3.0 ) THEN
GOTO S03Z035
ELSE
S00Z042:     SG := 1;
S00Z044:     IF ( SG = 1 ) OR ( SG = 2 ) THEN
S01Z044:       IF Current_state( SG, redamber ) THEN
S02Z044:         Set_sg( SG, green );
S03Z044:         GREEN_COUNT := 1;
S04Z049:         SG := SG + 1;
GOTO S00Z044
ELSE
S01Z045:       IF Current_state( SG, green ) AND ( GREEN_COUNT = 10 ) THEN
S02Z045:         Set_sg( SG, amber );
GOTO S04Z049
ELSE

```



```
        END
        ELSE
            GOTO S00Z019
        END
        ELSE
            GOTO S00Z017
        END
    END;
S00Z053: Set_cycle_second( cyc_sec );
S00Z054: cyc_length10 := cyc_length * 10;
S00Z055: RecVal( 1, cyc_length10 );
S00Z056: TRED1_10 := TRED[ 1, 1 ] * 100; TRED2_10 := TRED[ 2, 1 ] * 100;
S00Z057: RecVal( 2, TRED1_10 ); RecVal( 3, TRED2_10 )
PROG_ENDE: .
/*-----*/
```



**VisVAP Diagram for the Ramp  
Signal Controller in VISSIM:  
ALINEA Algorithm**

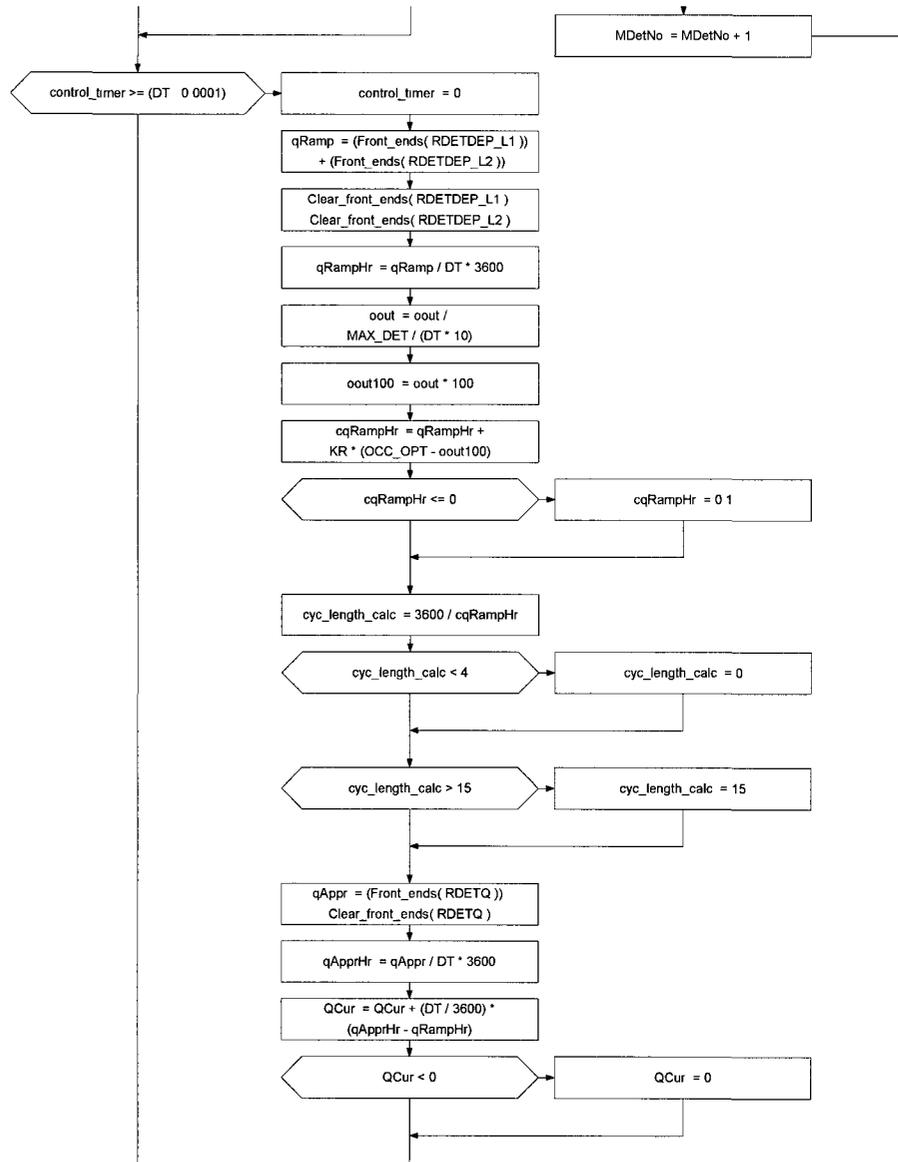
Program Comments

Initialize variables

Increment cycle second

Increment red counter if signal  
was red during the last time step  
Otherwise ignore

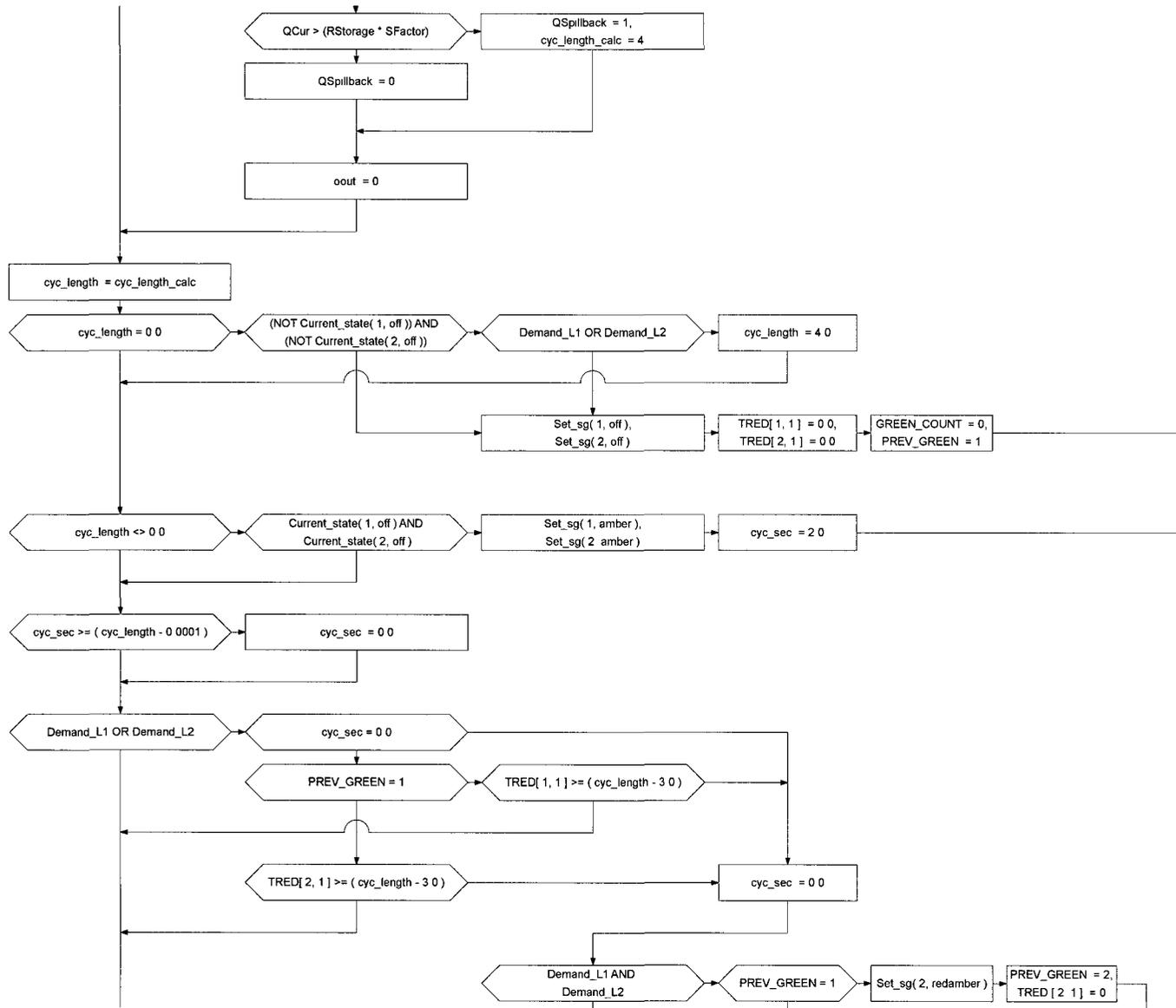
Store info on mainline occupancy



Compute cycle length

Adjust cycle length to fit within allowable range (4 seconds to 15 seconds)

Calculate current queue length



Adjust cycle length if queue spill-back is detected at the ramp entrance

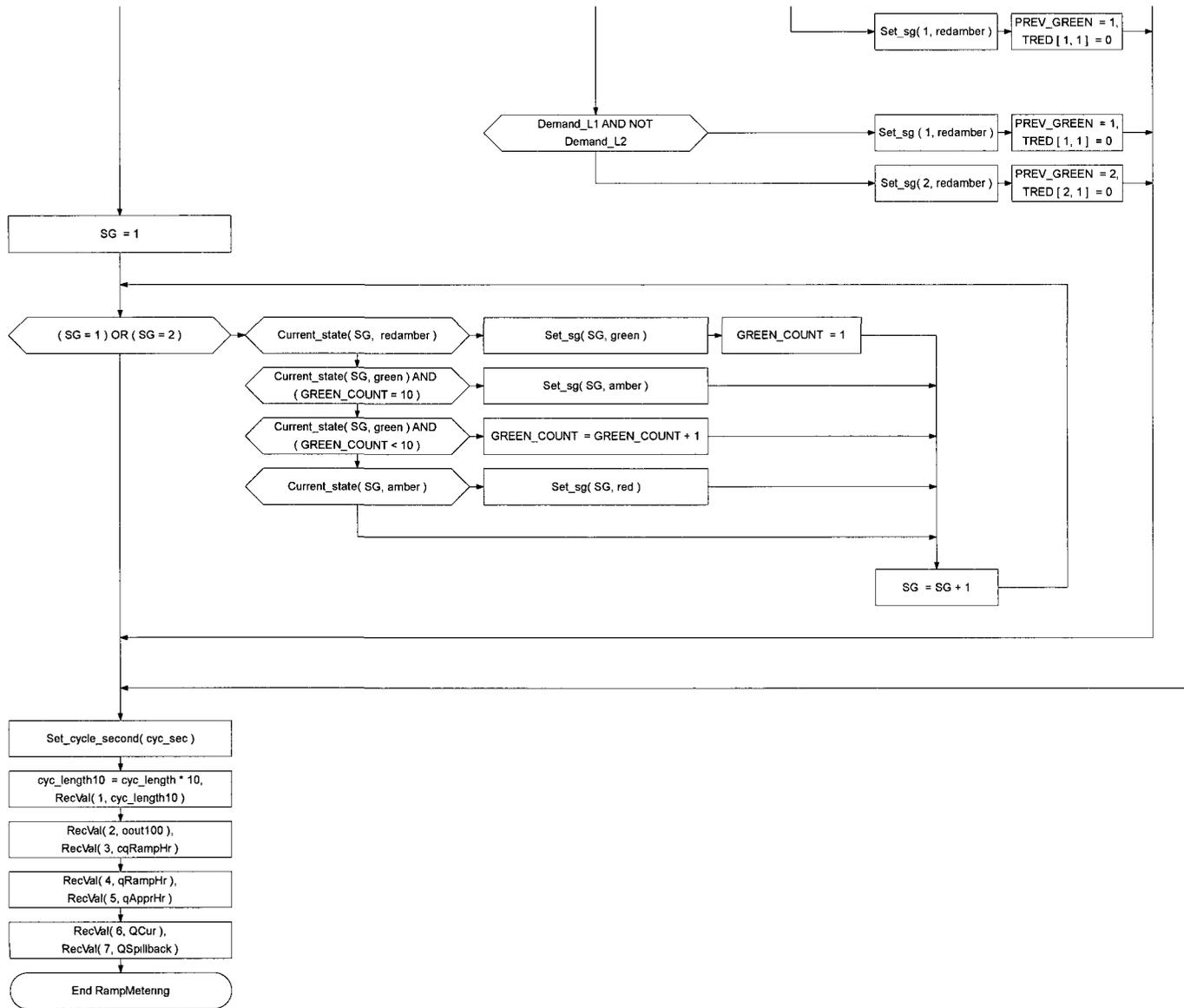
Turn ramp metering off if cycle length is set to zero AND there is no demand at the ramp meter

Turn signals to red if the meters are currently turned off and ramp metering is to be initiated

If current cycle second reaches the cycle length then reset it to 0

Set the signal indication

If the time since the last green phase >= the cycle length and a vehicle is present at the ramp meter, then end the red phase & allow the next vehicle to proceed



Switch signal states from red-amber to green to amber to red as appropriate  
 Note  
 Amber and red-amber times are set to 1 second by default

Send cycle second to VISSIM in order to be displayed in the signal times window  
 Send other variables to VISSIM

### PARAMETERS, ARRAYS, and EXPRESSIONS used in the VisVAP Diagram

PARAMETERS	Gen	Prog 1	Prog 2	Prog 3	Prog 4	Prog 5	Prog 6	Prog 7	Prog 8	Comment
RDETAPP_L1	101									Ramp detector, approaching stop bar, lane 1
RDETAPP_L2	102									Ramp detector, approaching stop bar, lane 2
RDETDEP_L1	201									Ramp detector, departing stop bar, lane 1
RDETDEP_L2	202									Ramp detector, departing stop bar, lane 2
RDETTQ	300									Ramp detector for queue management
MAX_DET	2									Number of mainline detectors
DT	30									Ramp metering control interval (seconds)
KR	70									Regulator parameter
OCC_OPT	28									Target (optimal) occupancy
RStorage		30	36	42	48	54				Ramp storage (vehicles)
SFactor	0.75									'Safety Factor' for ramp storage

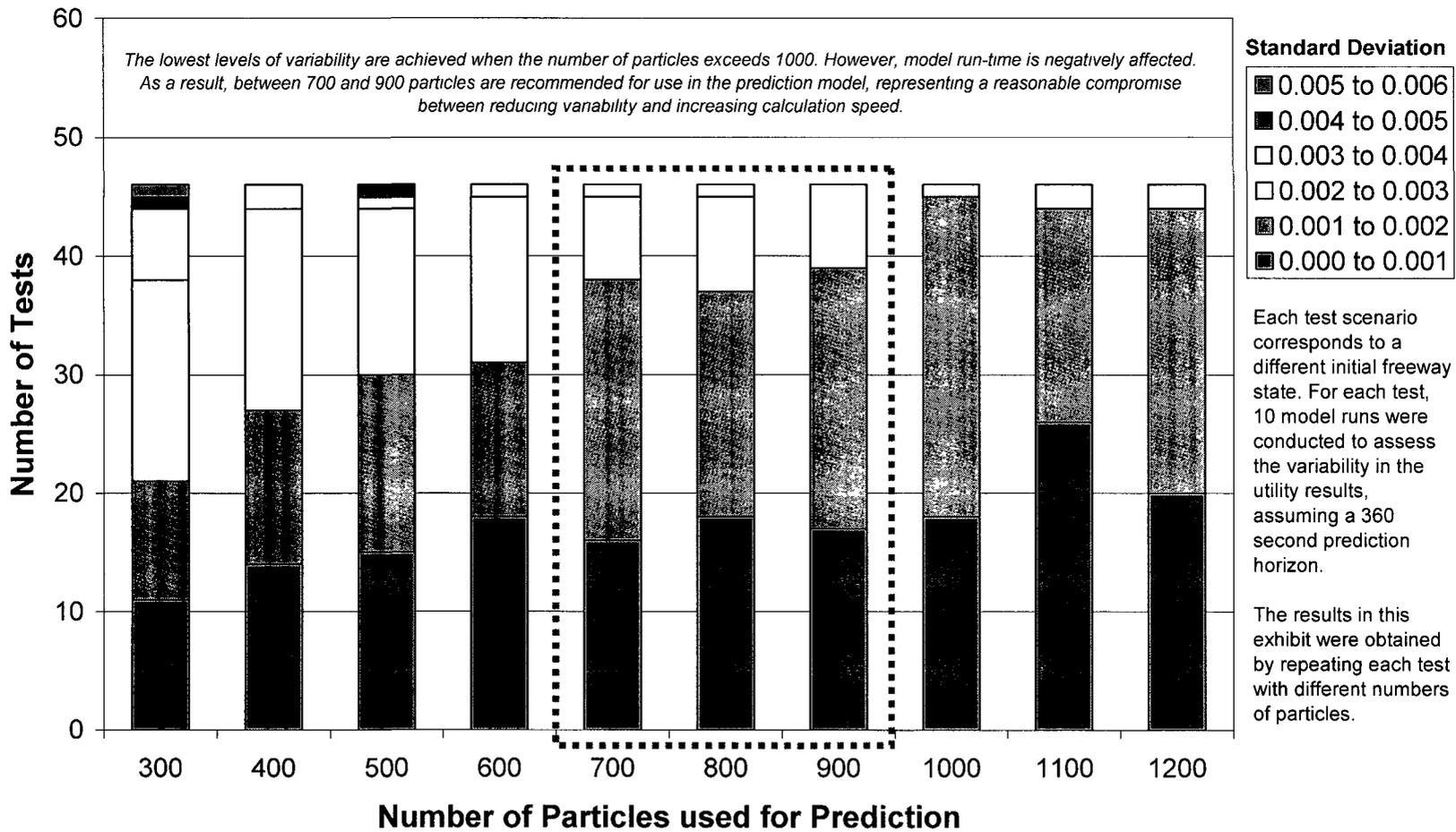
ARRAYS	Dim2	Dim1	[ 1 ]	Comment
TRED	2	1	0.0	Time since signal 1 turned red
TRED[2]			0.0	Time since signal 2 turned red
detNo	2	1	20	1st mainline detector
detNo[2]			30	2nd mainline detector

EXPRESSIONS	Contents	Comment
Demand_L1	Detection( RDETAPP_L1 )	Determines if someone is waiting at the stop bar in Lane 1
Demand_L2	Detection( RDETAPP_L2 )	Determines if someone is waiting at the stop bar in Lane 2

# APPENDIX N

NUMBER OF PARTICLES NEEDED FOR PREDICTION MODE

## Standard Deviation of Freeway Utility Effect of Number of Particles



# **APPENDIX O**

## **ON-LINE PERFORMANCE OF THE FREEWAY TRAFFIC MODEL**

## **O. ON-LINE PERFORMANCE OF THE FREEWAY TRAFFIC MODEL**

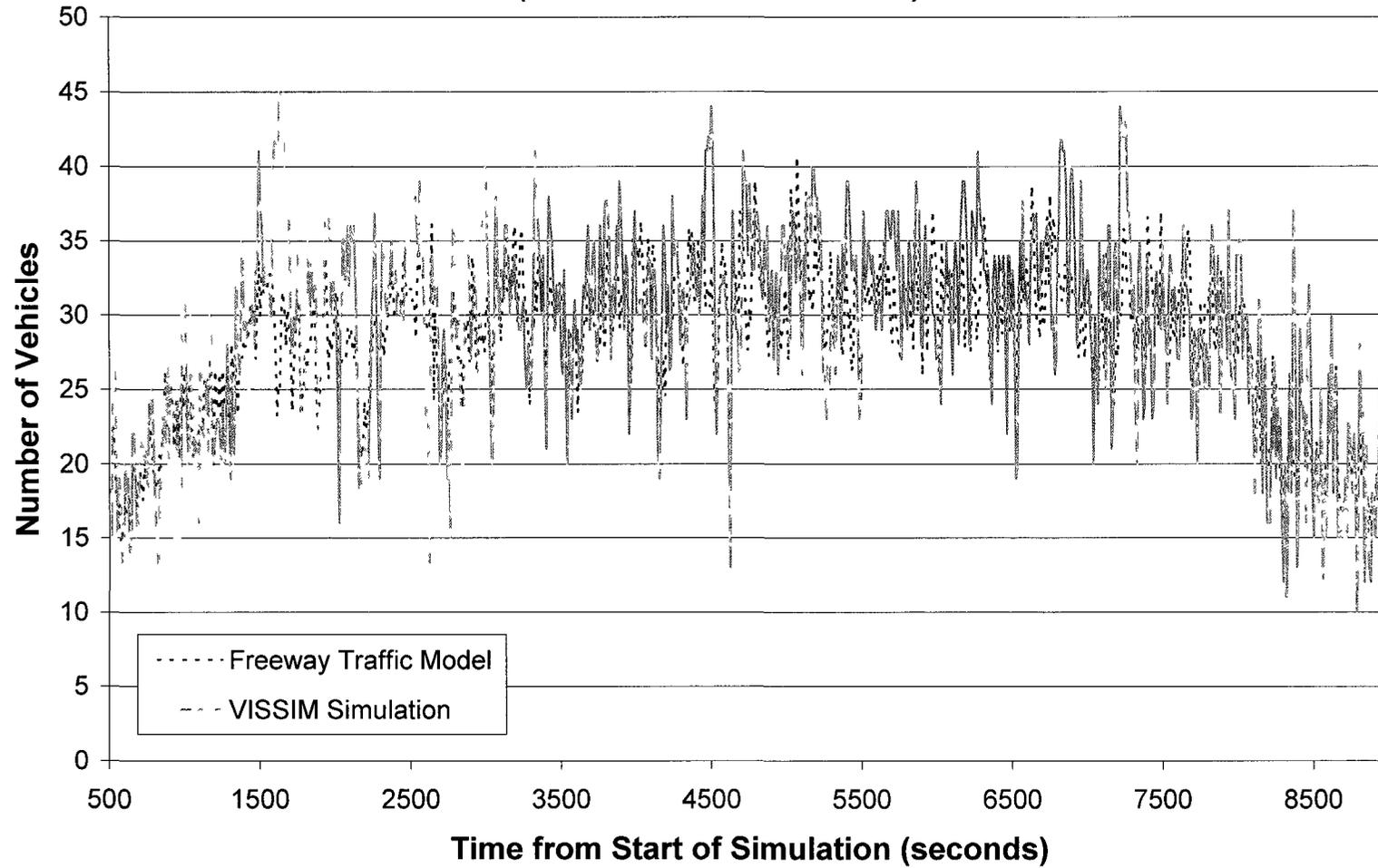
This appendix provides sample results which demonstrate the on-line performance of the Freeway Traffic Model (FTM) when incorporated as part of the ramp control algorithm. The figures correspond to Simulation Run 8 for the algorithm tests where the performance has been optimized for efficiency only.

The first set of figures show the number of vehicles per road segment estimated by the Freeway Traffic Model (operating in tracking mode), compared to the actual number of vehicles observed during the VISSIM simulation. Only selected segments are presented, corresponding to the section of highway experiencing mainline congestion (as well as the immediately upstream and downstream segments).

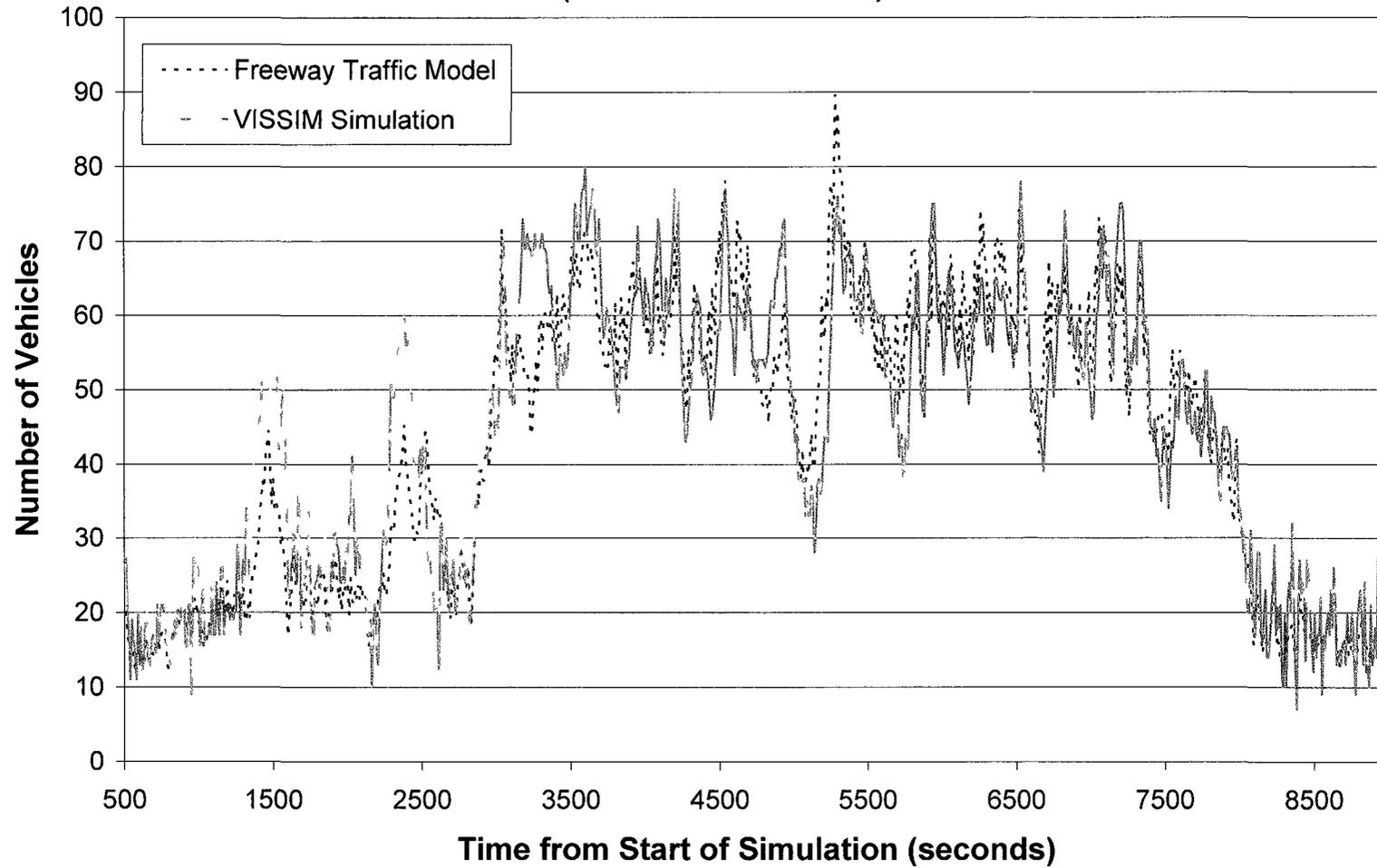
The second set of figures show the ramp queues estimated by the Freeway Traffic Model under ramp metering operation compared to the actual ramp queues measured during the VISSIM simulation. In developing these figures, it has been assumed that each queued vehicle occupies approximately 8.25 m. This assumption is required since VISSIM measures ramp queues in terms of distance, while the Freeway Traffic Model measures ramp queues in terms of stored vehicles.

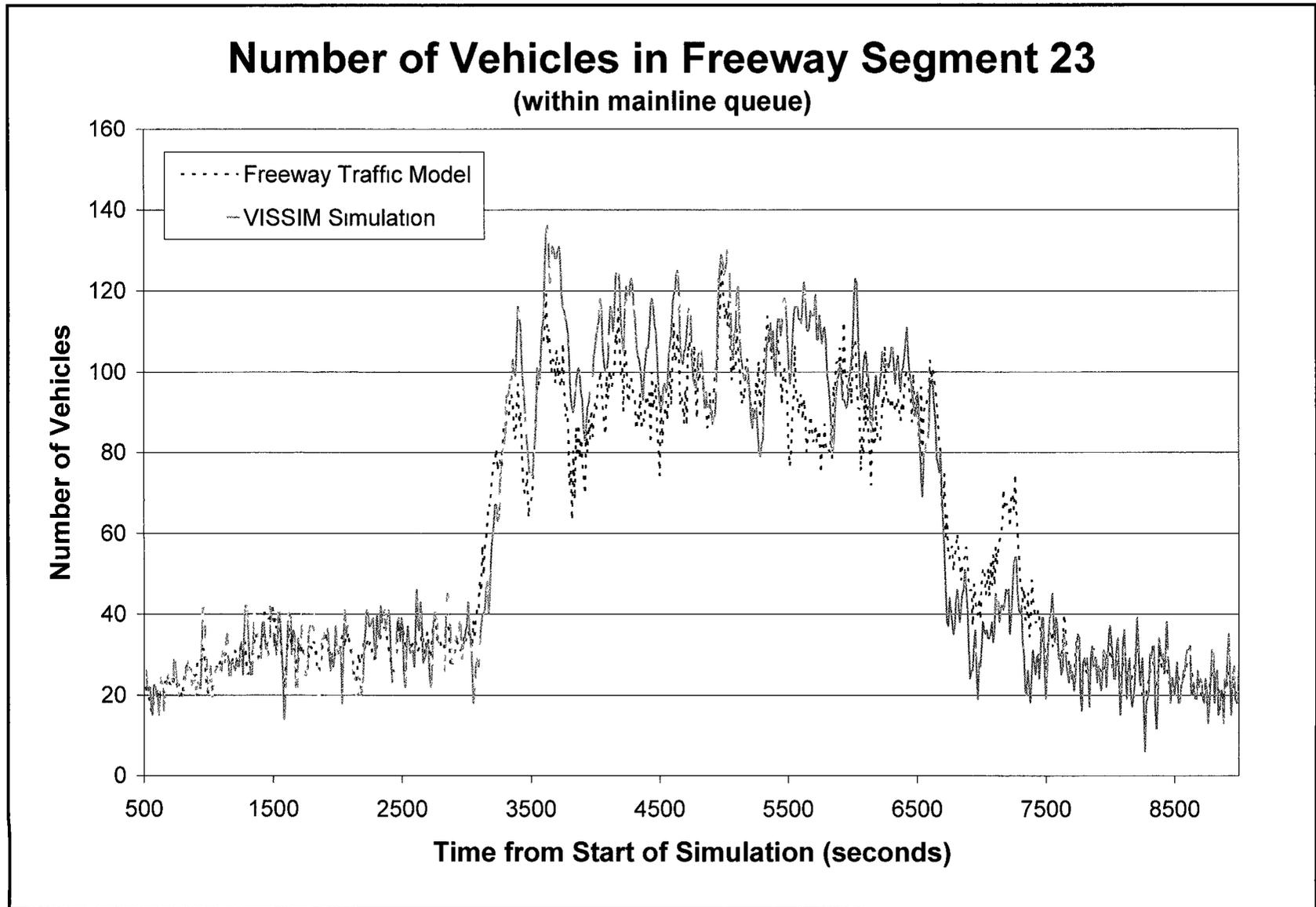
Based on the results presented, the Freeway Traffic Model appears to do a reasonable job of tracking both mainline congestion and ramp queues, with a level of accuracy appropriate for ramp metering applications.

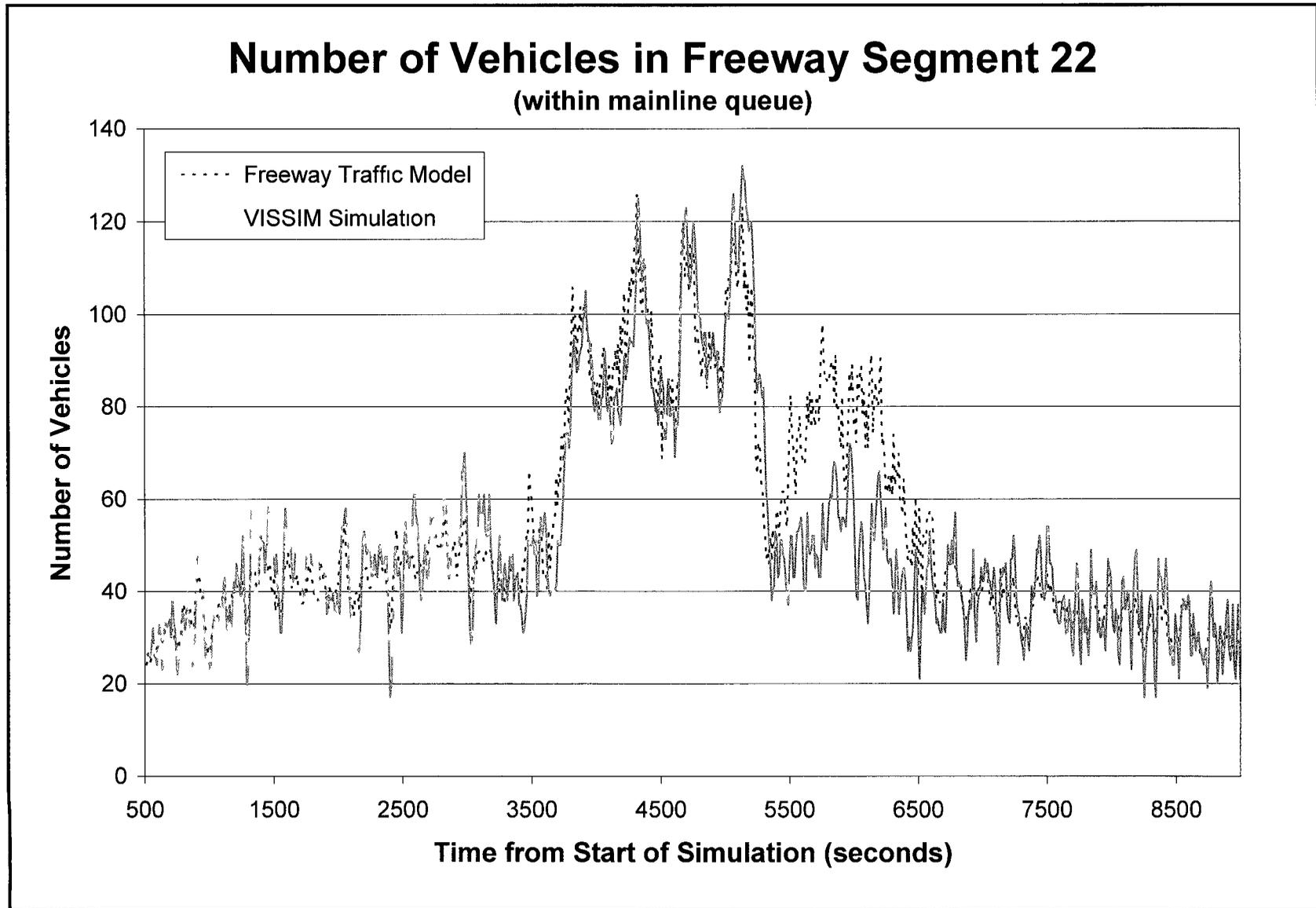
## Number of Vehicles in Freeway Segment 25 (downstream of bottleneck)

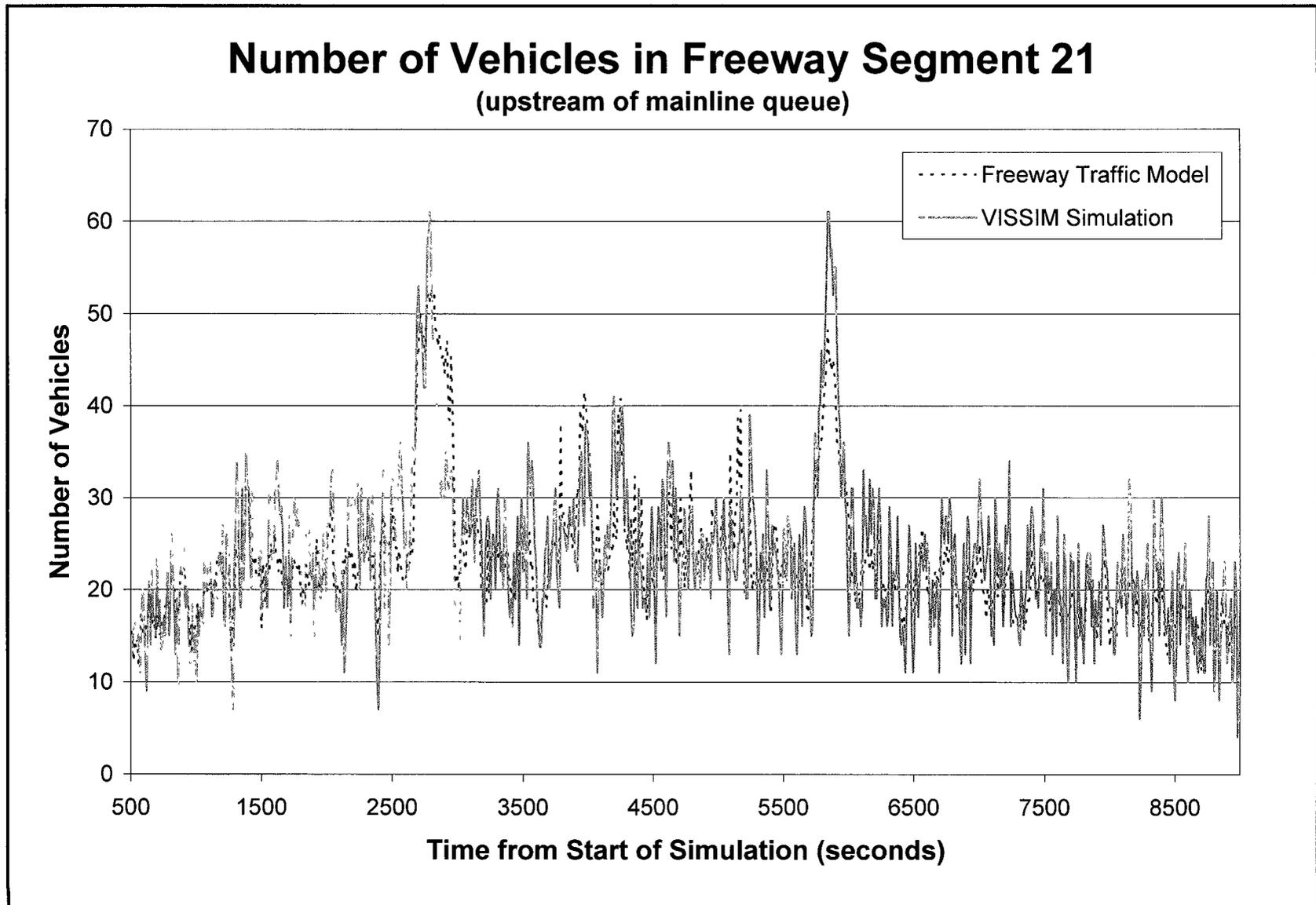


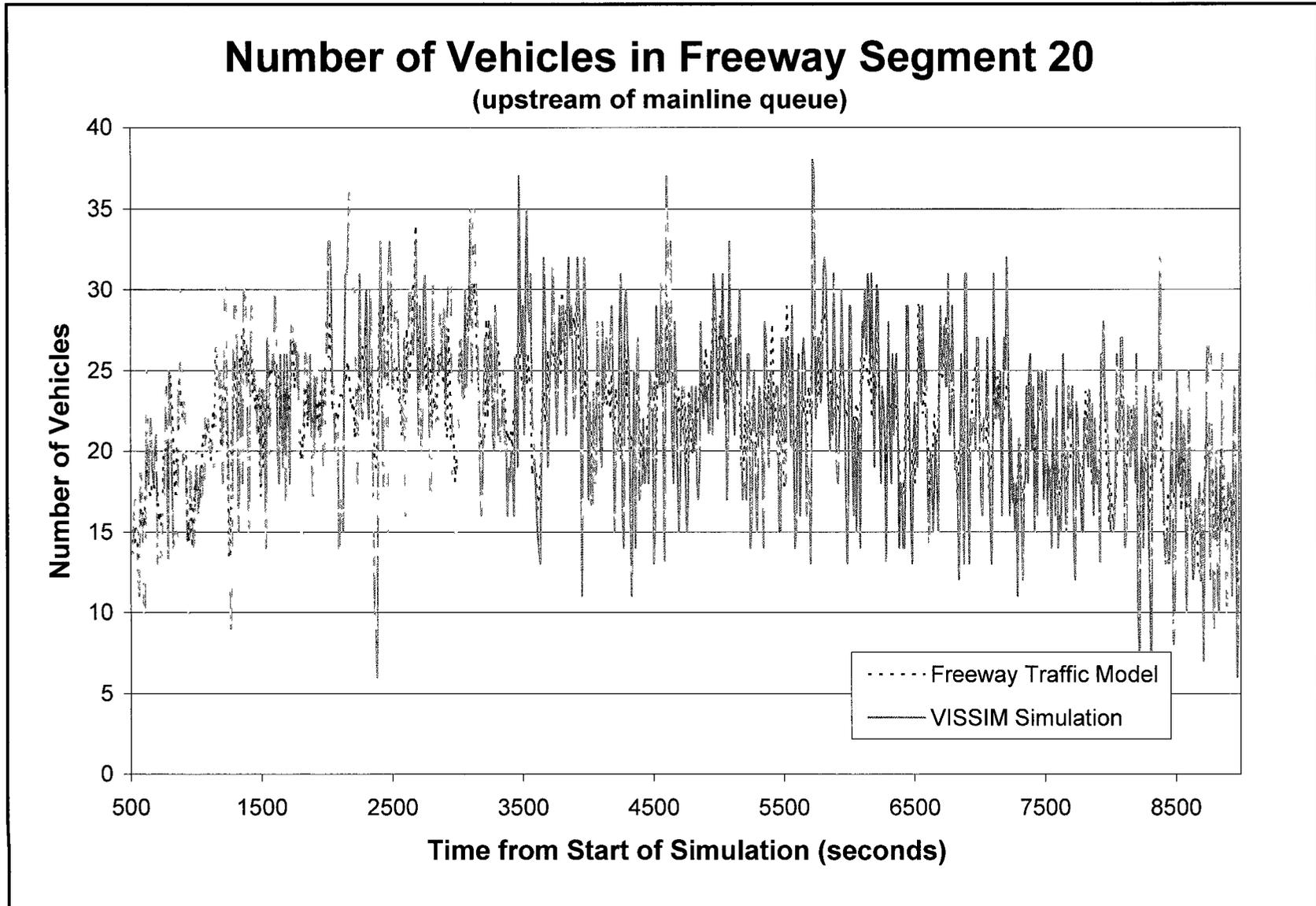
## Number of Vehicles in Freeway Segment 24 (at start of bottleneck)

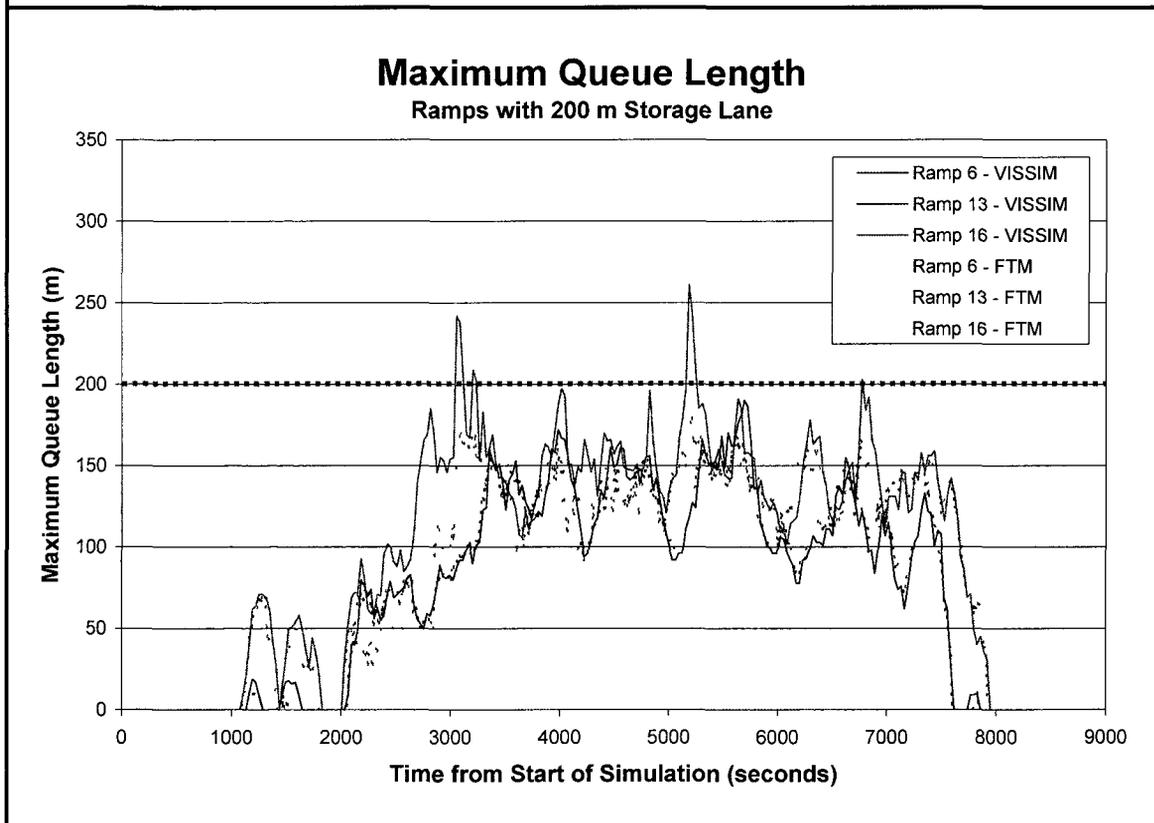
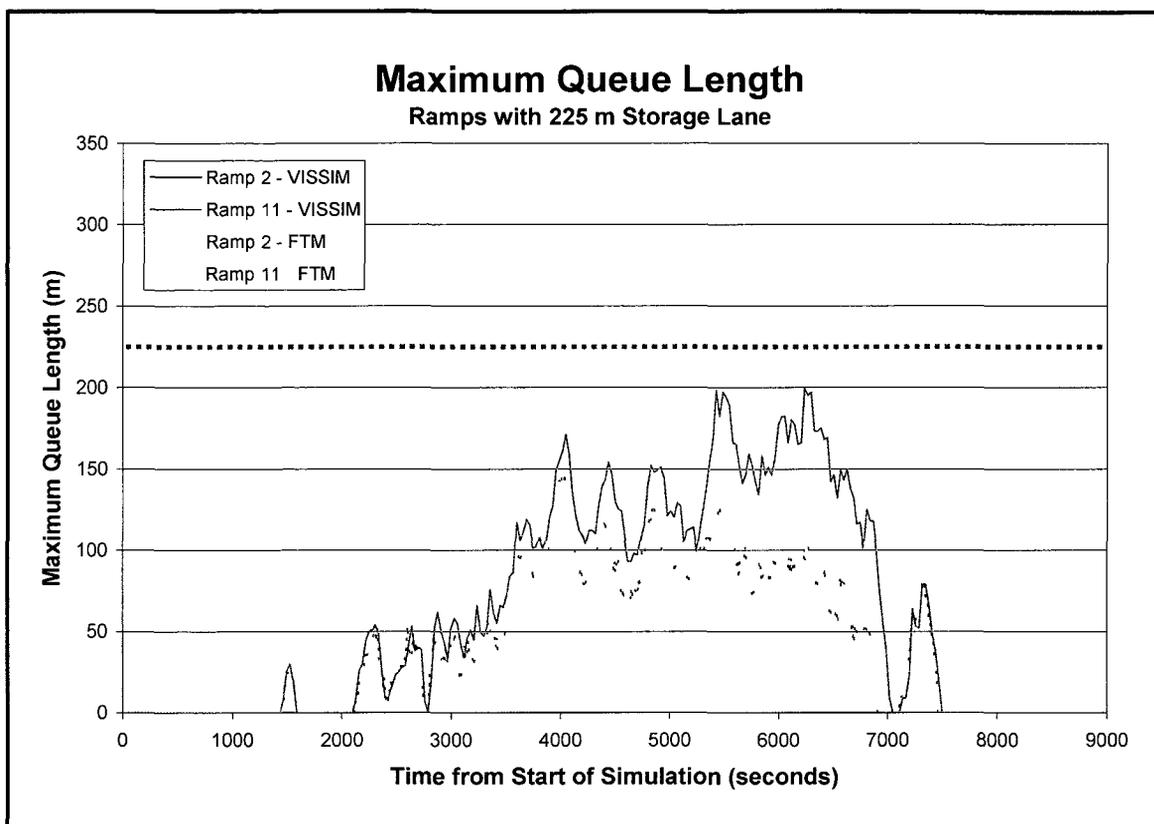


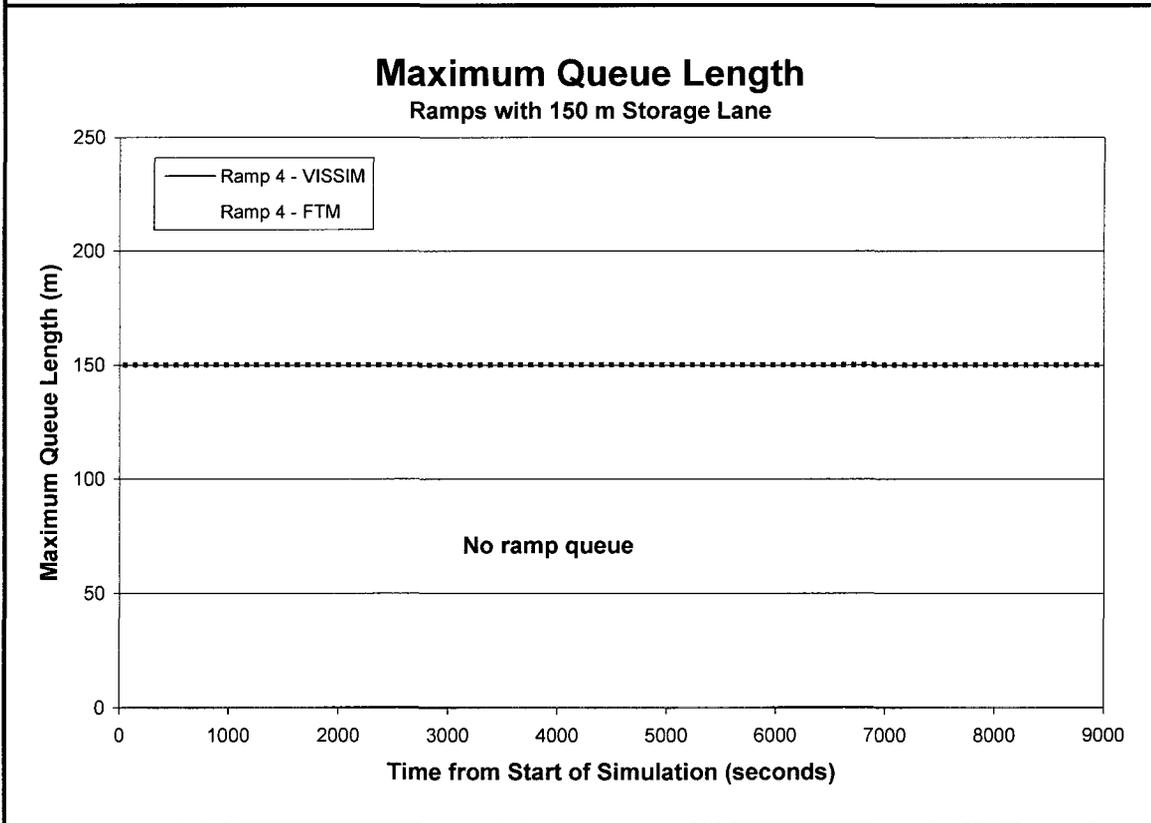
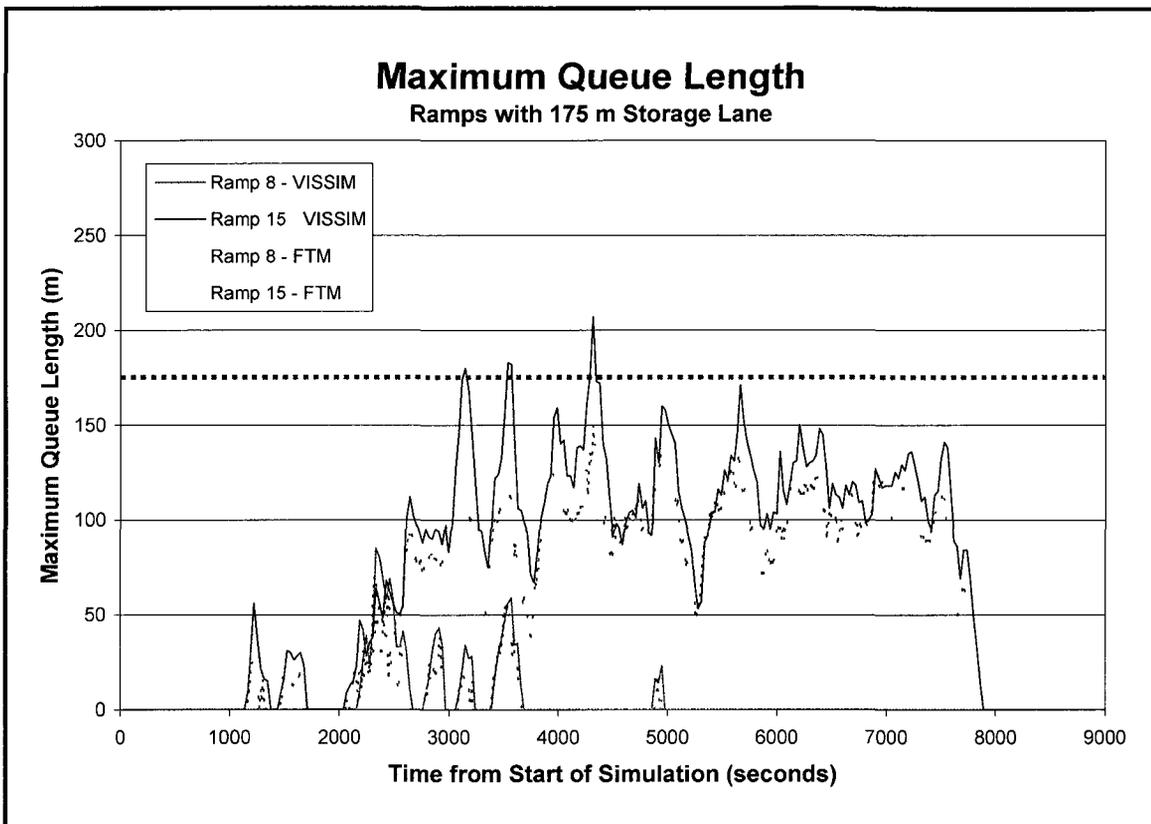


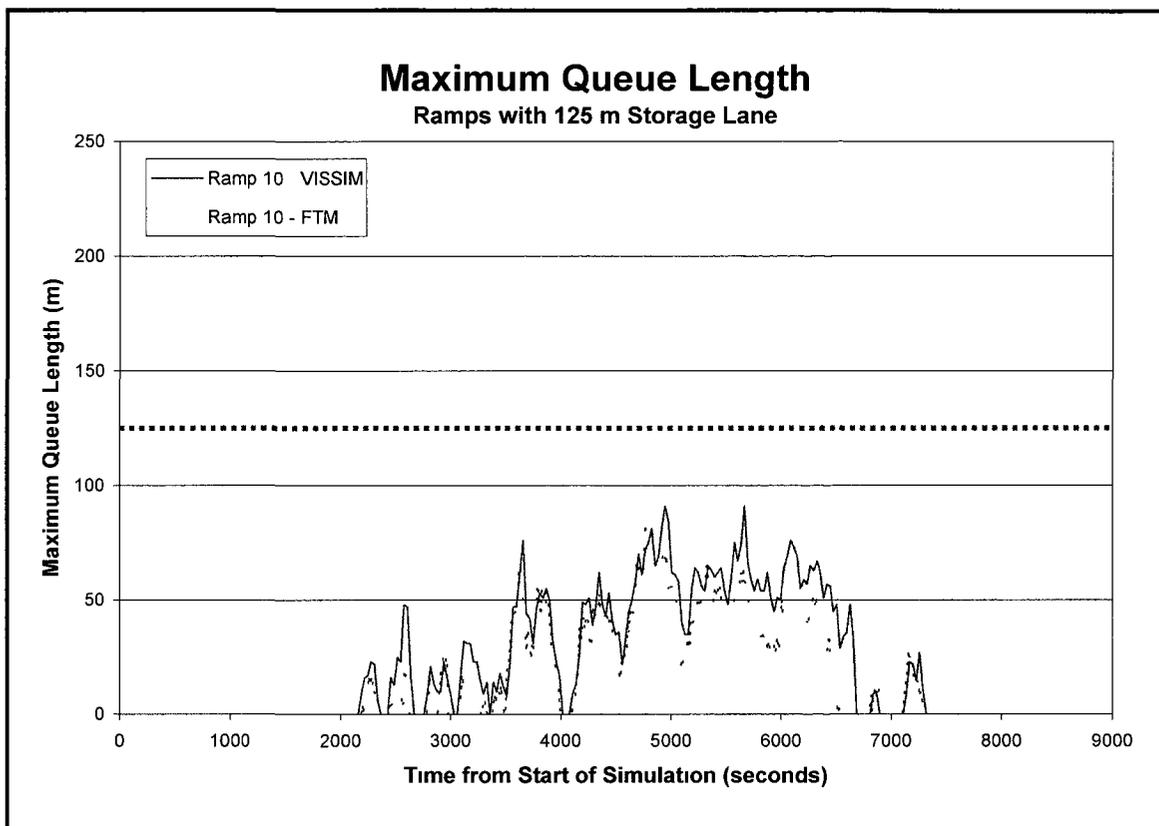






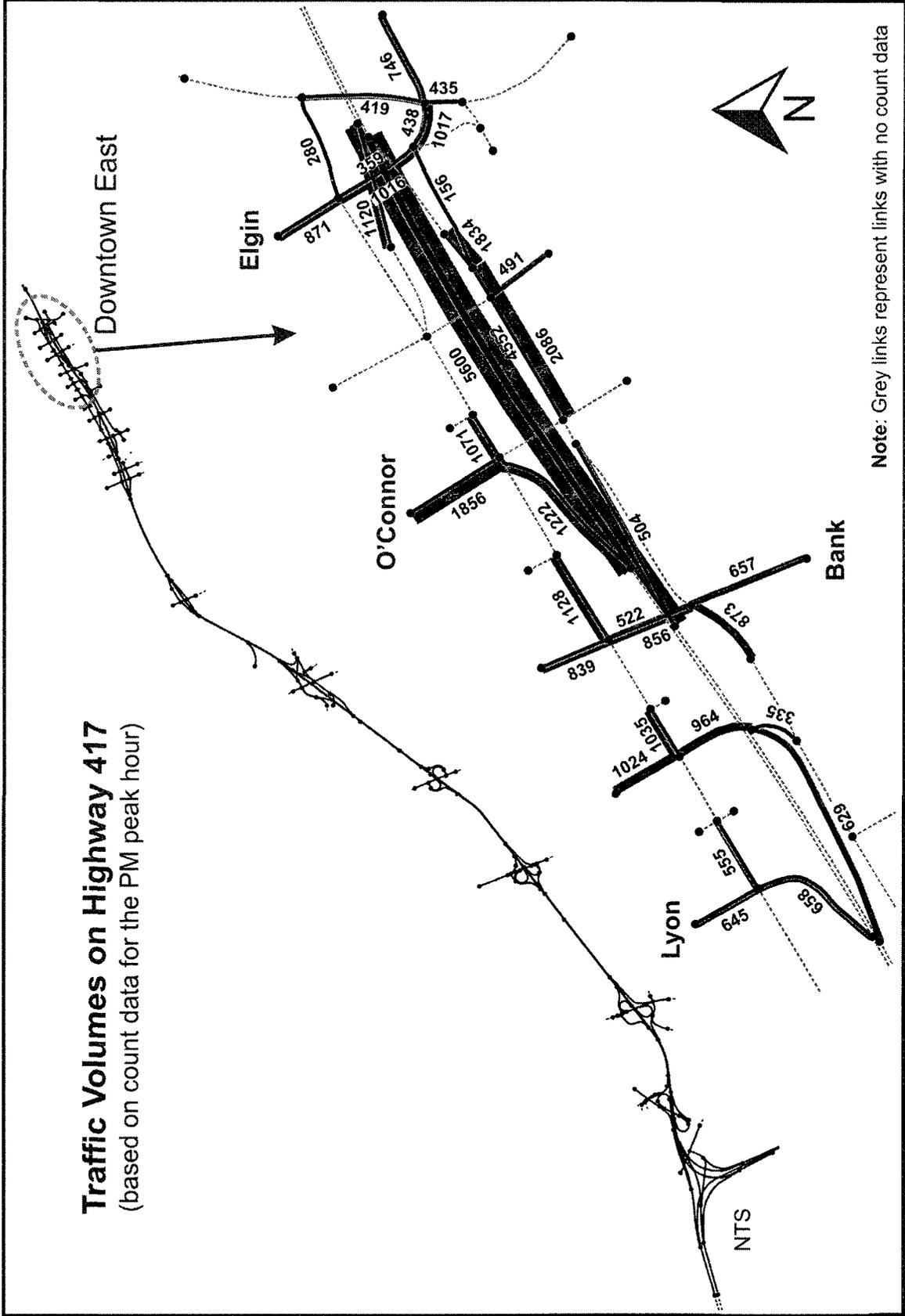




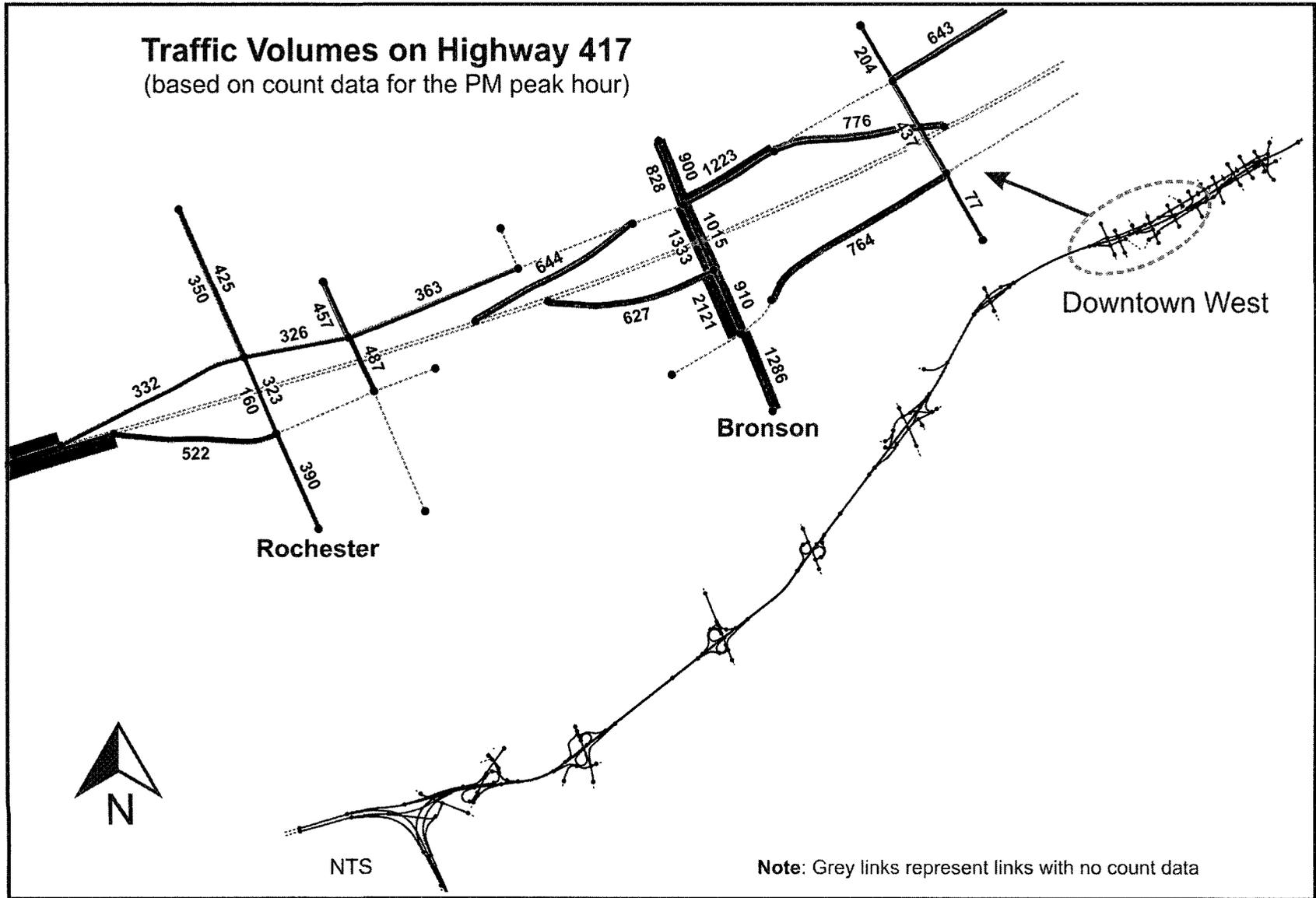


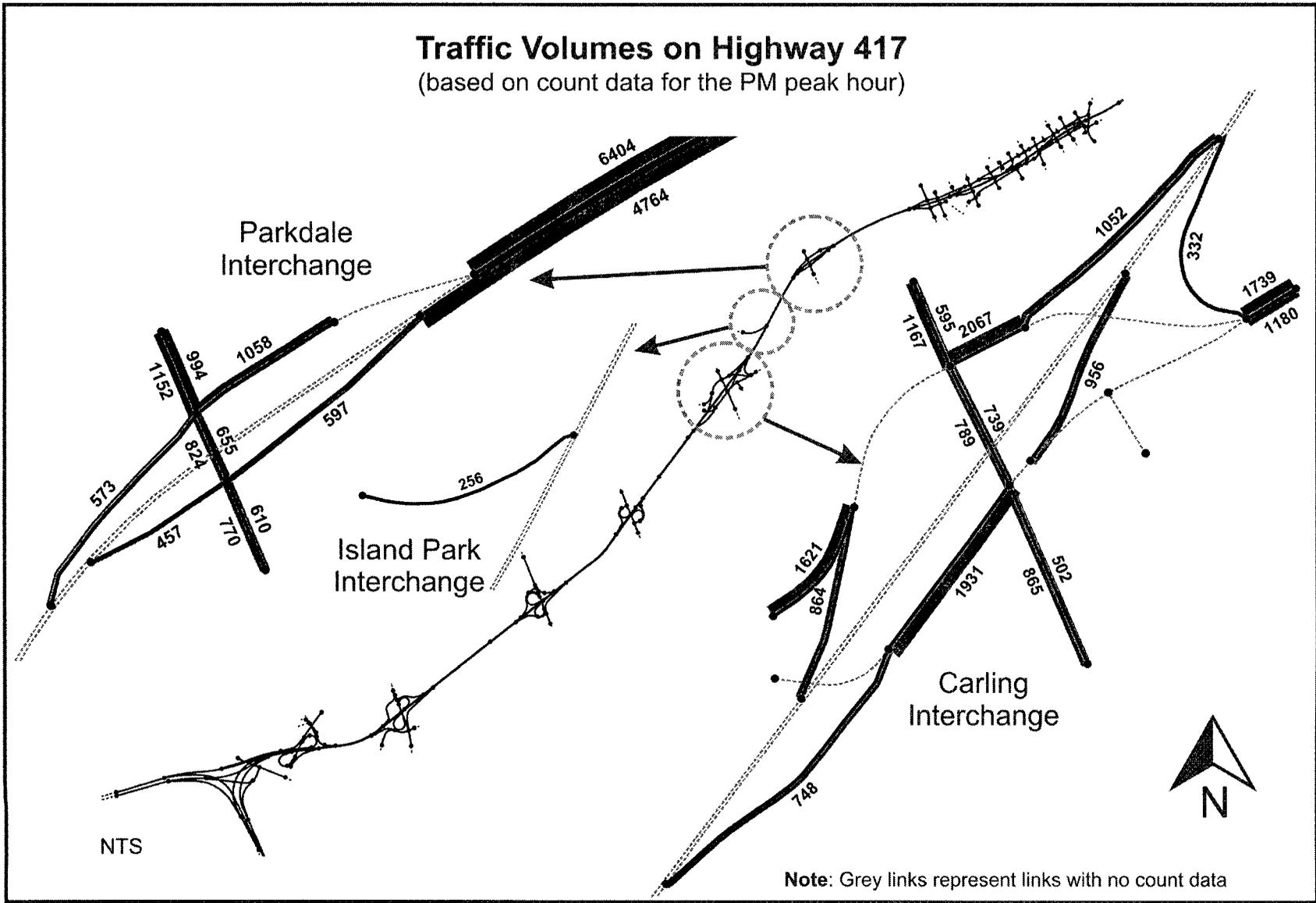
# **APPENDIX P**

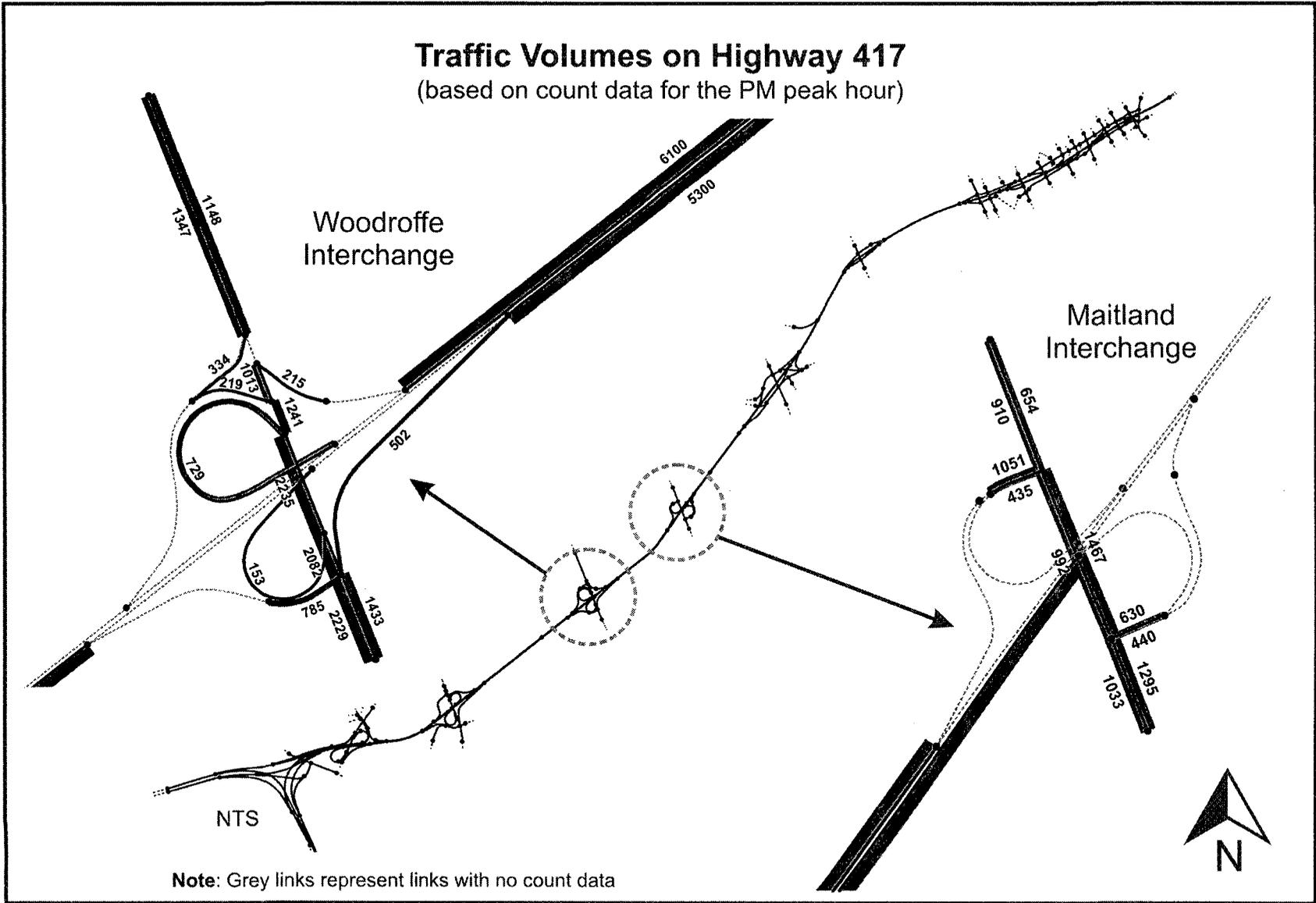
## **DEVELOPMENT & APPLICATION OF THE OTTAWA MODEL**

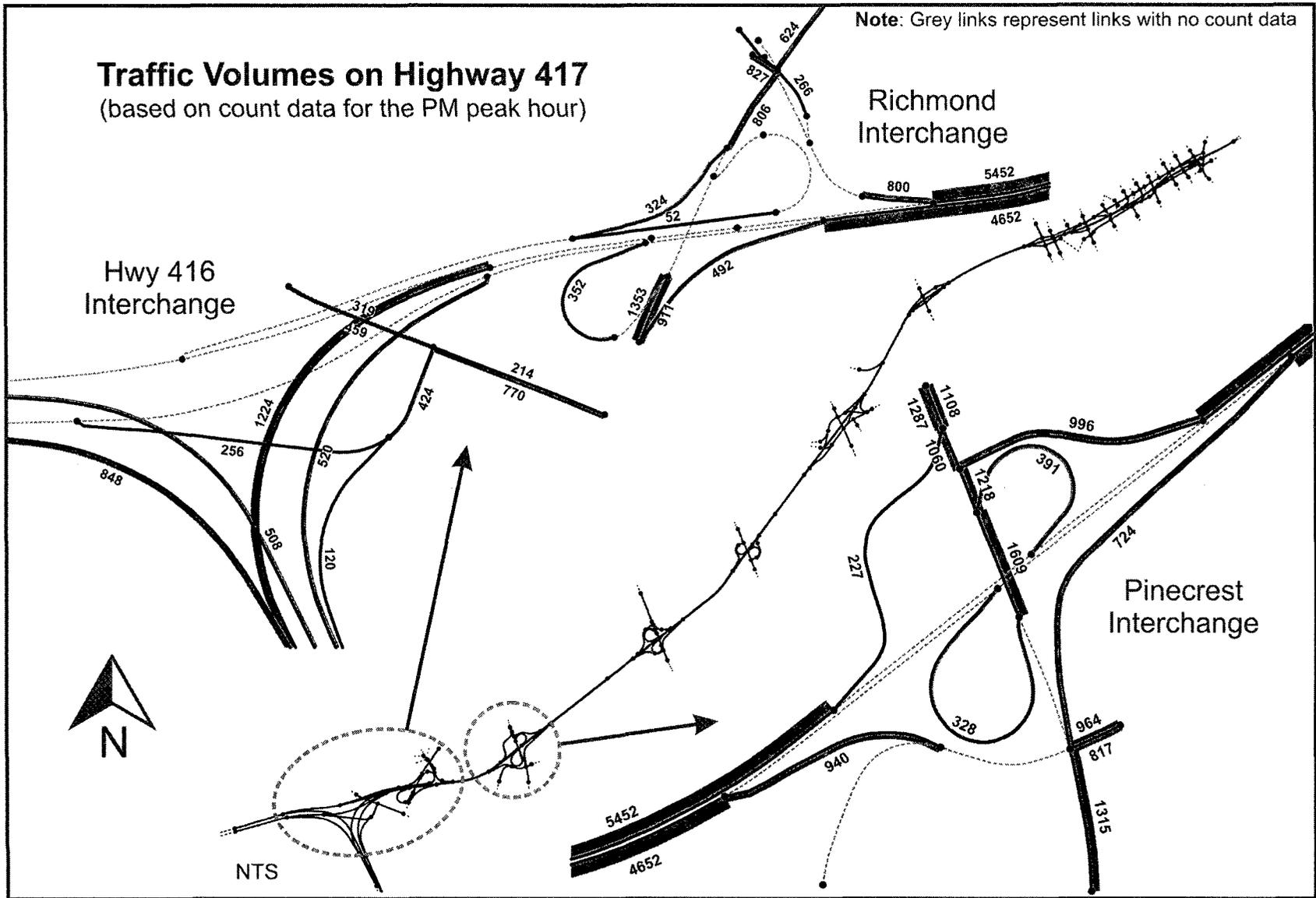


### Traffic Volumes on Highway 417 (based on count data for the PM peak hour)









## Ottawa Base Network Assumptions

### *General*

- Assumptions used in the Ottawa model are generally consistent with the VISSIM test network described in Appendix K
- Many of the VISSIM parameters are set at their default value (as defined in VISSIM 5.10-11), except as specifically noted

### *Simulation Parameters*

- Simulation resolution set to 10 time steps per simulation second

### *Vehicle Characteristics*

- VISSIM default values used for all vehicle characteristics
- Transit operations not considered
- Heavy vehicles assumed to represent 2.5% of the total travel demand, except on Highway 417 where a 5% truck percentage was assumed

### *Intersection Signal Control*

- Signal timing plans based on data from the City of Ottawa
- Signals initially coded using the simple NEMA editor, and later converted to the “Ring Barrier Controller” format in order to support a controller frequency of 10, similar to that used in the VAP logic for the ramp meters (the controller frequency refers to the number of passes through the control logic per simulation second, and must be consistent for all signals within the model)
- Signals coded to operate under semi-actuated control, with detectors used to call the non-coordinated phases
- Impact of pedestrian activations not considered (at freeway interchanges, pedestrian activity is unlikely to control the signal timing)
- Reaction to amber signals based on the “continuous check” decision model

### *Route Closures*

- Implemented to address unrealistic freeway diversion causing gridlock in the arterial network
- Closures include:
  - Westbound Carling off-ramp to Westbound Carling on-ramp
  - Westbound Parkdale off-ramp to westbound Parkdale on-ramp

### *Link Characteristics*

- 3.5 m lanes
- Lane change / emergency stop position for link connectors generally set equal to the VISSIM defaults (5 m for the emergency stop distance and 350 m for the lane change distance). Key exceptions are noted below.<sup>1</sup>
  - Off-ramp connector at freeway-ramp junctions
    - Lane change distance set so that drivers become aware of a freeway exit roughly 1500 m before reaching it
    - Emergency stop position assumed to be located 100 m downstream from the start of the deceleration lane
  - Mainline connector at the downstream end of auxiliary speed change lanes (including both acceleration and deceleration lanes)
    - Lane change distance set to a value greater than the length of the auxiliary lane to deter through vehicles from using it
  - On-ramp connector at freeway-ramp junctions
    - Only an issue when ramp metering is implemented and the on-ramp is widened to two lanes for vehicle storage, dropping to one lane as the ramp enters the freeway
    - Lane change distance set to coincide with the location of the ramp meter so that vehicles queued upstream of the meter on the two-lane section do not respond to the downstream lane drop before passing the meter (otherwise, vehicles would not queue evenly in both storage lanes)
  - Intersection connector for turning movements with an auxiliary lane
    - Emergency stop position assumed to be equal to the length of the auxiliary storage lane minus 10 m. This assumption encourages drivers to join the end of the queue, rather than trying to merge into the queue closer to the stop bar and disrupting traffic flow (a particular issue when the storage lane is full)

### *Network Fine-Tuning*

- To improve the model calibration results, several small adjustments were made at certain locations to address specific issues. As examples:
  - At one or two locations, adjustments were made to the treatment of right-turns-on-red to improve the movement capacity
  - Priority rules were added in several instances to prevent intersection blockages and improve weaving operations over short arterial segments
  - Lane change / emergency stop distances were modified to address unrealistic weaving behaviour on selected intersection approaches

---

<sup>1</sup> In most cases, the assumptions are similar to those adopted in the VISSIM test network. For additional information on the rationale for these assumptions, refer to Table K-2 in Appendix K.

## Driving Behaviour Parameters – Freeways

**Driving Behavior Parameter Sets**

No.: 3 Name: Freeway (free lane selection)

Following | Lane Change | Lateral | Signal Control

Look ahead distance  
 min.: 0.00 m  
 max.: 250.00 m  
 6 Observed vehicles

Look back distance  
 min.: 0.00 m  
 max.: 150.00 m

Temporary lack of attention  
 Duration: 1.00 s  
 Probability: 1.00 %

Car following model  
 Wiedemann 99

Model parameters

CC0 (Standstill Distance):	1.50	m
CC1 (Headway Time):	1.10	s
CC2 ('Following' Variation):	4.00	m
CC3 (Threshold for Entering 'Following'):	-8.00	
CC4 (Negative 'Following' Threshold):	-0.50	
CC5 (Positive 'Following' Threshold):	0.50	
CC6 (Speed dependency of Oscillation):	11.44	
CC7 (Oscillation Acceleration):	0.25	m/s <sup>2</sup>
CC8 (Standstill Acceleration):	3.50	m/s <sup>2</sup>
CC9 (Acceleration at 80 km/h):	1.50	m/s <sup>2</sup>

OK Cancel

**Driving Behavior Parameter Sets**

No.: 3 Name: Freeway (free lane selection)

Following | Lane Change | Lateral | Signal Control

General behavior: Free Lane Selection

Necessary lane change (route)

	Own	Trailing vehicle
Maximum deceleration:	-4.00 m/s <sup>2</sup>	-3.00 m/s <sup>2</sup>
-1 m/s <sup>2</sup> per distance:	200.00 m	200.00 m
Accepted deceleration:	-1.00 m/s <sup>2</sup>	-0.50 m/s <sup>2</sup>

Waiting time before diffusion: 60.00 s

Min. headway (front/rear): 0.50 m

To slower lane if collision time above: 0.00 s

Safety distance reduction factor: 0.60

Maximum deceleration for cooperative braking: -7.50 m/s<sup>2</sup>

OK Cancel

### Driving Behaviour Parameters – Arterial & Collector Roads

**Driving Behavior Parameter Sets**

No	Name
1	Urban (motorized)
3	Freeway (free lane selection)
6	Urban - Major Weave

No: 1 Name: Urban (motorized)

Following | Lane Change | Lateral | Signal Control

Look ahead distance

min: 0 00 m  
max: 250 00 m

4 Observed vehicles

Look back distance

min: 0 00 m  
max: 150 00 m

Temporary lack of attention

Duration: 0 00 s  
Probability: 0 00 %

Car following model: Wiedemann 74

Model parameters

Average standstill distance	2 00 m
Additive part of safety distance	2 00
Multiplic part of safety distance	3 00

OK Cancel

**Driving Behavior Parameter Sets**

No	Name
1	Urban (motorized)
3	Freeway (free lane selection)
6	Urban - Major Weave

No: 1 Name: Urban (motorized)

Following | Lane Change | Lateral | Signal Control

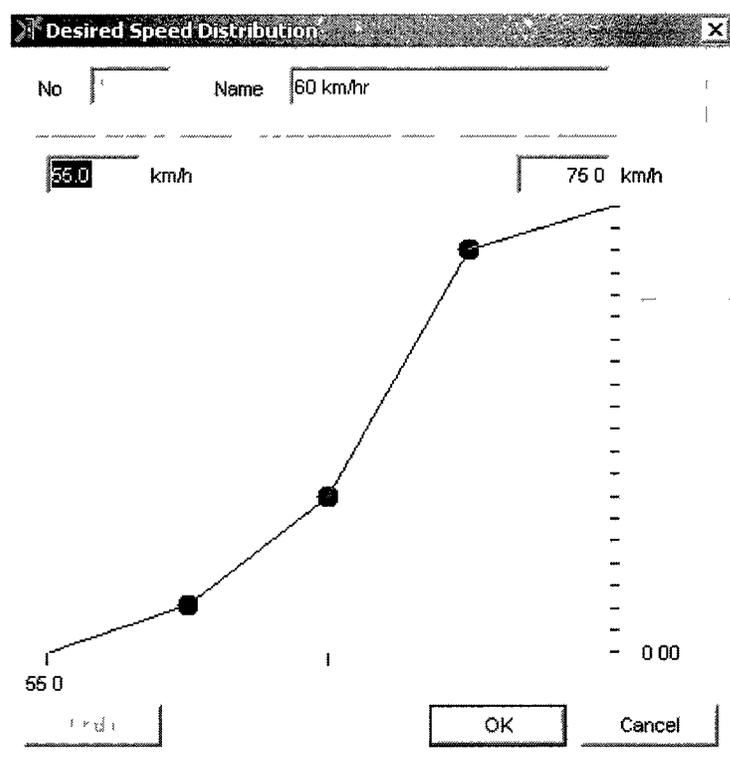
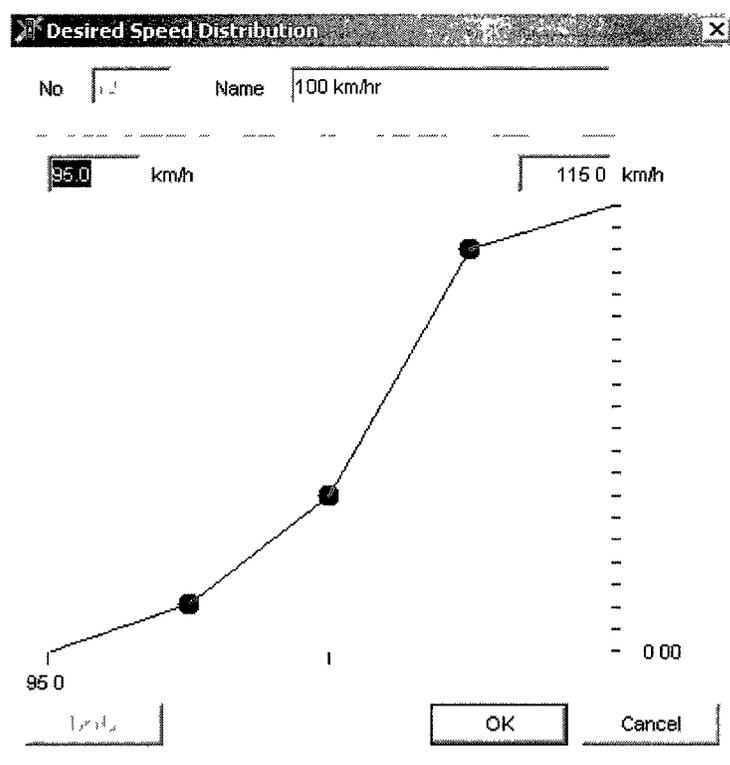
General behavior: Free Lane Selection

Necessary lane change (route)

	Own	Trailing vehicle
Maximum deceleration	-4 00 m/s <sup>2</sup>	-3 00 m/s <sup>2</sup>
-1 m/s <sup>2</sup> per distance	100 00 m	100 00 m
Accepted deceleration	-1 00 m/s <sup>2</sup>	-1 00 m/s <sup>2</sup>
Waiting time before diffusion		60 00 s
Min headway (front/rear)		0 50 m
To slower lane if collision time above		0 00 s
Safety distance reduction factor		0 60
Maximum deceleration for cooperative braking		-3 00 m/s <sup>2</sup>

OK Cancel

## Speed Assumptions



Typical Speed Distributions

- Right turn speed generally set at 15 km/hr
- Left turn speed generally set at 25 km/hr
- Speed distributions developed for each posted speed limit & applied as appropriate (i.e. wherever the speed limit changes)
- For the freeway mainline, a 110 km/hr speed distribution was applied (ranging from a minimum speed of 90 km/hr to a maximum speed of 125 km/hr)
- For off-ramps, an 80 km/hr speed distribution was applied near the ramp gore, followed by a speed distribution corresponding to the posted speed limit further downstream on the ramp
- In the majority of cases, there is no posted speed limit on the Highway 417 on-ramps within the model limits. Unless posted otherwise, a 50-60 km/hr speed distribution was generally applied on portions of the ramp with significant curvature, with a 70-80 km/hr speed distribution applied elsewhere. At a point roughly corresponding to the location of the ramp meter (i.e. 100-150 m upstream of the ramp gore), a 100 km/hr speed distribution was applied, allowing drivers to accelerate up to merging speeds before entering the freeway

## Dynamic Assignment Parameters

**Dynamic Assignment**

Trip chain file: ?.flt

Matrices

Traffic comp.	Matrix	
1, Cars&Trucks	Demand_Peak2.fma	Edit...
1, Cars&Trucks	Demand_Peak3.fma	New...
1, Cars&Trucks	Demand_Peak4.fma	Delete
1, Cars&Trucks	Demand_Unload1.fma	

Cost file: CostFile.bew  
 Check Edges

Path file: PathFile.weg  
 Check Edges

Archive files

Evaluation interval: 600 \$

Store costs Extended...

Search new paths Extended...

Store paths (and volumes)

Kirchhoff exponent: 3.50

Logit scaling factor: 1.50000

Logit lower limit: 0.00100

Scale Total Volume to 100.0 %

Correction of overlapping paths

Avoid Long Detours: 1.25

Use VISSIM's virtual memory

Convergence...

Route Guidance...

Create Static Routing

OK Cancel

**Edge Cost Values**

Exponential Smoothing With Smoothing Factor: 0.20

MSA (Method of Successive Averages), so far: iterations

OK Cancel

## Dynamic Assignment Parameters

**Path Search**

Reject paths with total costs higher by  % than the total cost of the best path

Limit number of paths to  paths per parking lot relation

Search paths for O-D pairs with zero volume

Stochastic edge penalization for alternative paths search

Spread

Passes

OK Cancel

**Convergence**

Travel Time on Paths  %

Travel Time on Edges  %

Volume on Edges  veh

OK Cancel

*Convergence based on one criterion only as recommended in the VISSIM User's Manual*

**Parking Lot Selection**

Decision Situation:

\* Parking Cost

+  \* Attraction

+  \* Distance from desired zone [m]

+  \* Distance from current position [m]

+  \* Current parking availability

= Efficiency

OK Cancel

**Cost Coefficients**

\* Travel Time [s]

+  \* Distance [m]

+  \* Link Cost

OK Cancel

*Coefficients for computing the generalized link cost. Since the link cost is zero for all links, the attractiveness of each link is simply a function of the link travel time.*

*Used to compute which parking lot will serve as the trip destination for the rare situation where a zone has more than one parking lot. By setting the distance coefficient to non-zero, the model will tend to assign vehicles to the closest parking lot.*

Convergence Evaluation

File: C:\Network Drive\Ottawa\_Model\VISSIM\_ClipppedModel\Ottawa\_FinalMay9.inp  
 Date: Thursday, May 12, 2011 3:19:24 AM  
 VISSIM: 5.10-11 [21194]

TimeFrom (Class from)	TimeTo (Class to)	Volume Difference								
		0	2	5	10	25	50	100	250	500
		2	5	10	25	50	100	250	500	~
Edges:										
0	600	232	198	159	157	15	0	0	0	0
600	1200	174	143	186	189	66	3	0	0	0
1200	1800	178	177	166	178	44	18	0	0	0
1800	2400	200	165	187	150	49	10	0	0	0
2400	3000	182	152	180	189	46	12	0	0	0
3000	3600	217	165	190	157	26	6	0	0	0
3600	4200	189	155	155	201	52	9	0	0	0
4200	4800	208	142	152	175	65	12	7	0	0
4800	5400	167	186	186	164	28	18	12	0	0
5400	6000	192	142	180	180	47	17	2	1	0
6000	6600	216	158	157	174	52	4	0	0	0
6600	7200	209	181	153	152	55	11	0	0	0
Paths:										
0	600	1519	78	47	24	9	7	2	1	0
600	1200	1470	113	54	27	13	5	4	1	0
1200	1800	1507	103	37	18	11	5	5	1	0
1800	2400	1504	103	40	18	11	6	4	1	0
2400	3000	1513	116	54	31	13	5	5	1	0
3000	3600	1539	113	44	20	11	5	5	1	0
3600	4200	1551	100	50	15	11	6	4	1	0
4200	4800	1515	124	45	34	10	5	4	1	0
4800	5400	1550	99	51	17	10	6	4	1	0
5400	6000	1562	91	43	21	11	7	2	1	0
6000	6600	1528	113	53	33	6	8	2	1	0
6600	7200	1573	98	34	20	9	7	3	0	0

TimeFrom (Class from)	TimeTo (Class to)	Travel Time Difference											
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	200%
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	200%	~
Edges:													
0	600	689	12	2	0	2	0	0	0	0	0	0	0
600	1200	689	12	1	1	0	0	0	0	0	0	0	2
1200	1800	690	12	2	1	0	0	0	0	0	0	0	0
1800	2400	679	23	1	0	1	0	1	0	0	0	0	0
2400	3000	665	32	5	1	0	0	0	0	0	0	0	2
3000	3600	650	46	4	3	0	2	0	0	0	0	0	0
3600	4200	657	42	4	2	0	0	0	0	0	0	0	0
4200	4800	649	48	3	2	2	0	1	0	0	0	0	0
4800	5400	640	59	2	1	0	0	0	0	0	1	2	0
5400	6000	635	64	4	1	1	0	0	0	0	0	0	0
6000	6600	638	62	3	0	2	0	0	0	0	0	0	0
6600	7200	618	78	6	0	2	0	0	1	0	0	0	0
Paths:													
0	600	1687	0	0	0	0	0	0	0	0	0	0	0
600	1200	1687	0	0	0	0	0	0	0	0	0	0	0
1200	1800	1687	0	0	0	0	0	0	0	0	0	0	0
1800	2400	1686	1	0	0	0	0	0	0	0	0	0	0
2400	3000	1737	1	0	0	0	0	0	0	0	0	0	0
3000	3600	1732	6	0	0	0	0	0	0	0	0	0	0
3600	4200	1724	14	0	0	0	0	0	0	0	0	0	0
4200	4800	1726	12	0	0	0	0	0	0	0	0	0	0
4800	5400	1694	44	0	0	0	0	0	0	0	0	0	0
5400	6000	1659	79	0	0	0	0	0	0	0	0	0	0
6000	6600	1624	120	0	0	0	0	0	0	0	0	0	0
6600	7200	1596	148	0	0	0	0	0	0	0	0	0	0

TimeFrom	TimeTo	Duality gap
0	600	0.00
600	1200	0.00
1200	1800	0.00
1800	2400	0.00
2400	3000	0.00
3000	3600	0.00
3600	4200	0.00
4200	4800	0.00
4800	5400	0.00
5400	6000	0.00
6000	6600	0.00
6600	7200	0.00

Figure P-1 Dynamic Assignment Convergence Results – Base Network (no metering)

**Table P-1 Peak Hour Volume Calibration Targets Adopted in this Study**

Criteria and Measures	Calibration Acceptance Targets
<b>Hourly Flows, Model Versus Observed</b>	
Individual Link Flows	
Within 15%, for 700 veh/h < Flow < 2700 veh/h	> 85% of cases
Within 100 veh/h, for Flow < 700 veh/h	> 85% of cases
Within 400 veh/h, for Flow > 2700 veh/h	> 85% of cases
Sum of All Link Flows	Within 5% of sum of all link counts
GEH Statistic < 5 for Individual Link Flows*	> 85% of cases
GEH Statistic for Sum of All Link Flows	GEH < 4 for sum of all link counts

\*The GEH statistic is computed as follows:

$$GEH = \sqrt{\frac{(E - V)^2}{(E + V) / 2}}$$

where:

E = model estimated volume

V = field count

**Adapted from:** *FHWA Traffic Analysis Toolbox*, Volume III (Dowling et al. 2004), which cites the original source as: "Freeway System Operational Assessment," *Paramics Calibration and Validation Guidelines* (Draft), Technical Report I-33, Wisconsin DOT, District 2, June 2002.

**Table P-2 Calibration Assessment: Comparison of Simulated and Observed Peak Hour Volumes**

Measurement		Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Average	Required Value <sup>1</sup>	Target Met?
Individual Links / Turns	<b>Volume Difference &lt; Criteria - Links</b>													
	% Links Passing													
	Volume < 700 (68 links)	90%	93%	90%	90%	90%	94%	88%	91%	93%	90%	91%	>85%	Yes
	700 < Volume < 2700 (81 links)	90%	89%	89%	89%	89%	90%	88%	90%	88%	85%	89%	>85%	Yes
	Volume > 2700 (12 links)	67%	67%	67%	67%	67%	67%	67%	67%	67%	67%	67%	>85%	No <sup>2</sup>
	Overall links (161 links)	88%	89%	88%	87%	88%	90%	86%	89%	88%	86%	88%	N/A	N/A
	<b>Volume Difference &lt; Criteria - Turns</b>													
	% Turns Passing													
	Volume < 700 (131 turns)	86%	87%	87%	86%	86%	86%	86%	86%	86%	85%	86%	N/A	N/A
	700 < Volume < 2700 (31 turns)	81%	84%	84%	84%	84%	84%	81%	84%	81%	84%	83%	N/A	N/A
	Overall turns (162 turns)	85%	86%	86%	86%	86%	86%	85%	86%	85%	85%	86%	N/A	N/A
	<b>GEH &lt; 5</b>													
	% Links Passing	86%	86%	85%	86%	86%	88%	85%	86%	86%	83%	86%	>85%	Yes
% Turns Passing	79%	78%	79%	79%	80%	80%	78%	79%	80%	78%	79%	N/A	N/A	
<b>Correlation Coefficient</b>														
Link Volumes	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	N/A	N/A
Turn Volumes	0.976	0.977	0.977	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.975	0.976	N/A	N/A
Combined Volumes	<b>Sum of Link Volumes</b>													
	Absolute % Difference	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<5%	Yes
	GEH Statistic	0.6	0.3	3.3	0.4	1.1	1.6	0.5	1.0	1.8	0.7	1.1	<4	Yes
	<b>Sum of Turn Volumes</b>													
	Absolute % Difference	2%	2%	2%	2%	2%	1%	2%	2%	2%	3%	2%	N/A	N/A
GEH Statistic	5.3	5.4	4.6	4.8	4.7	3.8	5.3	5.1	4.1	7.0	5.0	N/A	N/A	

<sup>1</sup> Required value to meet calibration target. N/A indicates that no calibration target applies (results provided for information only)

<sup>2</sup> Although the target is not met, the number of links in this category is quite low. Upon review of the count data, it was concluded that certain counts may be subject to error.

Results from this assessment based on link volumes and turning movements presented for illustrative purposes only. Results suggest that calibration results for link volumes are acceptable. Turning movement volumes are typically more difficult to replicate and are therefore expected to have poorer calibration results. Although this trend holds, the drop in accuracy is not substantial and the results suggest that the model also does an acceptable job at simulating turns.

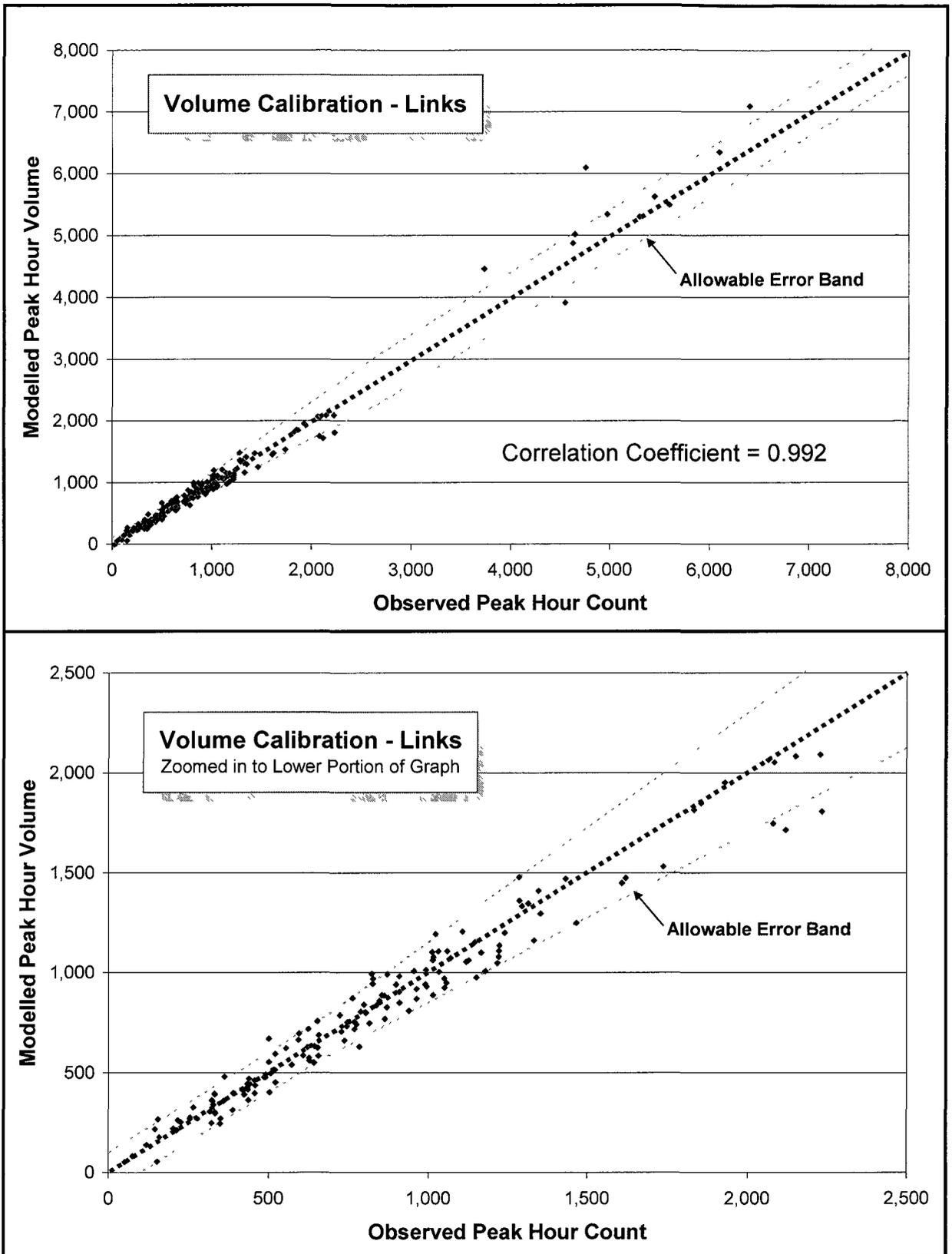


Figure P-2 Typical Peak Hour Calibration Results – Links (Run 8)

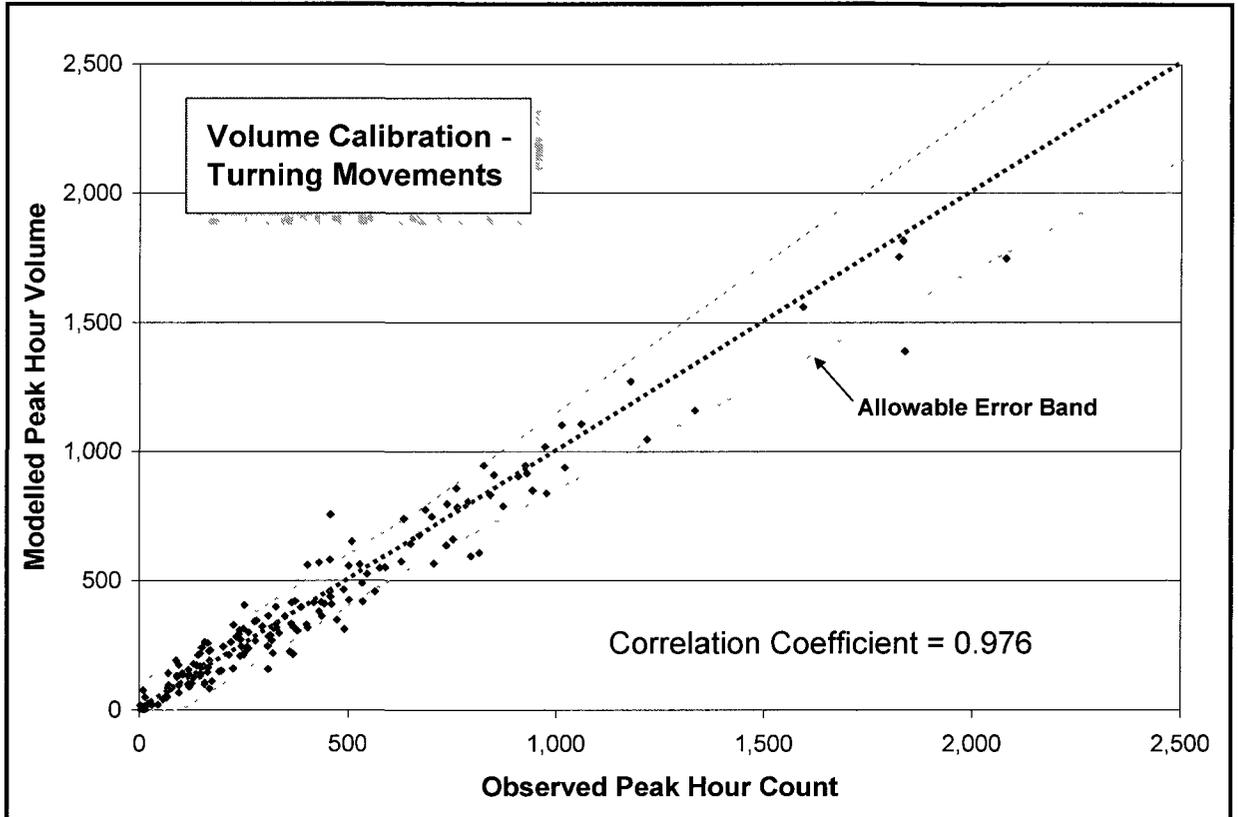


Figure P-3 Typical Peak Hour Calibration Results – Turning Movements (Run 8)

**Table P-3 Summary of Network Performance – Base Ottawa Network with No Ramp Metering**

<b>Performance Measure</b>	<b>Run 1</b>	<b>Run 2</b>	<b>Run 3</b>	<b>Run 4</b>	<b>Run 5</b>	<b>Run 6</b>	<b>Run 7</b>	<b>Run 8</b>	<b>Run 9</b>	<b>Run 10</b>	<b>Average</b>
# of vehicles that have left the network	75,037	75,417	75,773	75,305	75,528	75,608	75,097	75,620	75,625	75,472	75,448
# of vehicles in network at end of simulation	3881	3450	3096	3678	3503	3385	3758	3299	3303	3225	3458
Total distance traveled [km]	362,266	365,024	365,958	363,308	365,876	365,961	362,588	366,736	366,490	366,588	365,079
Total travel time [h]	6589	6126	5800	6351	6136	5969	6409	5934	5912	5970	6120
Average speed [km/h]	55.0	59.6	63.1	57.2	59.6	61.3	56.6	61.8	62.0	61.4	59.8
Total delay time [h]	2806	2318	1981	2558	2320	2150	2624	2109	2090	2147	2310
Average delay time per vehicle [s]	128	106	90	117	106	98	120	96	95	98	105
Number of stops	260,231	185,333	169,139	228,603	190,537	178,457	238,732	173,973	164,754	161,859	195,162
Average number of stops per vehicle	3.3	2.4	2.1	2.9	2.4	2.3	3.0	2.2	2.1	2.1	2.5
Total stopped delay [h]	546.4	499.3	483.5	528.2	512.2	487.2	532.1	490.8	485.2	486.1	505.1
Average stopped delay per vehicle [s]	24.9	22.8	22.1	24.1	23.3	22.2	24.3	22.4	22.1	22.2	23.0

\* Excludes the first 500 seconds of the simulation while the network is being loaded

Typical Congestion Profile – Highway 417 Westbound, No Metering (Run 2)

### Highway 417 Westbound - Congestion Profile

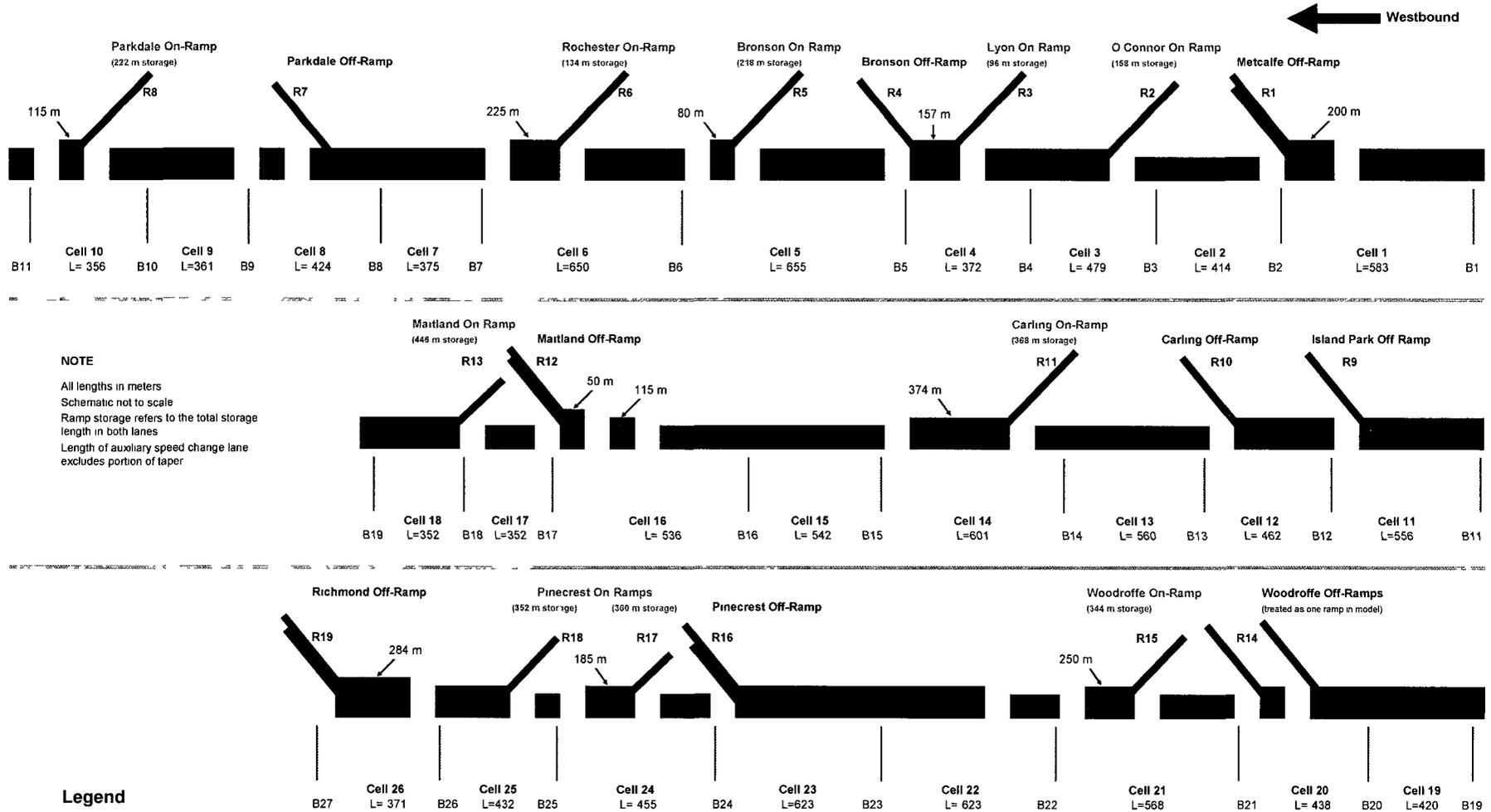
Link Length	Description	300	600	900	1200	1500	1800	2100	2400	2700	3000	3300	3600	3900	4200	4500	4800	5100	5400	5700	6000	6300	6600	6900	7200	
1495	100	97	100	98	101	100	100	101	99	97	97	98	96	96	91	97	95	100	97	98	98	100	100	97	100	100
1495	100	98	100	98	101	99	98	100	98	98	98	98	96	96	91	97	95	100	97	98	98	100	100	97	100	100
1495	100	97	100	101	100	100	98	100	97	97	97	98	96	96	91	97	95	100	97	98	98	100	100	97	100	100
1495	100	100	100	100	100	100	93	98	97	96	96	96	93	93	88	100	97	100	98	98	98	100	100	100	100	100
1495	93	101	102	100	98	98	97	98	97	97	97	97	97	97	92	100	98	100	98	98	98	100	100	100	100	100
1496	100	97	100	101	98	98	97	98	97	97	97	97	97	97	92	100	98	100	98	98	98	100	100	100	100	100
1496	100	98	100	98	98	98	98	98	98	98	98	98	98	98	92	100	98	100	98	98	98	100	100	100	100	100
12510	54	98	100	98	90	92	98	97	97	97	97	97	97	97	92	100	98	100	98	98	98	100	100	100	97	100
896	100	98	100	97	94	93	97	98	98	97	97	97	97	97	92	100	98	100	98	98	98	100	100	100	97	100
896	100	101	105	100	98	94	98	98	98	98	98	98	98	98	92	100	98	100	98	98	98	100	100	100	100	100
896	100	100	105	101	101	98	98	98	97	101	98	100	98	97	101	101	103	104	98	97	101	105	105	104	100	100
896	100	100	105	102	102	99	100	100	97	102	98	102	98	100	100	103	103	104	100	97	101	105	105	104	100	100
896	100	100	105	103	102	101	103	102	97	102	93	102	101	101	98	104	104	104	101	98	105	105	105	104	100	100
12428	52	100	105	103	103	102	104	102	99	102	98	102	102	100	98	103	105	104	102	98	105	105	105	104	100	100
892	100	101	103	101	97	99	102	101	98	98	94	98	101	98	97	97	101	101	101	97	103	103	101	102	100	100
892	100	102	104	100	98	100	103	100	100	97	100	102	97	100	97	103	101	102	100	105	104	102	103	100	100	100
892	100	103	105	104	100	102	102	104	102	101	100	101	103	101	100	100	104	100	102	102	105	105	104	105	100	100
892	100	104	105	103	102	103	104	104	102	102	102	102	102	102	98	99	104	101	104	103	104	105	105	105	100	100
892	80	103	104	104	103	100	102	100	101	101	102	101	102	98	98	104	102	104	100	98	105	105	105	100	100	100
12333	30	101	102	102	102	101	103	100	98	98	101	98	99	100	98	102	101	102	98	98	102	102	103	103	100	100
891	100	97	94	97	98	98	98	94	93	92	94	93	93	93	88	97	98	97	97	97	97	98	98	98	98	100
891	57	98	94	98	98	98	98	98	98	92	90	98	91	92	92	97	98	98	98	98	98	98	98	98	98	98
12312	37	100	97	98	101	98	98	98	98	97	93	98	98	97	98	98	98	98	98	98	98	98	98	98	98	98
881	100	101	100	101	102	100	101	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
881	100	100	102	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
881	100	100	103	105	104	104	104	104	98	103	101	97	104	104	100	100	100	100	100	100	100	100	100	100	100	100
881	100	100	104	105	104	105	105	104	98	103	100	98	104	104	100	100	100	100	100	100	100	100	100	100	100	100
881	100	100	105	105	100	100	100	100	98	103	100	98	104	104	100	100	100	100	100	100	100	100	100	100	100	100
12159	49	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
883	80	98	92	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
884	100	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
884	100	97	94	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
884	100	101	100	102	98	97	92	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
884	52	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
11958	74	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
872	100	100	102	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
872	100	100	101	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
872	25	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
873	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100







### Highway 417 Cell Configuration for the Freeway Traffic Model



**NOTE**  
 All lengths in meters  
 Schematic not to scale  
 Ramp storage refers to the total storage length in both lanes  
 Length of auxiliary speed change lane excludes portion of taper

**Legend**

- VISSIM Link
- VISSIM Connector
- Cell boundary with traffic sensor
- Cell boundary without traffic sensor

\* Traffic sensors also assumed to be located on all on- and off-ramps near the ramp gore

**Table P-4 Introduction of Ramp Metering on Highway 417: Summary of Key Assumptions used in the Ramp Control Algorithm**

Parameter	Assumed Value	Comments
Freeway Traffic Model Parameters <sup>1</sup>		
<ul style="list-style-type: none"> <li>Speed-density relationship for the uncongested travel speed</li> </ul>	$V_{ff} = 115$ $\alpha = -0.92$	Represents a small change from the VISSIM test network, which employed values of $V_{ff} = 112$ and $\alpha = -0.54$
<ul style="list-style-type: none"> <li>Minimum speed</li> </ul>	Gaussian Distribution Mean = 15 km/hr Std Dev = 3 km/hr	Test network also employed a Gaussian distribution, however, the mean was assumed to be slightly higher at 20 km/hr, and the standard deviation was also slightly higher at 5 km/hr
<ul style="list-style-type: none"> <li>Density within the traffic queue</li> </ul>	Gaussian Distribution Mean = 60 veh/km/lane Std Dev = 8 veh/km/lane	Test network employed a similar distribution, however, the mean was set at 58 veh/km/lane
<ul style="list-style-type: none"> <li>Noise term for speed measurements</li> </ul>	Gaussian Distribution Mean = 0 km/hr Std Dev = 6 km/hr	The sensor noise was assumed to be slightly lower than that used in the VISSIM test network. With less uncertainty in the sensor readings, tracking performance should be enhanced, however, more particles will be discarded as being inconsistent with the evidence
<ul style="list-style-type: none"> <li>Congested capacity</li> </ul>	Initial distribution assumed to have a mean of 2000 vphpl	In general, the mainline capacity on Highway 417 is lower than that used in the VISSIM test network due to changes in the driver behaviour parameters that were introduced to improve the model calibration results (in the test network, the initial distribution for the congested capacity had a mean of 2300 vphpl)
<ul style="list-style-type: none"> <li>Initial freeway density</li> </ul>	Gaussian Distribution Mean = 17 veh/km/lane Std Dev = 3 veh/km/lane	Reflects higher traffic loading when the algorithm is initialized, resulting in higher density
<ul style="list-style-type: none"> <li>Initial freeway speed</li> </ul>	Gaussian Distribution Mean = 99 km/hr Std Dev = 2 km/hr	Reflects higher traffic loading when the algorithm is initialized, resulting in lower speed
<ul style="list-style-type: none"> <li>Flow breakdown model</li> </ul>	Same as used in the VISSIM test network	The breakdown model only applies to on-ramps with an auxiliary speed change lane. For lane additions, it was assumed that the probability of breakdown is negligible

<b>Parameter</b>	<b>Assumed Value</b>	<b>Comments</b>
Data collection interval	20 seconds	Same as used in the VISSIM test network
Model update interval	10 seconds	Same as used in the VISSIM test network
Control interval	60 seconds	Same as used in the VISSIM test network
Prediction horizon	300 seconds	A slightly shorter prediction horizon was used (300 seconds versus 360 seconds) to reduce the simulation run-time and also reduce the extent of variability in the utility predictions
Number of unique control intervals over the prediction horizon	1	Same as used in the VISSIM test network
Maximum change in the metering rate from one control interval to the next	Unconstrained	Same as used in the VISSIM test network
Number of particles used for tracking	15,000	The number of particles was increased substantially from the 5000 particles used in the VISSIM test network. This was found to be necessary due to the greater level of congestion in the Ottawa model, which the Freeway Traffic Model has more difficulty tracking (due to the relatively simplistic way that shockwaves are modelled). Even with this number of particles, fewer than 5 particles were being carried forward when the network was most congested, suggesting the need for improvements in the model formulation.
Number of particles used for prediction	1200	The number of particles was increased substantially from the 800 particles used in the VISSIM test network. With a lower number of particles, the algorithm was having difficulty determining the optimal solution due to stochastic variation in the utility results. There is some evidence that an even greater number of particles could be justified, however, the impact on run-time would be significant.

Parameter	Assumed Value	Comments
Utility weights – efficiency only scenario	Freeway Congestion: 0.55 Ramp Delay: 0.45	<p>Represents a minor variation from the weights used in the VISSIM test network. It was determined that a slightly lower weighting for freeway performance (0.55 vs. 0.60) was warranted to improve the overall efficiency results (potentially due to the higher level of congestion in the Ottawa model)</p> <p>In addition, a slightly different approach was used for calculating the freeway utility. Rather than using the weighted corridor travel speed as the basis for the utility calculation, the utility for each individual freeway cell was computed based on the cell speed and weighted together based on the cell volume.</p>
Utility weights – efficiency+equity scenario	Freeway Congestion: 0.50 Ramp Delay: 0.35 Equity: 0.15	Values are slightly different from those used in the VISSIM test network.
Ramp storage	Storage adjustments introduced at individual ramps	<p>As observed with the VISSIM test network, the algorithm sometimes fails to meter ramps hard enough initially in order to avoid queue spillback later in the prediction horizon. This issue arises due to the fact that the ramp metering rates are assumed to be constant over the prediction horizon (when in reality, new rates are introduced every 60 sec.). To avoid this situation, a storage adjustment factor of roughly 1.5 to 2 was applied at ramps with lower travel demand. At higher volume ramps (i.e. Carling &amp; Maitland), virtually no adjustment was required. Although the adjustment factors were found to produce the desired results, the ad hoc nature of the changes is less than desirable, suggesting the need for improvements</p>
Utility rounding	To the nearest 1000th	<p>When the algorithm was applied in the test network, an adjustment was added to the utility so that the algorithm would choose the scenario with the least restrictive metering in cases where the utility was identical. For the Ottawa model, this adjustment was dropped to reduce the model run-time</p>

Parameter	Assumed Value	Comments
Trigger to over-ride the calculated metering rates to address queue spillback	10 km/hr (lower volume ramps) 20 km/hr (higher volume ramps)	The Carling & Maitland on-ramps are both defined as high volume ramps, since the ramp demand exceeds 850 vph. At such ramps, a higher speed threshold is needed for triggering the over-ride mechanism since the remaining storage will tend to fill up more quickly, requiring action to be taken sooner
Speed threshold for defining queue spillback in the Freeway Traffic Model	10 km/hr	This value is slightly less than the value used for the VISSIM test network (15 km/hr). In the Ottawa model, the speed of drivers entering the ramp is influenced by upstream traffic signals, and also the length of the ramp (which in some cases, is relatively short). As a result, some vehicles were observed to enter the ramp at speeds of less than 15 km/hr, even though the ramp queue was not in danger of spilling back

<sup>1</sup> Only includes parameters that are different from those employed in the hypothetical VISSIM test network.

Note that in applying the new ramp control algorithm in the Ottawa model, the Freeway Traffic Model was modified to fix a minor coding error related to the calculation of mainline flow within traffic queues (was not properly accounting for traffic exiting the freeway). In addition, a constraint was added so that the maximum flow within the queue cannot exceed the congested capacity, even in shockwave conditions.

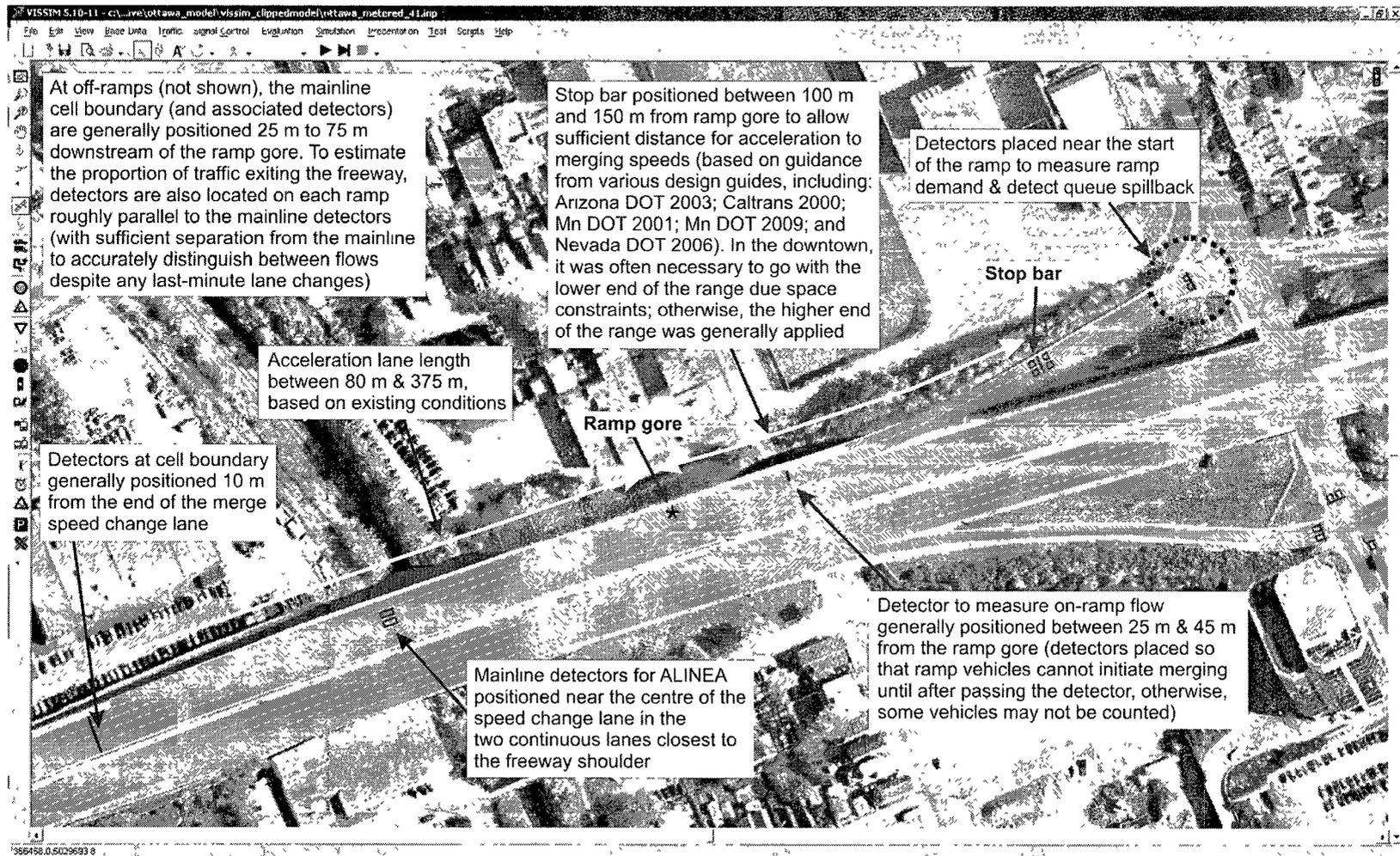
#### **A note regarding high volume ramps:**

At most ramps, vehicles are released from the stop bar in an alternating fashion, with one lane proceeding and then the other. When queue spillback is detected, the minimum cycle length (i.e. 4 sec) is applied, resulting in a maximum metering rate of 900 vph. However, at high volume ramps such as Carling and Maitland, the maximum metering rate is insufficient to reduce the queue. As a result, during spillback conditions, two vehicles are released at a time (one from each lane). Under such conditions, a cycle length of 6 sec was applied, giving a flow rate of 1200 vph which was generally found to be acceptable.

Within the simulation, vehicles departing the meter appeared to have no trouble merging down to a single lane prior to entering the freeway, even with two vehicles released at a time. However, the feasibility of the merge length for real-world networks was not confirmed.

With the introduction of tandem metering at high volume ramps during queue spillback conditions, it was necessary to increase the length of the check-in detector from 2 m to 10 m. Otherwise, the meter would occasionally turn off despite the existence of a ramp queue since vehicles were too slow moving up to be detected by the sensor.

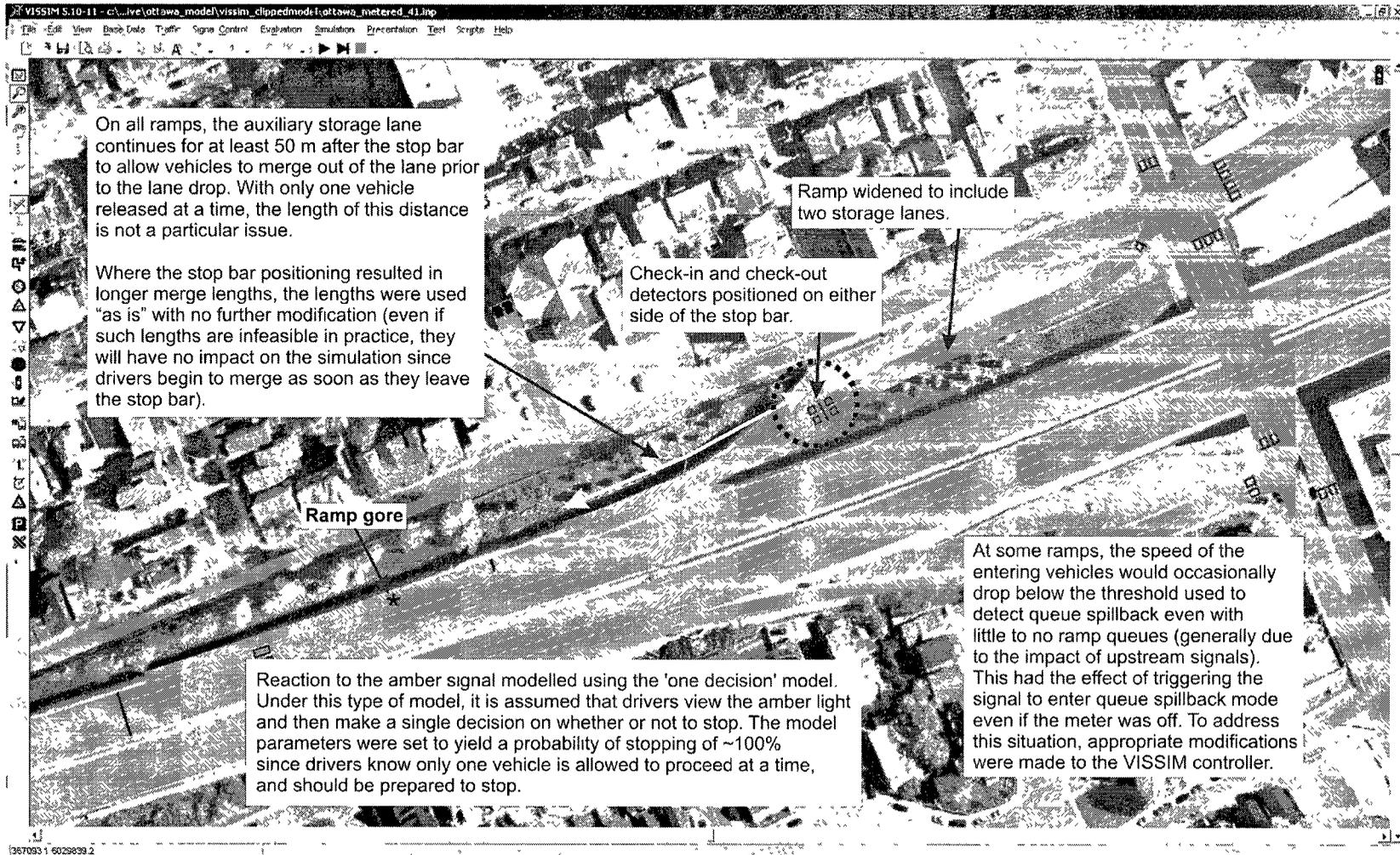
## Implementation of Ramp Metering on Highway 417 Westbound Assumptions & General Configuration



**Additional Notes:** All VISSIM detectors were assumed to have a smoothing factor of 1 (rather than the default value of 0.25) for adjusting the detector occupancy rate, since any required adjustments are undertaken directly by the ramp controller.

## Implementation of Ramp Metering on Highway 417 Westbound

### Assumptions & General Configuration



**Table P-5 Summary of Network Performance – ALINEA Algorithm**

<b>Performance Measure</b>	<b>Run 1</b>	<b>Run 2</b>	<b>Run 3</b>	<b>Run 4</b>	<b>Run 5</b>	<b>Run 6</b>	<b>Run 7</b>	<b>Run 8</b>	<b>Run 9</b>	<b>Run 10</b>	<b>Average</b>
# of vehicles that have left the network	76,028	75,931	75,907	76,063	76,099	75,935	76,194	75,899	76,014	76,079	76,015
# of vehicles in network at end of simulation	2904	3028	2881	2807	2766	2943	2689	2937	3001	2853	2881
Total distance traveled [km]	368,584	369,449	368,679	369,515	369,193	368,535	370,276	368,526	369,390	370,355	369,250
Total travel time [h]	5661	5651	5598	5419	5441	5565	5425	5561	5645	5455	5542
Average speed [km/h]	65.1	65.4	65.9	68.2	67.9	66.2	68.2	66.3	65.4	67.9	66.6
Total delay time [h]	1817	1800	1755	1568	1591	1725	1566	1719	1794	1594	1693
Average delay time per vehicle [s]	83	82	80	72	73	79	71	78	82	73	77
Number of stops	161,895	139,330	149,505	125,171	132,362	131,567	119,484	129,327	136,966	120,210	134,582
Average number of stops per vehicle	2.1	1.8	1.9	1.6	1.7	1.7	1.5	1.6	1.7	1.5	1.7
Total stopped delay [h]	514.6	511.9	501.2	478.2	476.5	502.1	485.7	493.5	517.3	499.1	498.0
Average stopped delay per vehicle [s]	23.5	23.3	22.9	21.8	21.8	22.9	22.2	22.5	23.6	22.8	22.7

\* Excludes the first 500 seconds of the simulation while the network is being loaded

**Table P-6 Summary of Network Performance – New Algorithm (Efficiency Only)**

<b>Performance Measure</b>	<b>Run 1</b>	<b>Run 2</b>	<b>Run 3</b>	<b>Run 4</b>	<b>Run 5</b>	<b>Run 6</b>	<b>Run 7</b>	<b>Run 8</b>	<b>Average</b>
# of vehicles that have left the network	75,824	75,958	75,654	75,949	75,998	76,018	75,607	75,861	75,859
# of vehicles in network at end of simulation	3109	3010	3222	2872	2871	2880	3214	2909	3011
Total distance traveled [km]	367,816	369,585	367,229	369,167	368,955	368,570	366,905	367,984	368,276
Total travel time [h]	5697	5676	5905	5541	5465	5537	6027	5627	5684
Average speed [km/h]	64.6	65.1	62.2	66.6	67.5	66.6	60.9	65.4	64.9
Total delay time [h]	1860	1823	2076	1694	1618	1697	2201	1790	1845
Average delay time per vehicle [s]	85	83	95	77	74	77	101	82	84
Number of stops	155,581	145,792	164,314	139,261	126,412	133,075	173,201	147,120	148,095
Average number of stops per vehicle	2.0	1.8	2.1	1.8	1.6	1.7	2.2	1.9	1.9
Total stopped delay [h]	496.4	506.5	498.3	490.1	473.1	482.1	513.7	494.3	494.3
Average stopped delay per vehicle [s]	22.6	23.1	22.7	22.4	21.6	22.0	23.5	22.6	22.6

\* Excludes the first 500 seconds of the simulation while the network is being loaded

**Table P-7 Summary of Network Performance – New Algorithm (Efficiency + Equity)**

<b>Performance Measure</b>	<b>Run 1</b>	<b>Run 2</b>	<b>Run 3</b>	<b>Run 4</b>	<b>Run 5</b>	<b>Run 6</b>	<b>Run 7</b>	<b>Run 8</b>	<b>Average</b>
# of vehicles that have left the network	75,815	75,804	75,731	75,546	75,899	75,713	75,523	75,678	75,714
# of vehicles in network at end of simulation	3118	3164	3084	3319	2970	3185	3356	3313	3189
Total distance traveled [km]	368,080	368,528	367,919	366,500	368,149	366,931	366,207	366,261	367,322
Total travel time [h]	5706	5723	5707	5926	5599	5810	6015	5924	5801
Average speed [km/h]	64.5	64.4	64.5	61.9	65.8	63.2	60.9	61.8	63.4
Total delay time [h]	1867	1880	1871	2104	1759	1986	2195	2103	1971
Average delay time per vehicle [s]	85	86	85	96	80	91	100	96	90
Number of stops	145,061	144,096	146,767	172,618	132,391	159,016	176,629	184,905	157,685
Average number of stops per vehicle	1.8	1.8	1.9	2.2	1.7	2.0	2.2	2.3	2.0
Total stopped delay [h]	468.4	478.2	472.9	496.4	456.5	490.9	500.6	514.2	484.8
Average stopped delay per vehicle [s]	21.4	21.8	21.6	22.7	20.8	22.4	22.8	23.4	22.1

\* Excludes the first 500 seconds of the simulation while the network is being loaded