ARIZONA DEPARTMENT OF TRANSPORTATION

REPORT NUMBER: FHWA-AZ99-462

RHODES - ITMS CORRIDOR CONTROL PROJECT

Final Report

Prepared by:
Douglas Gettman
Larry Head
Pitu Mirchandani
Systems and Industrial Engineering Department
University of Arizona
Tucson, Arizona 85721

May 1999

Prepared for:
Arizona Department of Transportation
206 South 17th Avenue
Phoenix, Arizona 85007
in cooperation with
U.S. Department of Transportation
Federal Highway Administration

The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Arizona Department of Transportation or the Federal Highways Administration. This report does not constitute a standard, specification, or regulation. Trade or manufacturer's names which may appear herein are cited only because they are considered essential to the objectives of the report. The U.S. Government and the State of Arizona do not endorse products or manufacturers.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-AZ99-462		

4. Title and Subtitle	5. Report Date
RHODES-ITMS CORRIDOR CONTROL PROJECT	**MAY 1999**
	6. Performing Organization Code

7. Author	8. Performing Organization Report No.
Douglas Gettman, Larry Head; and Pitu Mirchandani	

9. Performing Organization Name and Address	10. Work Unit No.
Systems and Industrial Engineering Department **The University of Arizona** **Tucson, Arizona 85721**	11. Contract or Grant No. SPR-PL-1(51)462

12. Sponsoring Agency Name and Address	13.Type of Report & Period Covered
ARIZONA DEPARTMENT OF TRANSPORTATION **206 S. 17TH AVENUE** **PHOENIX, ARIZONA 85007** ADOT Project Manager: Stephen R. Owen, P.E.	**Final Report 9/96 -5/99**
	14. Sponsoring Agency Code

15. Supplementary Notes

Prepared in cooperation with the U.S. Department of Transportation, Federal Highway Administration

16. Abstract

The *RHODES-ITMS Corridor Control* project addresses **real-time control of ramp meters** of a freeway segment, with consideration of the traffic volumes entering and leaving the freeway from/to arterials, and the regulation of these volumes via real-time setting of ramp metering rates.

Current approaches to traffic-responsive control of ramp meters include (a) time-of day control, (b) locally responsive strategies and (c) area-wide *linear programming* based approaches (currently implemented in parts of Europe). None of these approaches are both real-time responsive to traffic conditions and consider the multiple objectives of minimizing freeway travel times and decreasing congestion/queues at the interchanges.

A control system, MILOS (Multi-objective Integrated Large-scale Optimized ramp metering System), was developed that determines ramp metering rates based on observed traffic on the freeway and its interchange arterials. MILOS temporally and spatially decomposes the ramp-metering control problem into three hierarchical subproblems: (1) monitoring and detection of traffic anomalies (to schedule optimizations at the lower levels of the control hierarchy), (2) optimization to obtain area-wide coordinated metering rates, and (3) real-time regulation of metering rates to adjust for local conditions.

Simulation experiments were performed to evaluate the MILOS hierarchical system against (a) "no control" (i.e., when no ramp metering is in effect) , (b) a locally traffic-responsive metering policy, and (c) an area-wide LP optimization problem re-solved in 5-minute intervals. Three test scenarios were simulated (1) a short "burst" of heavy-volume flows to all ramps, (2) a three-hour commuting peak, and (3) a three-hour commuting peak with a 30-minute incident occurring somewhere in the middle of the corridor. The performance results indicate that MILOS is able to reduce freeway travel time, increase freeway average speed, and improve recovery performance of the system when flow conditions become congested due to an incident.

17. Key Words	18. Distribution Statement	23. Registrant's Seal
Real-time traffic-adaptive control; Adaptive Ramp-metering; Freeway Traffic Control; Hierarchical Control; Simulation; Optimization.	Document is available to the U.S. Public through the National Technical Information Service, Springfield, Virginia, 22161	

19. Security Classification	20. Security Classification	21. No. of Pages	22. Price
Unclassified	Unclassified	208	

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
	LENGTH			
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
ml	miles	1.61	kilometers	km
	AREA			
in^2	square inches	645.2	millimeters squared	mm^2
ft^2	square feet	0.093	meters squared	m^2
yd^2	square yards	0.836	meters squared	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	kilometers squared	km^2
	VOLUME			
ft oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	meters cubed	m^3
yd^3	cubic yards	0.765	meters cubed	m^3

NOTE: Volumes greater than 1000 L shall be shown in m^3.

	MASS			
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams	Mg

APPROXIMATE CONVERSIONS FROM SI UNITS

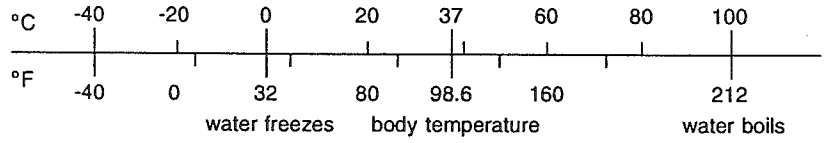
Symbol	When You Know	Multiply By	To Find	Symbol
	LENGTH			
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
	AREA			
mm^2	millimeters squared	0.0016	square inches	in^2
m^2	meters squared	10.764	square feet	ft^2
m^2	meters squared	1.19	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	kilometers squared	0.386	square miles	mi^2
	VOLUME			
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	meters cubed	35.315	cubic feet	ft^3
m^3	meters cubed	1.31	cubic yards	yd^3
	MASS			
g	grams	0.035	ounces	oz
kg	kilograms	2.205	pounds	lb
Mg	megagrams	1.102	short tons (2000 lb)	T

TEMPERATURE (exact)

Symbol	When You Know	Do The Following	To Find	Symbol
°F	Fahrenheit temperature	°F - 32 ÷ 1.8	Celcius temperature	°C

TEMPERATURE (exact)

Symbol	When You Know	Do The Following	To Find	Symbol
°C	Celcius temperature	°C x 1.8 + 32	Fahrenheit temperature	°F



METER: a little longer than a yard (about 1.1 yards)
LITER: a little larger than a quart (about 1.06 quarts)
GRAM: a little more than the weight of a paper clip
MILLIMETER: diameter of a paper clip wire
KILOMETER: somewhat further than 1/2 mile (about 0.6 mile)

*SI is the symbol for the International System of Measurement

RHODES-ITMS Corridor Control Project

PREFACE

This report documents the work performed on the *Corridor Control* Subproject of the *RHODES-ITMS* Project. This research effort was funded by the Arizona Department of Transportation (ADOT) and the Maricopa Association of Governments (MAG). Essentially, the scope of the Project was to develop a method to optimally control, in real time, the ramp meters on a segment of a freeway. The control architecture used was based on extensions of the hierarchical control concepts developed for the surface street network in the previous *RHODES* Project funded by ADOT and the Pima Association of Governments.

The *Corridor Control* subproject was directed by the principal investigators, **Pitu B. Mirchandani** and **Larry Head,** both of the Systems and Industrial Engineering Department at the University of Arizona. This report is largely based on the dissertation written by Dr. **Douglas Gettman** whose Ph.D. research was supported by the project.

In addition, Drs. Gettman, Head, and Mirchandani wish to acknowledge their appreciation to the Project's **Technical Advisory Committee** (TAC) whose continual active participation, technical input and support resulted in *MILOS* (the real-time ramp-metering system described in this report) being very relevant to freeway ramp-metering control. The following individuals served on the TAC at various times:

Jim Decker	Traffic Operations, City of Tempe
Tim Wolfe	ADOT Technology Group
Dan Powell	ADOT District 1
Tom Parlante	ADOT Traffic Engineering
Glenn Jonas	ADOT Freeway Management
Jim Shea	ADOT Traffic Engineering
Sarath Joshua	Maricopa Association of Governments (previously at ATRC, ADOT)
Paul Ward	Maricopa Association of Governments
Roy Turner	Maricopa Association of Governments
Pierre Pretorius	Maricopa County Transportation and Development Agency
Don Wiltshire	Maricopa County Transportation and Development Agency
Alan Hansen	Federal Highway Administration
Tom Fowler	Federal Highway Administration
Jessie Yung	Federal Highway Administration
Steve Owen	RHODE-ITMS Project Manager, Arizona Transportation Research Center (ATRC), ADOT

RHODES-ITMS Corridor Control Project

EXECUTIVE SUMMARY

The RHODES-Integrated Traffic Management System (ITMS) Program addresses the design and development of a real-time traffic adaptive control system for an integrated system of freeways and arterials. The overall program was initiated in December 1993, jointly funded by the Arizona Department of Transportation (ADOT) through the State Planning and Research Program budget and the Maricopa Association of Governments (MAG). The *RHODES-ITMS* program is overseen by ADOT's Arizona Transportation Research Center.

Subsequently, in September 1996, the *RHODES-ITMS Corridor Control* research project was initiated which specifically addresses **real-time control of ramp meters** of a freeway segment, with consideration of the traffic volumes entering and leaving the freeway from/to arterials, and the regulation of these volumes via real-time setting of ramp metering rates. This is the final report for the *RHODES-ITMS Corridor Control Project.*

RESEARCH CONCEPTS

Current approaches to controlling ramp meters to respond to varying traffic conditions, those reviewed in this report, include (a) time-of day control, (b) locally responsive strategies (one such strategy is currently under consideration by ADOT), and (c) area-wide *linear programming* (LP) based approaches (currently implemented in parts of Europe). None of these approaches are both fully responsive in real-time to prevailing and predicted traffic conditions and consider the multiple objectives of minimizing freeway travel times and decreasing congestion/queues at the interchange ramps and the corresponding arterial intersections.

In this research, a control system was developed, referred to as MILOS (Multi-objective Integrated Large-scale Optimized ramp metering System), that determines ramp metering rates based on observed and predicted traffic on the freeway and its interchange arterials. MILOS has an hierarchical architecture to address the complexities of the real-time freeway management problem, namely, (a) the dynamic and stochastic nature of state changes, and (b) the existence of multiple objectives. MILOS temporally and spatially decomposes the ramp-metering control problem into three hierarchical subproblems: (1) monitoring and detection of traffic anomalies (to schedule optimizations at the lower levels of the control hierarchy), (2) optimization to obtain area-wide coordinated metering rates, and (3) real-time regulation of metering rates to adjust for local conditions.

The area-wide coordination problem at the second level of the hierarchical control system is modeled as a "*quadratic programming*" (QP) optimization problem that considers the impact of queue growth on the adjacent interchanges. A multi-criterion objective function is used to trade-off between freeway travel times and congestion/queues at the interchanges. The resulting nominal solution of the second-level area-wide optimization

problem is then provided to the third-level control function which locally adjusts these nominal ramp-meter rates.

The third-level problem, referred to as predictive-cooperative real-time (PC-RT) rate regulation problem, modifies, if necessary, the ramp metering rates based on local traffic at each interchange. The PC-RT algorithm is based on a linear programming formulation that uses a linearized approximation of a macroscopic freeway flow model (in terms of dynamic difference equations). The PC-RT algorithm pro-actively utilizes opportunities to disperse queues or hold back additional vehicles when freeway and ramp traffic conditions are appropriate. The cost coefficients of the LP optimization objectives are based on the multi-objectives trade-offs considered in the second-level area-wide coordination problem.

The optimization runs of the area-wide coordination problem and the PC-RT rate regulation problem at each ramp are scheduled for execution by the highest-level ramp-demand/freeway-flow monitoring system that is based on concepts from "*statistical process control*" in production systems. Basically, this system functions as follows: When the monitored conditions are within the expected variances in ramp demands and freeway flows, no optimization run is scheduled to obtain new ramp metering rates; when the conditions are outside the expected variances then either the PC-RT algorithm (LP) is run if the deviations are not too large, or the area wide QP is run when the deviations are large, to obtain new ramp metering rates.

RESULTS

Simulation experiments were performed to evaluate the MILOS hierarchical system against (a) "no control" (i.e., when no ramp metering is in effect) , (b) a locally traffic-responsive metering policy currently under consideration by ADOT, and (c) an area-wide LP optimization problem re-solved in 5-minute intervals. The simulation model was of a small freeway corridor in metropolitan Phoenix, Arizona (seven miles of State Route 202 with 7 off-ramps, 4 controllable on-ramps, and one freeway-freeway on-ramp (hypothesized as controllable). Three test scenarios were simulated (1) a short "burst" of heavy-volume flows to all ramps, (2) a three-hour commuting peak, and (3) a three-hour commuting peak with a 30-minute incident occurring somewhere in the middle of the corridor.

The performance results indicate that MILOS is able to reduce freeway travel time, increase freeway average speed, and improve recovery performance of the system when flow conditions become congested due to an incident. Specifically, when comparing with the "no control" case, freeway travel times were lowered by 8% - 36%, speeds were increased by 3% - 18%, and recovery times were reduced by 6% - 25%. It also performed better than the area-wide LP optimization that has been reported to perform well in Europe. Locally responsive strategy performed well in light to moderate traffic volumes, and, in fact, had lower freeway travel times and faster speeds than MILOS; however, for heavier volumes and incidents it had larger ramp-queues and longer recovery times than MILOS.

AREAS OF FUTURE WORK

This research project identified several interesting future research, development and deployment efforts. Development of (1) an algorithm to decompose a region into subnetworks for MILOS control, (2) a model on route diversion and (3) methods to estimate ramp demands and interchange turning probabilities are promising research areas. Integration of MILOS with traffic-adaptive surface-street signal control and incident/anomaly detection systems are developmental efforts that could make traffic management even more real-time responsive. Finally, field testing and the deployment of MILOS (and its future enhancements) should be an on-going effort towards the ITS goal of implementing advanced traffic management systems that are safer, more efficient and beneficial to the traveling public.

ACKNOWLEDGMENTS

The *RHODES-ITMS Corridor Control Project* was completed in January 1999. Project oversight was provided by a Technical Advisory Committee (TAC) comprising of representatives from key agencies. The project was administered by the Arizona Transportation Research Center of ADOT. The following individuals served on the TAC at various times:

Steve Owen	RHODES-ITMS Project Manager, Arizona Transportation Research Center (ATRC), ADOT
Tim Wolfe	ADOT Technology Group
Dan Powell	ADOT District 1
Tom Parlante	ADOT Traffic Engineering
Glenn Jonas	ADOT Freeway Management
Jim Shea	ADOT Traffic Engineering
Alan Hansen	Federal Highway Administration
Tom Fowler	Federal Highway Administration
Jessie Yung	Federal Highway Administration
Sarath Joshua	MAG (previously at ATRC, ADOT)
Paul Ward	MAG
Roy Turner	MAG
Pierre Pretorius	Maricopa County Transportation and Development Agency
Don Wiltshire	Maricopa County Transportation and Development Agency
Jim Decker	Traffic Operations, City of Tempe

Table of Contents

List of Figures

List of Tables

Chapter 1: Problem overview

Introduction

Traffic delay due to congestion on freeways and surface streets was approximated at 1.2 billion vehicle-hours in the United States in 1984 and projected to reach 6.9 billion vehicle-hours by 2005 [Lindley, 1987]. User costs associated with traffic delay were estimated at $100 billion in 1990 for the U.S. [Euler, 1990]. Total trips and commuter miles are expected to grow significantly in most metropolitan areas as the trend towards suburban sprawl continues. Construction of additional freeway lanes and wider surface streets is certainly needed to respond to such societal needs. However, in many situations, it is not possible to address capacity needs in dense urban areas with new construction. In these situations, capacity increases are possible only by adding modes of travel (rail, subway, etc.) or reducing traveler delays with more efficient management of the available system capacity.

Introduction of traffic management devices and systems (i.e. traffic signal systems, ramp metering systems, lane channelization, HOV, etc.) has been shown to reduce delays and increase capacity [FHWA, 1985]. Early estimates of the impacts of ITS technologies range from 10% reduction in emissions (relative to the projected increase in total vehicle-miles traveled) to 20% savings in vehicle-delay and 30% reduction in stops [Mobility 2000, 1989]. These results are realized *without* significant spending on road widening and adding miles of freeway. Benefit/cost ratios of 16:1 and 22:1 have been reported for investment in ITS technologies in Los Angeles and Texas, respectively [Mobility 2000, 1989].

One of the most significant contributors to total vehicle-hours of delay is the daily commute of travelers on the freeway from their homes to their place of business and vice-versa. On average, over 38% of total vehicle-hours of delay are recurrent, i.e. occurring during the commuting hours, often referred to as the *peak* periods [Lindley, 1987]. Accidents and anomalous events, sometimes referred to as *nonrecurrent* congestion-related delay, account for the majority of the remaining delay factors. Thus, the largest

impacts on the reduction of total system delay for *freeway operations*, are realized by efficiently managing the critical peak times and managing incident conditions effectively.

The fundamental freeway management problem

The central problem in freeway management can be described best by presenting the *fundamental* diagram of freeway flow. Figure 1-1 illustrates speed versus volume on a typical freeway segment in Phoenix, AZ [Technical Advisory Committee, 1997]. In this figure, the upper concentration of points represents the uncongested flow "regime", where traffic flows smoothly, i.e. without significant travel delays. The lower, less populated collection of points represents 15-minute intervals when this freeway section was congested, incurring traffic delays to travelers in and upstream of this section.

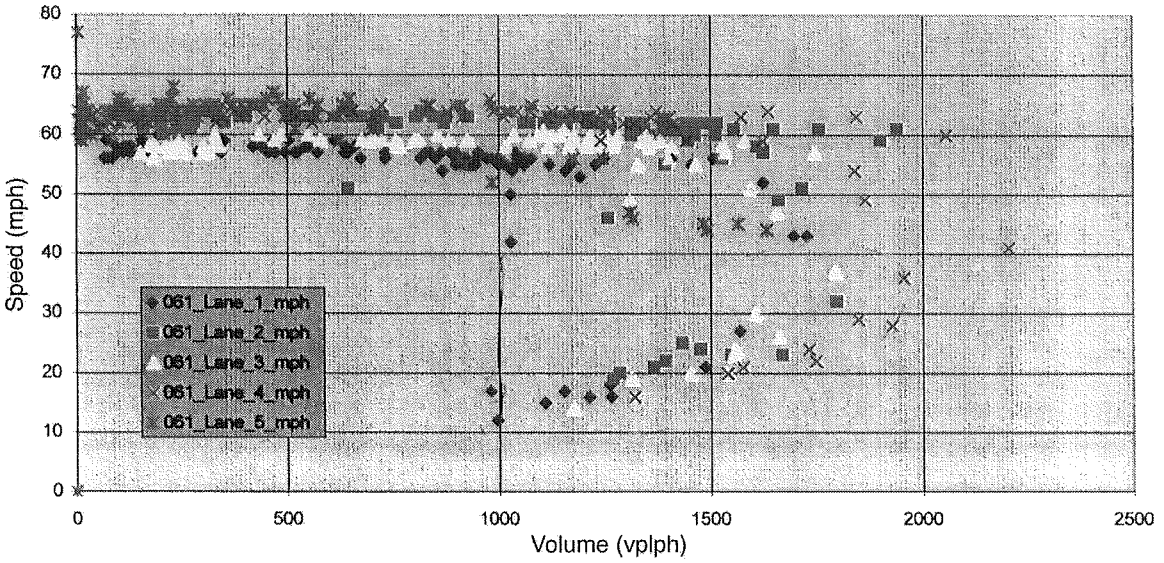


Figure 1- 1. Empirical speed-volume measurements

Figures 1-2 and 1-3 depict the time-series of the same points in Figure 1-1. These figures demonstrate that as the volume rises it becomes increasingly more precarious that the speed will drop sharply and the system will "transition" to the *congested* flow regime. As the system approaches closer and closer to the maximum flow rate, the transition can be initiated given increasingly smaller shocks [Newell, 1993]. That is to say that the higher the volume becomes, the more sensitive the system is to small anomalies in flow such as anomalies induced by merging *platoons* of vehicles.

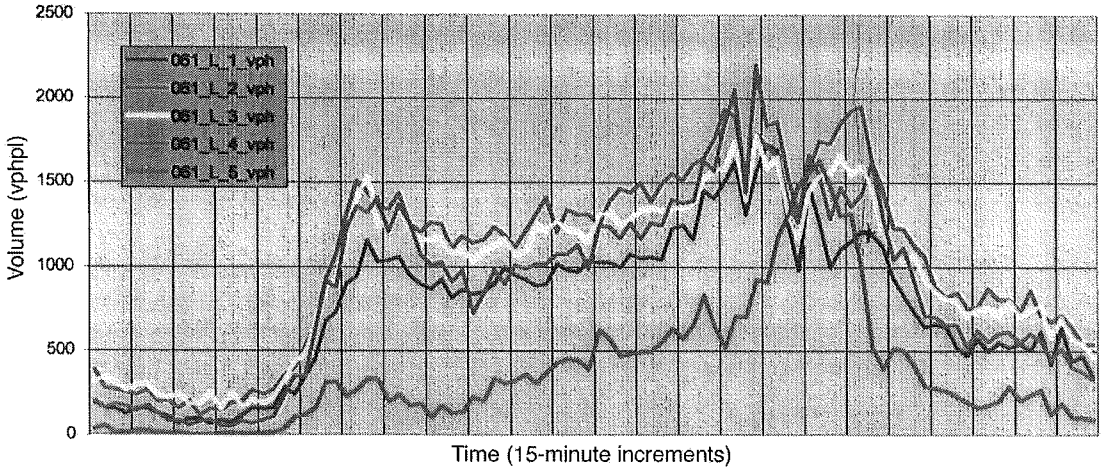


Figure 1- 2. 15-minute flow time-series indicating congestion

This increased susceptibility to shocks is due to the fact that the time between adjacent vehicles in a single lane (i.e. *headways*) *must* become smaller when the volume increases. Thus, when any interruption of the flow process occurs, drivers have less time to react to changes in the speed of the vehicle they are following and tend to over-react; braking sharply. This sharp braking can cause immediate transition to congested flow, since the volume is not reduced at the same rate as the speed.

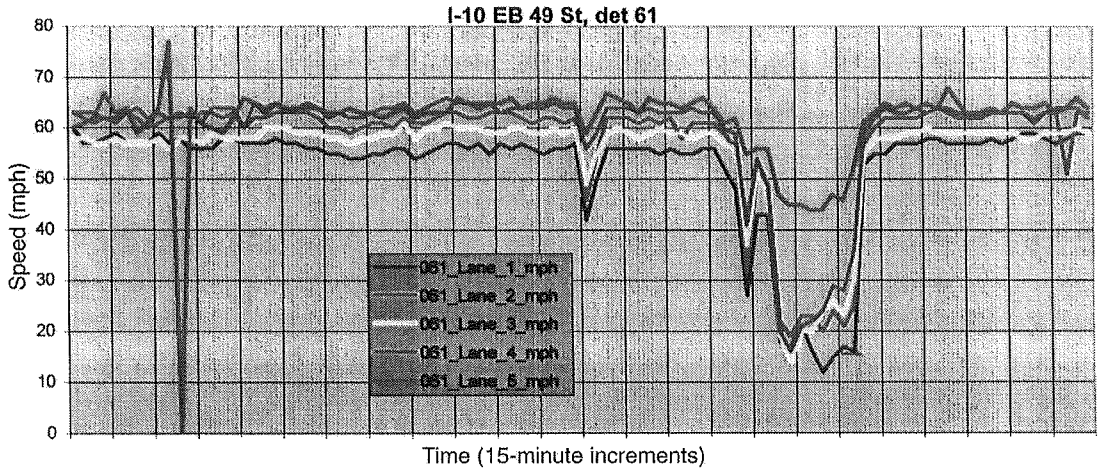


Figure 1- 3. 15-minute speed time-series indicating congestion

Given this fact, there is an obvious difficulty in operating the freeway system at the highest volumes (and best use of available capacity) near the upper end of the characteristic curve of Figure 1-1. This point is the *most susceptible* to transitioning into

the undesirable congested flow regime. The difficulty is amplified because, as indicated by Figures 1-2 and 1-3, capacity reductions usually occur when demand is greatest during the peak morning and afternoon commuting times. Thus, without some form of demand/capacity management, the sheer volume of demand for the freeway system drives the network into congestion.

There are several basic *technological* forms of freeway management available to address the freeway flow trade-off:

(a) Advising travelers to avoid certain sections and/or change their departure times,

(b) Advising travelers (unilaterally) to maintain a certain speed, or

(c) Restricting access to the freeway system at certain locations.

Solution (a) describes passive advanced traveler information systems (ATIS) methods which is outside the scope of this dissertation. Solution (b) has been investigated [Karaaslan, et al., 1990; Smulders, 1993], but is likely to have compliance problems in the U.S. Automated highway systems (AHS) eliminate this compliance problem, but are relatively far from mass implementation due to regulatory concerns and cost issues [Bender, 1991]. Solution (c) is generally referred to as *ramp metering*, and is the primary topic of this research.

Issues in application of ramp metering as a method of freeway management

Ramp metering systems have existed since the early 1960's and have been used effectively in many municipalities [Carlson, 1979; Marsden, 1981; Jacobson, 1989; Haj-Salem et al., 1990; Hallenbeck and Nisbet, 1993; Wright, 1993] and in others with less conclusive benefits [Lipp et al., 1992], but there is general consensus that ramp metering systems can provide substantial benefits in throughput, travel-time, and congestion reduction when applied *appropriately*. In fact, the effectiveness of on-ramp metering has been substantial enough that a recent study proposed *main-line* metering as a tool for congestion management [Haboian, 1997].

There does *not* exist a consensus, however, of what constitutes the most *effective* metering rate(s), arising from the fact that the freeway management problem is difficult to solve to optimality. Consider the following complicating factors:

(a) The state variables (i.e. volume, density, speed, ramp queues, etc.) change *dynamically over time;*

(b) Although the behavior of *individual* travellers is somewhat deterministic in that the drivers know where they are going, or at least have a trip purpose, the behavior of traffic as a *stream* is stochastic and difficult to predict over long time horizons;

(c) Flow *anomalies* (e.g. accidents, friction effects) occur at random, unpredictable intervals;

(d) The ramp metering problem is a *multi-dimensional* one, with a large number of state and control variables;

(e) The state variables are only *partially observable* at a limited number of fixed locations where detectors are installed, and

(f) There are *multiple stakeholders* and, consequently, *multiple objectives* need to be addressed in any freeway management policy.

The concerns of multiple stakeholders arise when you consider the fact that the freeway system exists *embedded inside* and *interacting with* a larger network of surface-streets and other modes of transportation. Previous research has not considered the complexity associated with considering multiple stakeholders by:

(1) assuming that the freeway system exists in virtual *isolation* from the larger surface-street network (e.g. usually assuming that ramp queuing capacity is infinite), and

(2) choosing a single *system-optimal* optimization criterion that considers only the effects of metering decisions on freeway conditions.

Freeway management policies developed or proposed to date have included, however, considerations of dynamic state changes, stochasticity, multi-dimensionality, unpredictability, and partial-observability in the freeway management problem.

The question remains, however, whether a *system-optimal* policy for the freeway control problem *alone* is "system-optimal" for the *entire* transportation network. Here we indicate the transportation network as the entire system of freeways and surface streets in a metropolitan area or municipality (or collection thereof). In fact, it is entirely plausible that the freeway management policy "optimal" to the freeway conditions could be counterproductive to the entire transportation network because the interactions between the two surface street system and the freeway system are neglected.

This is especially true for ramp metering methods that (inevitably) create *queues* at the freeway access points. These queues, if not suitably managed, can interfere with operation of the surface-street system by extending into the adjacent interchange (commonly known as *spillback*). Thus, the objectives at the ramp interface of the surface-street manager and the freeway system manager conflict. The surface-street manager would like to keep the ramp queue as short as possible and the freeway manager would like to keep the queue as long as possible typically during congested conditions.

The multi-objective approach to freeway system management

The central issue addressed by this research is the consideration of the important interaction between the surface-street system and the freeway system and their traffic objectives in the development of a freeway access control (ramp metering) system. This problem is addressed by using a *multi-objective* solution methodology. The trade-off solutions produced by this solution method are defined by *combining* the two conflicting objectives into a single, multi-criterion objective as opposed to other methods that enumerate Pareto solutions [Haimes, et al., 1990]. In addition, because of the relative size of the carrying capacity of the freeway with respect to the adjacent surface-street system, the trade-off solution point is selected to *maintain* freeway performance that is at least as good as management policies that do *not* consider the interactions of the two sub-systems. It will be shown in Chapter 9, via simulation, that acceptable freeway performance similar to area-wide control methods that do *not* consider the effects on the interchanges can be obtained by implementing a compromise solution while, at the same time, providing queue management.

Research methodology

To mitigate the complicating factors of the multi-objective ramp queue management issue, this research uses a structured approach based on previous work in freeway ramp metering control systems, but utilizing new technologies where appropriate. Table 1-1 indicates the characteristic of the research methodology that addresses each of the complicating factors.

Complicating factor	**Mitigating control system characteristic**
Dynamic state changes	Rolling-horizon optimization Temporal-spatial decomposition
Stochasticity	SPC-based anomaly detection Temporal-spatial decomposition
Multi-dimensionality	Temporal-spatial decomposition
Unpredictability	Predictive scenario optimization SPC-based anomaly detection
Partial-observability	Predictive scenario optimization Rolling-horizon optimization
Multiple objectives	Multi-objective criterion functions Cost coefficient trade-off weights

Table 1- 1. Characteristics of the proposed freeway control system

In brief, Table 1-1 identifies the characteristics of a hierarchical control system that decomposes the large-scale freeway ramp metering into a series of optimization problems of varying temporal and spatial resolution. The optimization problems are re-solved as the parameters and conditions of the system change to continually adjust the control strategy to the real-time behavior of the system. In addition, to mitigate the unpredictability of the future system state, a predictive scenario-based optimization scheme is implemented in real-time to prepare the local subsystem for the next short-term stochastic disturbance.

Summary of the forthcoming chapters

The remainder of this document is structured as follows; Chapter 2 presents a brief overview of previous work on the ramp metering problem. Chapter 3 outlines the hierarchical structure of the research methodology and the temporal/spatial decomposition of the control problem. Although hierarchical treatment of freeway management is not new, the specific hierarchy proposed in this project is novel, in particular the identification of subnetworks from a large-scale freeway system, and the basis for interaction between the area-wide layer and the locally traffic-responsive layer are new. Chapter 4 presents a popular and useful model of freeway traffic flow modified slightly to more accurately represent the ramp-freeway interface under the presence of congestion. Chapter 5 presents the area-wide coordination component of the hierarchical control system that considers the impact of queue growth on the adjacent interchanges in the optimization model. This optimization model is based on models available in the literature but incorporates several additions: (1) a new multi-criterion objective function and trade-off structure, (2) an alternative treatment of queue growth constraints, and (3) modeling of demands from surface-street interchange flows.

Chapter 6 presents the locally traffic-reactive, predictive-cooperative real-time rate regulation algorithm that provides additional capacity at the freeway/surface-street interface. The basis for this optimization model is not new (i.e. linearization of the nonlinear macroscopic flow model of Chapter 4), but the formulation of the scenario-based linear-programming problem is new. The link to the solution of the area-wide coordination problem of Chapter 5 using the dual information is entirely novel.

Chapter 7 presents the statistical process control concepts used to monitor system operation and, in real-time, identify perturbations to the system states. This structure of demand estimation and fluctuation identification in the context of freeway management systems is an entirely new treatment of this modeling/estimation/optimization procedure. Chapter 8 summarizes the hierarchical control system presented as components in Chapters 5, 6, and 7 and presents the algorithmic operation of the system. Chapter 9 presents a simulation experiment that evaluates the hierarchical system against several

other ramp metering policies on a relatively small, but realistic, freeway management problem in the metropolitan Phoenix, AZ area. Presentation of performance variance information comparing metering methods has not been done before in freeway management literature. Finally, Chapter 10 summarizes the results of the research.

Chapter 2: Ramp metering literature review

Costs and benefits of ramp metering

Ramp metering is the most widely used form of freeway control [Yagar, 1989]. Ramp metering limits the rate at which vehicles enter the freeway system, thus potentially reducing the possibility of bottlenecks, shock wave propagation, and congestion. A wide range of benefits are available from the use of ramp metering [Arnold, 1987; Yagar, 1989; McShane and Roess, 1990]:

(1) minimizing the total travel time of freeway users
(2) efficient use of freeway capacity
(3) discouraging routes with high societal costs
(4) reducing the variance of corridor trip times
(5) decreasing local freeway congestion and shock waves resulting from merging platoons
(6) decreasing the accident rate in freeway weaving sections.

Another study indicates that ramp meters can efficiently reduce system travel time, although the savings are *network dependent* [Hellinga and Van Aerde, 1997]. Of course there are disadvantages and adverse effects of ramp meters:

(1) encouraging longer trip distances on diversion routes
(2) favoring through traffic over local traffic and short trips
(3) modifying the evolved "status quo" of unobstructed freeway entry
(4) increasing the overall operating cost of the control system
(5) adversely affecting the surface street controller operation due to queue spillback and diversion to oversaturated locations

An operational study in the Denver area showed no statistically-significant improvement when a simple demand-capacity metering system was installed and evaluated [Lipp et al., 1991]. Ramp metering advantages may also be *strongly* dependent on the existence of good alternative routes, especially in the absence of effective queue management

strategies [Hellinga and Van Aerde, 1997]. Nevertheless, most large metropolitan areas have some type of ramp metering installed or currently under installation, indicating that practitioners have been convinced that the benefits (fiscal, social, temporal) of ramp metering outweigh the costs of implementation, maintenance, and the adverse effects mentioned above.

Types of ramp metering algorithms

The vast array of ramp metering algorithms developed to date can be classified into one of three general categories;

(1) *fixed-time* or time-of-day,

(2) *traffic-responsive*, and

(3) *hybrids* combining attributes of traffic-responsive and time-of-day algorithms.

Time-of-day metering algorithms

Time-of-day metering algorithms derive settings that apply during 10-30 minute intervals based on historical origin-destination flow rates and demand volumes for an entire commuting corridor or facility [Wattleworth and Berry, 1967; Messer, 1969; Yuan and Kreer, 1971; Wang, 1972; Wang and May, 1973; Chen et al., 1974; USDOT, 1976; Kahng et al., 1984].

The main drawback of fixed-time, time-of-day metering systems is the inability to handle non-recurrent incidents, accidents, special events, and fluctuations in traffic flow that may occur [Newman et al., 1970], since the actual demand may not be close to the demand used to derive the time-of-day metering rate. Recent studies have indicated that although time-of-day and day-to-day patterns exist, the variability of the actual flows from the historical average flows is significant enough to make some time-of-day settings ineffective [Rahka and Van Aerde, 1997].

Local traffic-responsive ramp metering algorithms

Traffic-responsive ramp metering algorithms measure variables such as speed, volume, and occupancy on the freeway and apply metering rates that keep the local freeway volume under capacity or at some desired set-point [Athans, 1969; Buhr et al., 1969;

Hardin, 1972; Estep, 1972; Pretty, 1972; HCM, 1985; Papageorgiou, 1989, 1991; Middelham and Smulders, 1991; Nihan, 1991; Nihan and Berg, 1992; Davis, 1993; Chang and Wu, 1994]. Other locally traffic-reactive ramp metering systems have been developed that merge vehicles into gaps in traffic [Drew et al., 1966; Wattleworth and Courage, 1968; Brewer et al., 1969] but such systems have not been widely implemented. Other types of traffic-reactive metering systems follow a pre-determined set of relationships between metering rates and traffic variable measurements. Examples of such systems are fuzzy and traditional rule-based expert systems and neural networks [Blumentritt et al., 1981; Rajan et al., 1986; Sasaki and Akiyama, 1987; Gray et al., 1990; Stephanedes et al., 1992; Zhang et al., 1994; Zhang and Ritchie, 1995; Papageorgiou et al., 1995].

The main drawbacks of using traffic-responsive ramp metering in a large-scale freeway are:

(1) the absence of *coordination* between adjacent ramp meters, and

(2) the absence of consideration of the *area-wide effects* of local changes to the metering rate.

Hybrid ramp metering control algorithms

Many hybrids and extensions of the basic traffic-responsive and time-of-day control methods have been developed. These extensions address the generally recognized issue that although day-to-day and time-of-day patterns exist, and can be exploited, their realization on a *specific* day and time may be significantly different from the assumed historical average pattern. Thus, hybrid ramp metering algorithms allow the system to follow the underlying trends, but still react to temporal and/or spatial flow irregularities.

The most straightforward extension of time-of-day methods is the use of a *rolling time horizon* and/or *periodic re-optimization* of the area-wide algorithm with new information [Messer, 1969; Drew et al., 1969; May, 1979; Papageorgiou, 1980, 1983; Kahng et al, 1984; Chang and Wu, 1994; Asakura, 1995]. However, the issue has been raised of how

long such re-optimization intervals should be. Most studies used *fixed* update intervals of 5-15 minutes.

The complexity of the large, multi-variable ramp metering problem has also been addressed by *decomposition* of the problem into smaller-scale descriptions of subsystems which are each optimized independently [Isaksen and Payne, 1973; Looze et al., 1978; Payne et al., 1979; Goldstein and Kumar, 1982; Papageorgiou, 1983; Kahng et al., 1984; Payne et al., 1985]. Many hybrids offer a combination of the rolling-horizon extension and spatial decomposition by establishing a *hierarchical* approach to the large-scale ramp metering problem, similar to the organizational structure shown in Figure 2-3 [Papageorgiou, 1983].

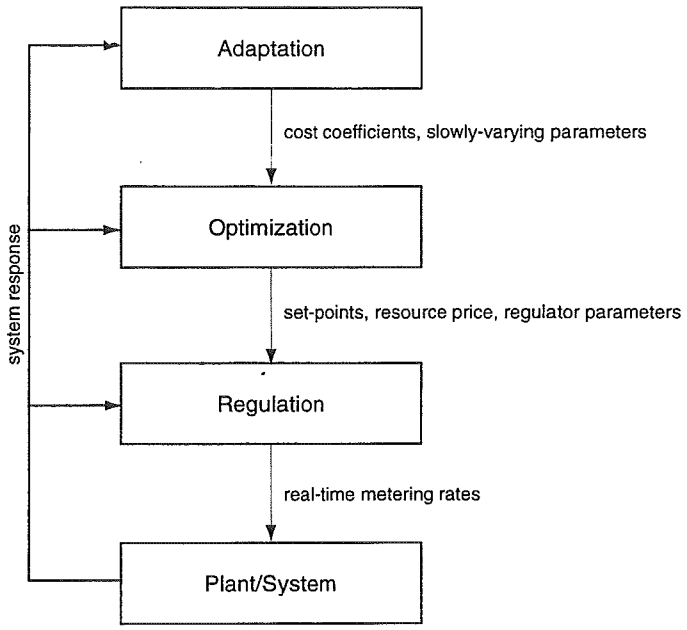


Figure 2- 1. Typical hierarchical control system structure

These systems combine locally-optimized traffic-responsive control with guidance from upper levels of the hierarchy regarding area-wide conditions, special events, and incidents [Drew et al., 1969; May, 1979; Papageorgiou, 1984; Payne et al., 1985]. Hybrid metering algorithms based on hierarchical control structures can also support the explicit consideration of optimization *modes* such as "normal flow", "congestion", and

"special-event" that are scheduled by the highest level(s) of the hierarchy [Papageorgiou, 1984; Pooran et al., 1994].

Integrated freeway/surface-street metering algorithms

Although there is a large body of work on freeway control algorithms, only exploratory work has been done to produce ramp metering solutions that integrate information from the surface-street system [Fan and Asmussen, 1990; Stephanedes and Chang, 1991, 1993; Pooran et al., 1992, 1994; Han and Reiss, 1994]. Some work has recently been proposed to develop freeway management solutions that derive *both* signal settings and metering rates in a commuting corridor of surface streets and freeway [Cremer et al., 1990; Chang et al., 1992; Papageorgiou, 1995; Zhang and Hobeika, 1997]. The failure to integrate the two systems has been due to the technological barriers that have restricted application of data-intensive ramp metering methods and the difficulty of modeling the two sub-systems together for optimization purposes [Van Aerde et al., 1987]. As such, no results of field implementation studies could be found in the literature.

However, as the technological barriers are being removed and real-time traffic information is becoming readily available, a new focus on improving the *system-wide* performance of the freeway and surface street network has emerged [Van Aerde and Yagar, 1988]. Critical data such as origin-destination (and/or route-proportional) matrices, time-varying demands, turning probabilities, and the like can be more reliably estimated on-line as the Intelligent transportation systems (ITS) "infrastructure" of communication networks and detection technology continues to be deployed.

Summary

For the past 35 years, much research has been done in the area of ramp metering control systems. Even simple metering systems installed in the field have been shown to be effective at improving freeway performance and having benefits that outweigh the installation and recurrent operating costs. However, metering systems sometimes have detrimental effects to the adjacent surface-streets when the ramp queue spills back into the interchange. Methods to address the spillback problem at the interface between the

freeway system and surface-street system have only recently been established in the research community and sparsely implemented in the field. Another drawback of ramp metering algorithms based on local traffic is the lack of consideration for the system-wide effects of the metering decisions and dis-proportionate queue growth rates [Benmohamed and Meerkov, 1994]. The remainder of this document describes a hierarchical freeway management system that builds on the successes of previous research in hybrid ramp metering algorithms and adds consideration of the important problem of queue management.

Chapter 3: Hierarchical ramp metering control system structure

Introduction

The ramp metering control system developed in this research is specifically designed to address the complicating factors of the freeway management problem. Recall from Chapter 1 that the freeway management problem is a difficult control and optimization problem because of these factors. Previous work in freeway ramp metering control systems has primarily focused on the complications caused by:

(1) dynamic state changes,
(2) stochasticity,
(3) multi-dimensionality, and
(4) partial observability

without consideration of *multiple objectives* or the *unpredictability* of the future traffic state.

The freeway control system developed in this research addresses all six of the complicating factors by establishing a hierarchical system of *layers* that

(1) addresses *embedded* spatial and temporal descriptions of the ramp metering control problem,
(2) considers concerns of *both* the freeway and surface-street systems in the optimization problem(s),
(3) plans *pro-active* metering rates in real-time to respond to possible future traffic states, and
(4) re-schedules optimizations based on the *stochastic fluctuations* of the demand processes.

Before detailing the characteristics of the hierarchical control model developed in this research, we review some concepts and previous research in hierarchical optimization.

Multi-*level* methods in hierarchical control

The hierarchical approach to system control has substantial fundamental research support, especially in the area of large-scale differential equation systems [Mahmoud,

1977; Sandell et al., 1978; Wilson, 1979; Papageorgiou and Schmidt, 1980; Bernassou and Titli, 1982; Papageorgiou, 1983]. Hierarchical control is particularly useful when the system being controlled has an appreciably large set of state variables, and/or and appreciably large set of control inputs. Large-scale control problems of this type are primarily difficult because of appreciable computation time required to solve for the "optimal" controls.

Such large-scale differential equation control problems are typically addressed by decomposing the problem using a multi-*level* approach. The multi-*level* approach creates a two-level optimization problem from a global optimization problem. Various methods have been proposed to solve the two-level optimal control problem including *interaction-prediction* and *interaction-balance* procedures [Sandell et al, 1978; Wilson, 1979; Papageorgiou and Mayr, 1982].

Multi-*layer* hierarchical control systems

A hierarchical control system can also describe a controller that solves the ramp metering problem at several embedded *layers* of aggregation. Thus, the multi-*layer* hierarchical approach typically indicates a structure where the targets, constraints, costs, and parameters of a given layer are communicated from a *higher-level* layer and the given layer communicates the targets, constraints, costs, and parameters to the *lower-level* layer(s) in the hierarchy [Mahmoud, 1977]. Few general theories exist to describe the effectiveness or expected performance of the multi-*layer* approach in system control since the definition and structure of such "layers" are problem-dependent [Sandell et al, 1978]. This approach has been implemented to address the freeway control problem with the layers being parameter estimation, incident detection, flow identification, and gap-acceptance ramp metering, respectively [Drew et al, 1969], although the gap-acceptance metering method has not found widespread acceptance. A later extension by Messer incorporated an LP-based area-wide coordination method at the "optimizing" layer of the hierarchy [Messer, 1971]. Modeling of the freeway control problem with a hierarchical, multi-layer approach has since persisted in the literature because of the natural way in which it addresses the complicating factors of the problem.

Set-point regulation methods

The hierarchical approach also applies to the development of *set-point* regulation control methods [Payne and Isaksen, 1973; Papageorgiou, 1983; Stephanedes and Chang, 1991]. A set-point regulation controller solves two separate optimization problems. One optimization problem (or problems) is solved to obtain the *set-point(s)* of the system. A second set of "optimization problems" are solved to obtain *control laws* that regulate the system state, under the influence of external disturbances, to operate at the set-points.

A third layer (*adaptation*) resides above the upper-layer optimization problem to modify the problem structure, parameters, and the like to the changing system conditions. The set-point regulation control method has been successfully applied in many areas of engineering such as chemical processing and aircraft control systems. Typically, because of the natural structure (i.e. geographic size, multi-dimensionality) of the system being controlled, the upper-layer control problem uses an *aggregate* model of the system to reduce processing requirements. Then at the lower-layer, the control problem is decoupled into *independent* subproblems that use a *more detailed* dynamic description of a *geographically smaller* portion of the problem given the assumptions of eqn. 3-3. Thus, because of the smaller size of the subproblems, more computational effort can be applied to solve each subproblem in real-time.

The MILOS hierarchical structure

This research addresses the system-wide ramp metering control problem by using a structured hierarchical framework hereafter referred to as the Multiobjective Integrated Large-Scale Optimized ramp control System (MILOS). This framework is based on the multi-*layer* approach to hierarchical process control using the set-point control method. Although neither the multi-layer approach to hierarchical control nor the set-point control method are contributions of this research, the hierarchical structure of the MILOS framework includes the following contributions:

(a) consideration of *multiple objectives* in the optimization problem(s);

(b) *integration* of information about the current conditions of the adjacent surface-street system;

(c) *prediction* of possible future system states in the development of pro-active real-time metering rates; and

(d) computability in real-time.

These primary characteristics of MILOS are driven by the structure of the real-world freeway/surface-street system. Thus, MILOS is composed of four hierarchically embedded, interactive subsystems:

(1) locally reactive, predictive-cooperative real-time control,

(2) area-wide coordination,

(3) anomaly detection / optimization scheduling, and

(4) subnetwork identification.

The structure of MILOS is a *pyramid* of modules that address smaller and smaller geographic areas of the large-scale ramp metering problems as one progresses lower in the hierarchy. The pyramid structure indicates that *one* optimization scheduler module schedules the solution of *several* area-wide coordination problems that in turn schedule the solution of several traffic-responsive real-time metering problems. This pyramid structure is illustrated in Figure 3-1, motivated by the structure of the RHODES hierarchical system for real-time *surface-street* traffic management and the RHODES-ITMS system developed at the University of Arizona [Head et al., 1992; Head and Mirchandani, 1993].

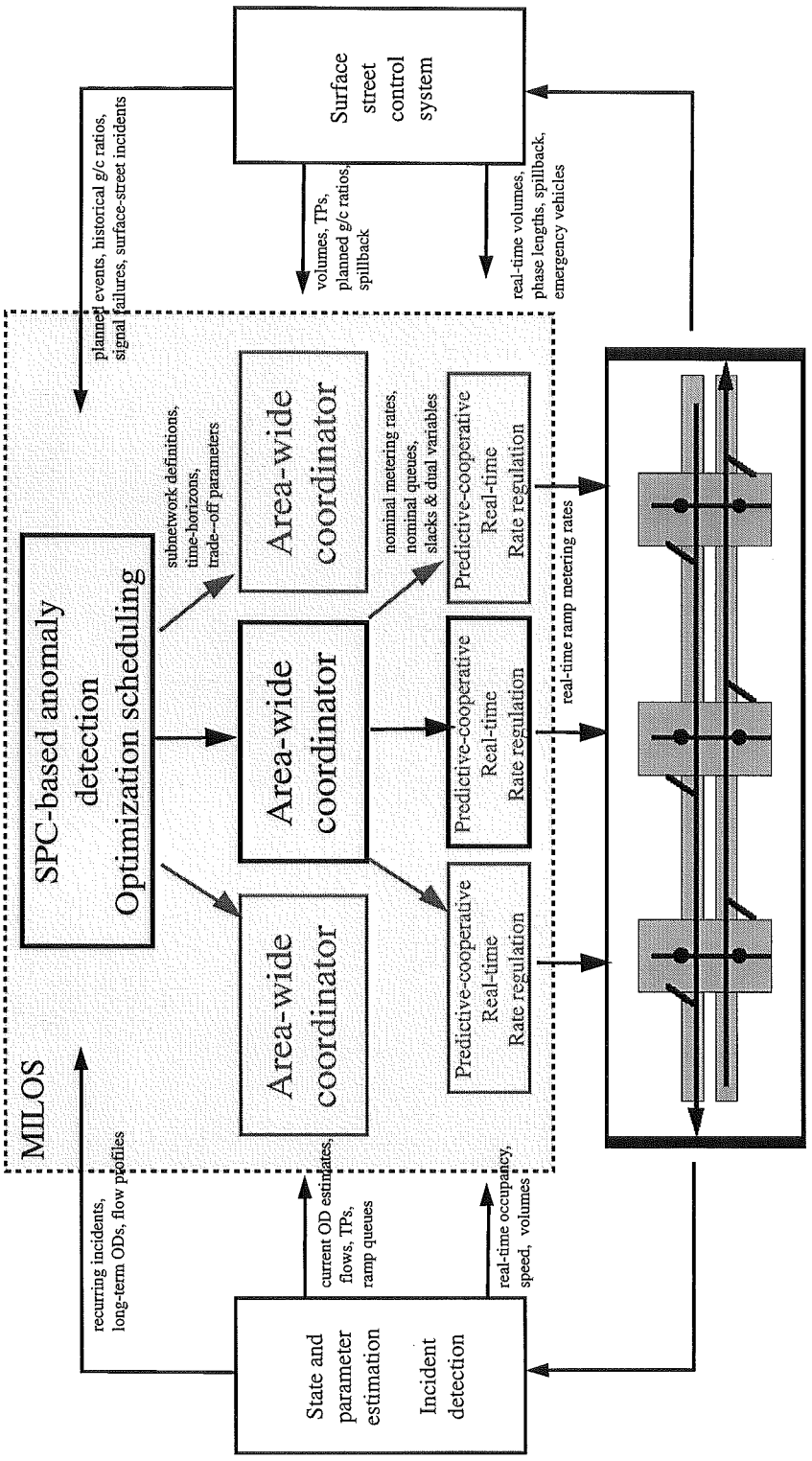


Figure 3- 1. The pyramid structure of the MILOS hierarchy

Modal decomposition of the MILOS hierarchy

MILOS can be considered to operate in several *modes*; strategic, tactical, and operational. At the highest level of the hierarchy, the *strategic* mode solves optimization problems with time horizons on the order of hours, days, and weeks, as well as responding to seasonal changes etc. The spatial influence of the strategic mode is the entire freeway and interchange network. The objectives of the strategic mode are to:

(1) identify the "optimal" sub-network definitions that lower-level processors use to solve de-coupled optimization problems,

(2) update parameters reflecting special events and long-term disturbances such as work zones,

(3) update the slowly varying parameters in the system, and

(4) determine the optimization time horizons for the lower-level problems.

The strategic mode is fulfilled by the *SPC anomaly detection* module and the *subnetwork identifier* module.

The *tactical* mode of the MILOS hierarchy solves optimization problems with time horizons of hours and minutes, using the subnetwork definitions and parameters passed from the strategic levels of the hierarchy. The spatial influence of a tactical-level module or optimization problem is "several " (e.g. 5-15) adjacent ramp meters and the associated surface-street interchanges. The objectives of the *tactical* mode are to:

(1) plan coordinated metering rates for recurrent congestion,

(2) identify short-term flow fluctuations that require re-solution of the area-wide and real-time optimization problems,

(3) react to changes in the relative congestion levels of the interchanges,

(4) balance queue growth rates in a given geographic sub-network, and

(5) respond to non-recurrent congestion generated by incidents.

The tactical mode is implemented by the *SPC-based anomaly detection* module and the *area-wide coordination* modules.

At the lowest level of the hierarchy, the *operational* mode solves optimization problems with time horizons of minutes, using the set-point metering rates and desired freeway

states provided by the tactical mode modules. The spatial influence of the operational level is a single ramp meter, a single interchange signal, and a small, *relatively predictable* subsection of the freeway. By "relatively predictable" it is meant that reasonable predictions for the next few minutes of flow can be made for this small section using a mathematical model. The objectives of the operational mode are to:

(1) reduce ramp queue lengths when not detrimental to freeway conditions,

(2) plan metering rates pro-actively based on prediction of possible future states,

(3) react to short-term flow fluctuations that could cause freeway congestion, and

(4) manage ramp queue spillback, if possible.

The operational mode is implemented by the *predictive-cooperative real-time control* modules that each solve optimization problems local to a single ramp meter.

Subnetwork identification

The majority of this research is focused on the area-wide optimization (tactical level) and traffic-responsive real-time control (operational level) algorithms. However, a role of the *strategic* mode in the MILOS framework is to identify the *problem boundaries* for the area-wide coordination and predictive-cooperative real-time control problems. Some research has been done to develop a method to determine boundaries for surface-street coordination problems [Moore and Jovanis, 1985], but little mention of such issues can be found in freeway control literature. For example, consider the large freeway network in Figure 3-2, where each node represents an interchange with a ramp meter (considering the unidirectional case only).

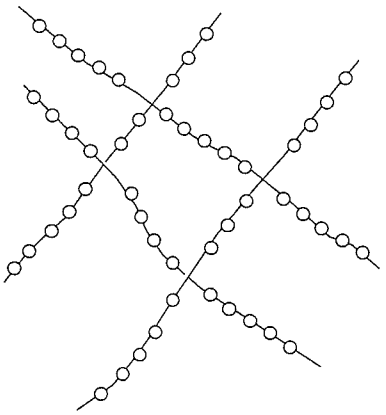


Figure 3- 2. Example freeway network

Consider an initial decomposition of the large-scale network problem into subsystems along each of the freeways, as shown by the boxes in Figure 3-3. Each of the subsystems would then be applied to a single area-wide coordination problem.

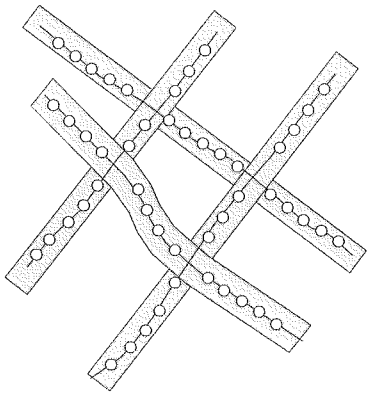


Figure 3- 3. Initial decomposition of freeway network

The question remains whether or not additional system performance could be gained by a different subnetwork structure, say for example, the structure indicated by Figure 3-4, where again each box indicates a separate area-wide coordination problem.

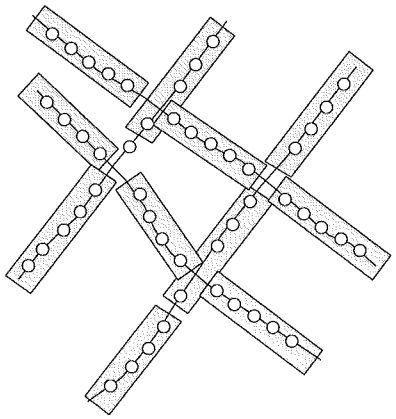


Figure 3- 4. Alternative decomposition of freeway network

No further development for the identification of subnetworks using analytical techniques was conducted in this research. Since many subnetwork definitions are easily, and "heuristically" pre-determined by the network topology, it was assumed that subnetworks for control is given. It should be noted that these subnetwork boundaries need not, and probably do not, coincide with political and jurisdictional boundaries [Fan and

Asmussen, 1990]. Thus, inter-agency cooperation would be necessary for cross-boundary coordination when a subnetwork crosses a boundary.

In this research , however, the subnetwork definitions at the area-wide coordination level will be taken as given by the traffic management decision-makers. It should be a topic of future work to develop the analytical *subnetwork identification* module of MILOS to advise freeway system managers of modifications to the subnetwork decomposition structure as current network conditions change.

SPC-based anomaly detection and optimization scheduling layer

The optimization procedures, at both the area-wide and real-time control layers, are continually re-evaluated using a *rolling-horizon* approach. Rolling-horizon approaches to traffic management have been proposed by many researchers in both surface-street and freeway control [May, 1979; Gartner, 1983; Chang et al, 1992; Head, et al., 1992; Sen and Head, 1997]. As the system evolves, the anomaly detector continually compares the observed freeway flows and ramp demands to the expected flows and demands. When a significant deviation from the expected state is detected, a new optimization run is scheduled immediately. The SPC anomaly detection module is based on the concept of control limits from the statistical process control (SPC) literature. This method is a completely novel approach to the "integrated" demand estimation and optimization scheduling problem and is discussed further in Chapter 7.

Area-wide coordination layer

The area-wide coordination layer provides the *tactical* decisionmaking of the MILOS hierarchy. The area-wide coordination level allocates medium-term (i.e. 10-20 minute) *target* or *nominal* ramp metering rates to maximize freeway throughput, balance ramp queue growth rates, and minimize queue spillback into the adjacent surface-street interchanges for a given subnetwork. The area-wide coordinator is based on a rolling-horizon implementation of a multi-criteria quadratic programming optimization problem. The area-wide coordinator interacts with the SPC anomaly detection module to identify over-capacity congestion conditions and to modify the optimization constraints and

criteria appropriately during incident conditions. The area-wide coordinator is sensitive to the needs of the adjacent surface streets by planning queue-growth rates according to the relative congestion level of each interchange. Several aspects of this formulation of the area-wide coordination problem are novel and discussed further in Chapter 5.

Predictive-cooperative real-time rate regulation layer

The predictive-cooperative real-time (PC-RT) rate regulation layer fulfills the *operational* mode of the MILOS hierarchy. The PC-RT optimization problems are based on a linearized description of the freeway state variables and are solved to minimize a linear measure of *additional* travel-time savings. This additional travel-time savings is above and beyond that due to the area-wide coordination solution by itself. At the *operational* layer, the system model is more detailed than at higher layers of aggregation [Papageorgiou, 1983; Payne et al., 1985; Fan and Asmussen, 1990]. Thus, linearization allows the PC-RT rate regulation module to plan, in real-time, several *pro-active* modifications to the nominal metering rates provided by the upper-layer area-wide coordination module based on predicted scenarios of possible ramp and freeway flows in the next few minutes. The scenario-based optimization structure of the PC-RT rate regulation module is a new treatment of the real-time ramp metering problem and its explicit connection to the solution of the area-wide coordination problem is completely new. These issues are discussed further in Chapter 6.

Integration of MILOS with necessary external systems

MILOS, as shown in Figure 3-1, is primarily an *optimization* system. Parameter estimation, especially turning-probability and route-proportional rate estimation, freeway detector data collection/filtering, incident detection and surface-street performance data all are taken as *inputs* to the MILOS hierarchy, and assumed to be "solved problems". Of course, the successful implementation of an optimization routine such as MILOS is highly dependent upon the reliability and accuracy of the external algorithms and systems. In particular, MILOS requires real-time turning-probabilities, demand flows, green splits, and queue lengths from the interchanges control system. Such information

requires the availability of a sufficiently intelligent real-time signal controller and communications network.

Summary

A Multiobjective Integrated Large-Scale Optimized ramp control System (MILOS) is developed in this research. The framework is based on the multi-*layer* approach to hierarchical process control using the set-point/regulation paradigm. The MILOS framework is specifically structured to address the complicating characteristics of dynamic state changes, stochasticity, multi-dimensionality, partial observability, the existence of multiple objectives, and unpredictability that are inherent to the large-scale freeway control problem. In addition, MILOS considers the effects of freeway control decisions on the adjacent surface-street system at each level of the hierarchy. MILOS is composed of four hierarchically embedded, interactive subsystems:

(1) area-wide coordination,

(2) predictive-cooperative real-time control,

(3) SPC-based anomaly detection and optimization scheduling, and

(4) subnetwork identification,

based upon the decomposition of the large-scale control system into its *strategic*, *tactical*, and *operational* processing modes. In the next chapter, a popular and useful macroscopic flow model is discussed that is used (in Chapter 9) to evaluate the results of implementing the MILOS systems in a simulated freeway environment.

Chapter 4: Freeway Macrosimulator

Model construction

A macroscopic freeway traffic simulator based on the enhanced FREFLO [Payne, 1971, 1979; Rathi et al., 1985] and META models [Papageorgiou, 1984; Cremer, 1989] is used to evaluate and compare various ramp metering strategies developed in this research. The FREFLO macroscopic traffic simulator is based on partial differential equation (PDE) description of freeway traffic flow as a fluid of density $\rho(x,t)$ and speed $v(x,t)$ where x indicates spatial variation and t indicates temporal variation of the fluid's density and speed such that

$$\frac{\partial v}{\partial x}+\frac{\partial q}{\partial t}=r-s$$
$$\frac{\partial v}{\partial t}=v\frac{\partial v}{\partial x}-\frac{1}{T}\left[v-v_e(\rho)+\nu\frac{\partial \rho}{\partial x}\right] \qquad \textbf{Eqn. 4- 1}$$

where r is the ramp meter input rate, s is the off-ramp (or end of freeway) output rate $q(x,t)$ is the flow rate, $v_e(\rho)$ is the equilibrium speed-density relationship, and ν is the aniticipation coefficient [Lighthill and Witham, 1955; Richards, 1956; Michalopolous et al., 1986, 1991, 1993]. This set of PDEs describes the conservation of vehicle flow through the freeway system and the dynamic relationship of speed and density. To evaluate $\rho(x,t)$ and $v(x,t)$ for various input rates r and exit rates s at each point along the freeway, the PDEs are discretized over space and time to obtain, using the simple Euler formula, the *difference* equation description of the system

$$\rho_j(k+1)=\rho_j(k)+\frac{T}{\Delta_j}\left(V_{IN,j}(k)-V_{OUT,j}(k)-s_j(k)+r_j(k)\right) \qquad \textbf{Eqn. 4- 2}$$

$$v_j(k+1)=v_j(k)+\frac{T}{\tau}\left(v_e\left(\rho_j(k)\right)-v_j(k)\right)+\frac{T}{\Delta_j}v_j(k)\left(v_{j-1}(k)-v_j(k)\right)$$
$$-\frac{\nu}{\tau}\frac{T}{\Delta_j}\left(\frac{n_{j+1}\rho_{j+1}(k)-n_j\rho_j(k)}{n_j\rho_j(k)+\kappa}\right)-\frac{\zeta T}{\Delta_j}\left(\frac{n_{j+1}r_{ON,j}(k)v_j(k)}{n_j\rho_j(k)+\kappa}\right) \qquad \textbf{Eqn. 4- 3}$$

$$V_{IN,j}(k) = \alpha \cdot V_{j-1}\left(\rho_{j-1}(k), v_{j-1}(k)\right) + (1-\alpha) \cdot V_j\left(\rho_j(k), v_j(k)\right)$$
$$V_{OUT,j}(k) = \alpha \cdot V_j\left(\rho_j(k), v_j(k)\right) + (1-\alpha) \cdot V_{j+1}\left(\rho_{j+1}(k), v_{j+1}(k)\right) \qquad \textbf{Eqn. 4- 4}$$
$$V_{\bullet,j}\left(\rho_j(k), v_j(k)\right) = \rho_j(k) \cdot v_j(k)$$

where $\rho_j(k)$ is the density, $v_j(k)$ is the mean speed, and $V_j(k)$ is the volume of vehicles in freeway section j at time k. Additional terms are added in (4-3) that are not represented in (4-1) for the speed PDE. $v_e(r_j(k))$ is an analytical speed-density characteristic such as

$$v_e\left(\rho_j(k)\right) = v_f\left(\left(1 - \frac{\rho_j(k)}{\rho_{MAX}}\right)^{l(3-2b_j)}\right)^m. \qquad \textbf{Eqn. 4- 5}$$

The parameters v_f, ρ_{max}, l, and m of (4-5), as well as the other parameters of (4-2), (4-3), and (4-4), must be calibrated from field data and may vary over time and location. In particular, the time-interval T must be selected such that $T < \frac{\min_j[\Delta_j]}{v_f}$ to ensure that the state updates are frequent enough that flows do not "skip" sections.

For simplicity, we assume that the parameters do not vary from location to location during a given simulation. In addition, it is reasonable to assume that variations of the parameters u_f, ρ_{MAX}, l, and m of (4-5) are much slower than traffic flow dynamics and thus can be assumed as constant over a simulation period. Full description of the derivation of the remaining parameters in Table 4-1 can be found in [Payne, 1971; Cremer, 1989; Papageorgiou, 1989].

Symbol	Value
Δ_j	length of section j (km)
T	time interval duration (hr),
α	$\in [0,1]$, spatial discretization parameter
τ	time constant (km/hr), approximately the segment free-flow travel time
κ	constant (veh/km) to improve performance of eqn. 4-3 at low densities
ν	anticipation coefficient (veh/km^2)
z	on-ramp friction coefficient
f	lane-drop friction coefficient
l	shaping parameter for speed-density characteristic
m	shaping parameter for speed-density characteristic
υ_f	mean free-flow speed in section j (km/hr)
ρ_{MAX}	maximum density in a single traffic lane (veh/km)
φ	Influence factor of merging slowing effect $\in [0,1]$
n_j	number of lanes in section j
$r_j(k)$	on ramp rate in section j at time k (veh/hr)
$S_j(k)$	off ramp rate in section j at time k (veh/hr)
b_j	speed limit $\in [0,1]$ in section j (km/hr)
ζ	Influence factor of lane drop slowing effect $\in [0,1]$ (see eqn.4-6)

Table 4- 1. Parameters in macroscopic simulation equations

Equation 4-2 describes the evolution of the density $\rho_j(k)$ of each freeway segment j. Freeway sections which do not contain on-ramps have $r_j(k) = 0$ in (4-2) and sections without off-ramps have $s_j(k) = 0$ in (4-2). By convention, when both an off-ramp and on-

ramp are at an interchange, we define a single section that contains both the on and off-ramp.

Equation 4-3 describes the evolution of the speed $v_j(k)$ in link j over time. The four main terms of this evolution equation are included from both theoretical and empirical considerations [Payne, 1971]. The first term of (4-3) keeps the simulated speed from straying too far from the analytical speed-density relationship, and thus corrects for errors from the speed predicted by the analytical function $v_e(\rho_j(k))$. The second term is the "anticipation" term, indicating that speed in link j changes to reflect density changes *downstream* in link $j+1$ due to car-following behavior. The third term of (4-3) is the "convection" term that represents the effect of vehicle arrivals from upstream link $j-1$ on the speed in link j. The fourth term represents the slowing effect of merging vehicles from on-ramps and is > 0 iff $r_j > 0$ [Papageorgiou, 1989]. An additional term

$$-\frac{\varphi T}{\Delta_j}\left(\frac{n_j - n_{j+1}}{n_j}\right)\left(\frac{\rho_j(k)}{\rho_{crit} n_j}\right)v_j(k)^2 \qquad \textbf{Eqn. 4- 6}$$

was added to (4-3) by Papageorgiou to represent the slowing effect from a lane-drop when $n_{j+1} < n_j$. Previous research has indicated that this addition more accurately reflects this slowing effect than (4-3) without this term [Papageorgiou, 1989].

Equation 4-4 accounts for the spatial discretization of the flow model. The flow rate out of section j, $V_{OUT,j}(k)$, is expressed as a weighted sum of the flow rate $\rho_j(k)*v_j(k)$ from section j and section $j+1$, $\rho_{j+1}(k)*v_{j+1}(k)$, such that $\alpha \in [0,1]$. Similarly, the flow rate into section j $V_{IN,j}(k)$ is expressed as a weighted sum of the flow rates from section $j-1$, $\rho_{j-1}(k)*v_{j-1}(k)$, and j, $\rho_j(k)*v_j(k)$, to smooth the behavior of the model. Equation 4-4 does not, however, represent the general case where a segment j can have n feeder flows $V_{j,IN,1}(k),\ldots, V_{j,IN,n}(k)$ and/or m receiver links $V_{j,OUT,1}(k),\ldots, V_{j,OUT,n}(k)$. Such cases are straightforward additions to the single-source, single-receiver model by using weighted averages from the multiple sources/sinks for computing upstream and downstream state variables, ρ_{j+1}, v_{j+1} and ρ_{j-1}, v_{j-1} . respectively [Papageorgiou, 1984].

Modeling flow in heavy congestion

(4-2), (4-3), and (4-4) have been shown to accurately reflect freeway traffic when calibrated to a specific location during periods of moderate congestion [Papageorgiou, 1983, 1989; Cremer, 1989]. However, in situations of *heavy* congestion, the equations must be modified to reflect the fact that vehicles cannot continue to flow from section to section when a section is at the maximum density. The main reason for the continuing flow from section to section even though the congestion is high is that the dynamic equation for the section speed $v_j(k)$ does not accurately represent the breakdown in speeds when a section becomes congested. Previous researchers have addressed this by substituting the equilibrium speed-density relationship $v_e(\rho_j(k))$ for the dynamic speed-density relationship during periods of high congestion [Rathi et al., 1985]. We take a similar approach here, adding a threshold γ to (4-4) such that if the density exceeds the threshold, the density at the maximum capacity flow rate, the flow into or out of section j transitions to the theoretical volume-density relationship. Hence, during periods of high congestion and modify the structure of the density evolution equation while continuing to compute speeds using (4-3).

Essentially, we add a threshold $\gamma$to (4-2) such that the flow into or out of section j is equal to zero during sufficiently high congestion in the adjacent section. Hence,

$$V_{IN,j}(k) = \alpha \cdot V_{j-1}\left(\rho_{j-1}(k), v_{j-1}(k)\right) + (1-\alpha) \cdot V_j\left(\rho_j(k), v_j(k)\right)$$

$$V_{OUT,j}(k) = \alpha \cdot V_j\left(\rho_j(k), v_j(k)\right) + (1-\alpha) \cdot V_{j+1}\left(\rho_{j+1}(k), v_{j+1}(k)\right) \qquad \textbf{Eqn. 4- 7}$$

$$\text{if} \quad \rho_j(k) \geq \gamma \quad V_{\bullet,j}\left(\rho_j(k), v_j(k)\right) = \rho_j(k) \cdot v_e\left(\rho_j(k)\right)$$

$$\text{otherwise} \qquad V_{\bullet,j}\left(\rho_j(k), v_j(k)\right) = \rho_j(k) \cdot v_j(k)$$

Equation 4-7 is the modified form of (4-4). This condition helps to more accurately model the stagnancy of flow when density becomes overcritical and speeds are very low.

Special conditions must be added for the implementation of (4-7) for sections at the beginning and end of a freeway facility, so that if a congestion wave is passing upstream, it does not stagnate in the first segment and limit the input flow rate. Thus, since we do not have a measurement of $\rho_0(k)$, we use $\rho_1(k-1)$. Given that the congestion wave is passing upstream, the previous measurement of ρ_1 estimates the current density of ρ_0, which is not available. This modification indicates that the source volume $V_0(k)$ must be reduced as the

congestion wave passes out of the system before it can return to the nominal value. Otherwise the density $\rho_1(k)$ in the first section will remain congested when the source volume $V_0(k)$ is reasonably large.

Dynamic modeling of ramp queues

In addition to the freeway state variables, it is also necessary to evaluate the queue lengths $q_i(k)$ at each ramp such that

$$q_i(k+1) = q_i(k) + T\left(d_i(k) - r_j(k)\right)$$
$$q_i(k) \geq 0 \qquad \textbf{Eqn. 4- 8}$$

where $r_i(k)$ is limited by

$$r_{MIN} \leq r_j(k) \leq \hat{r}_{MAX}$$
$$\hat{r}_{MAX} = \begin{cases} \min(d_i(k), r_{MAX}) & \text{if} \quad q_i(k) = 0 \\ r_{MAX} & \text{otherwise} \end{cases} \qquad \textbf{Eqn. 4- 9}$$

where $r_{i,MIN}$ and r_{MAX} are given minimum and maximum ramp metering rates, respectively. Since $r_i(k)$ can be set higher than the demand rate when a queue is present (as high as the saturation flow rate), we must restrict $q_i(k)$ to be non-negative since we cannot have negative queues. At this level of modeling, we do not consider driver behavior in the metering rate limitations. Thus, when a given metering rate is specified, (e.g. 456 veh/hr) it is assumed that drivers can implement this rate precisely (e.g. not 445 veh/hr or 500 veh/hr). Recent results have indicated that this is usually not true in the real system, especially at high metering rates when reaction time may be very close to the allocated green-time of each metering signal [Banks, 1992; Decker, 1997].

To reflect the fact that vehicles are slowed, and possibly stopped, when entering and exiting a freeway segment that is highly congested, we can add a soft-limiter

$$\xi\left(\rho_j(k)\right) = \frac{e^{\left(\frac{\rho_c - \rho_j(k)}{\rho_{MAX} - \rho_c}\right)}}{1 + e^{\left(\frac{\rho_c - \rho_j(k)}{\rho_{MAX} - \rho_c}\right)}} \qquad \textbf{Eqn. 4- 10}$$

to the on-ramp rate $r_i(k)$ and off-ramp rate $s_i(k)$. The shape of (4-10) is a sharp, but continuous, transition about the critical point ρ_c as illustrated in Figure 4-1.

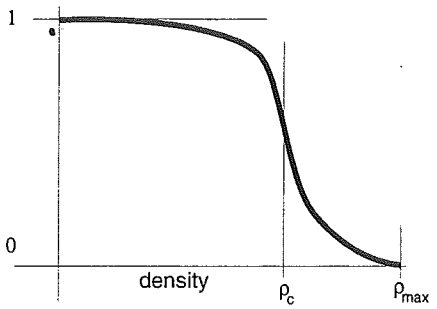


Figure 4- 1. Typical shape of soft-limiter function

The limiting effect is also applied to the off-ramp rates, since, under congested conditions, the off-ramp is also blocked after some critical density ρ_c is exceeded. Previous models without such blockage terms would underestimate the clearance time required for congestion to dissipate. Hence, (4-2) is further modified as

$$\rho_j(k+1) = \rho_j(k) + \frac{T}{\Delta_j}\left(V_{IN,j}(k) - V_{OUT,j}(k) + \xi\left(\rho_j(k)\right)\left(r_j(k) - s_j(k)\right)\right) \qquad \textbf{Eqn. 4- 11}$$

and (4-8) as

$$q_i(k+1) = q_i(k) + T\left(d_i(k) - \xi\left(\rho_j(k)\right)\cdot r_j(k)\right)$$
$$q_i(k) \geq 0 \qquad \textbf{Eqn. 4- 12}$$

to reflect the fact that a queue can develop at the ramp, even if the ramp metering rate is higher than the demand rate, when freeway congestion blocks vehicles merging from the ramp. ρ_c and ρ_{MAX} must be carefully chosen (calibrated) in equation to reflect realistic effects of queue-growth and restricted flow in the presence of freeway congestion. Nevertheless, their addition to the macro-simulation model is to more accurately represent congested conditions for comparison of various metering strategies in the evaluation experiment of Chapter 9. Some preliminary evidence of the positive effects of these additions are also shown in the benchmark test example of the area-wide coordination problem in Chapter 5.

Summary of macroscopic model

The macroscopic description of freeway traffic and ramp queues used to evaluate ramp metering strategies in this research project is a system of nonlinear difference equations based on the fluid-flow model given by

$$\rho_j(k+1) = \rho_j(k) + \frac{T}{\Delta_j}\left(V_{IN,j}(k) - V_{OUT,j}(k) + \xi\left(\rho_j(k)\right)\left(r_j(k) - s_j(k)\right)\right)$$

$$v_j(k+1) = v_j(k) + \frac{T}{\tau}\left(v_e\left(\rho_j(k)\right) - v_j(k)\right) + \frac{T}{\Delta_j} v_j(k)\left(v_{j-1}(k) - v_j(k)\right)$$

$$- \frac{\nu}{\tau}\frac{T}{\Delta_j}\left(\frac{n_{j+1}\rho_{j+1}(k) - n_j\rho_j(k)}{n_j\rho_j(k) + \kappa}\right) - \frac{\zeta T}{\Delta_j}\left(\frac{n_{j+1} r_{ON,j}(k) v_j(k)}{n_j\rho_j(k) + \kappa}\right)$$

$$V_{IN,j}(k) = \alpha \cdot V_{j-1}\left(\rho_{j-1}(k), v_{j-1}(k)\right) + (1-\alpha) \cdot V_j\left(\rho_j(k), v_j(k)\right)$$

$$V_{OUT,j}(k) = \alpha \cdot V_j\left(\rho_j(k), v_j(k)\right) + (1-\alpha) \cdot V_{j+1}\left(\rho_{j+1}(k), v_{j+1}(k)\right)$$

$$\text{if} \quad \rho_j(k) \geq \gamma \quad V_{\bullet,j}\left(\rho_j(k), v_j(k)\right) = \rho_j(k) \cdot v_e\left(\rho_j(k)\right)$$

$$\text{otherwise} \qquad V_{\bullet,j}\left(\rho_j(k), v_j(k)\right) = \rho_j(k) \cdot v_j(k)$$

$$q_i(k+1) = q_i(k) + T\left(d_i(k) - \xi\left(\rho_j(k)\right) \cdot r_j(k)\right)$$

$$q_i(k) \geq 0$$

$$r_{MIN} \leq r_j(k) \leq \hat{r}_{MAX}$$

$$\hat{r}_{MAX} = \begin{cases} \min(d_i(k), r_{MAX}) & \text{if} \quad q_i(k) = 0 \\ r_{MAX} & \text{otherwise} \end{cases}$$

$$\xi\left(\rho_j(k)\right) = \frac{e^{\left(\frac{\rho_c - \rho_j(k)}{\rho_{MAX} - \rho_c}\right)}}{1 + e^{\left(\frac{\rho_c - \rho_j(k)}{\rho_{MAX} - \rho_c}\right)}}$$

$$v_e\left(\rho_j(k)\right) = v_f\left(\left(1 - \frac{\rho_j(k)}{\rho_{MAX}}\right)^{l(3-2b_j)}\right)^m$$

This model has the parameters [ρ_c, φ, ρ_{crit}, κ, α, τ, ζ, l, m, ρ_{max}, v_f, T], the specific geometric and travel-behavior details [$d_{IN,j}(k)$, $s_{OFF,j}(k)$, Δ_j, n_j, b_j], and the control variables $r_j(k)$. The parameters *must* be calibrated precisely to the specific location characteristics and driving population of the intended application area to obtain reasonable real-world flow behavior from the model. Inappropriate choices for the model parameters can easily lead to "unstable" model performance and inconclusive results. This model has been modified slightly from previous instances of the model to reflect ; (a) the condition where queues are built at the freeway on-ramps when the density in a section is too high to allow the current on-ramp flow rate, and (b) the condition that off-ramp rates are also reduced when the density becomes large because vehicles cannot physically move to the off-ramp to exit the freeway system when the speed is near zero.

Stochastic effects and diversion behavior

Note that this macroscopic description of freeway flow is a deterministic model. Empirical data for real-world freeways indicates that the system does *not* evolve in a completely deterministic manner, but is highly affected by stochastic disturbances and flow fluctuations. We address this issue by treating the input streams $d_i(k)$ and the initial upstream freeway input(s) $V_0(k)$ as random variables, but leaving the evolution equations deterministic, as opposed to previous approaches based on adding acceleration noise to the dynamic speed equations [Weits, 1988]. It is shown empirically in Chapter 9 that considering the inputs $d_i(k)$ and $V_0(k)$ as random variables significantly improves the match of the macroscopic model to a stochastic, *microscopic* model of a study area in Phoenix, AZ. This microscopic simulation model, CORSIM, simulates travel of individual vehicles in one-second increments and is well accepted for extensive simulation testing and evaluation.

This macroscopic flow model also does not *explicitly* simulate diversion behavior or route modification, but this can be "easily" added by modifying/updating the route-proportional matrix (which determines the off-ramp rates $s_j(k)$ and demands $d_i(k)$) due to the current conditions. Of course, as has been indicated in previous work, estimation and re-estimation of route-proportional matrices and diversion rates is very difficult and is a subject of much research [Cremer and Keller, 1987; Madanat et al., 1995; Ashok and Ben-Akiva, 1993; Ding et al., 1996]. However, the ultimate success of a ramp metering control system such as MILOS in real-world freeway systems is highly dependent upon accurate and reliable turning-proportions and/or route-proportional matrix estimation. This research will assume that route-proportional matrices are given. In Chapter 7 we present a method that *could* be used to detect changes to the route proportions and/or turning-probabilities, but we do not further explore this possibility.

Summary

A macroscopic freeway traffic simulator based on the enhanced FREFLO and META models was developed for evaluation of various ramp metering strategies developed in this research project. The performance of the model was improved to represent highly congested conditions, especially the simulation of ramp queues. A term was added to the flow equations that represents the inability of vehicles to enter the freeway from the ramp when the freeway is so congested that no merging maneuver can occur. Specific simulation results showing the effects of the modeling enhancements are presented in

Chapter 9. The next three chapters develop the area-wide coordination, locally-reactive real-time optimization, and SPC-based anomaly detection and optimization scheduling layers of the MILOS hierarchical control structure.

Chapter 5: Area-wide Coordination Problem

Introduction

The area-wide coordination layer provides primarily the *tactical* decision-making of the MILOS hierarchy by providing *target* ramp metering rates based on area-wide conditions and aggregate traffic flows in each segment. The area-wide coordinator is based on two rate coordination problem formulations from the literature

(1) Yuan and Kreer's queue-balancing problem [Yuan and Kreer, 1971] and

(2) Wattleworth and Berry's throughput maximization problem [Wattleworth and Berry, 1968].

These formulations are significantly modified in the model presented here. By using a multi-criterion objective function, we combine the two conflicting objectives to address both total system performance and user-specific performance benefits. The objective function also includes

(a) consideration of the specific differences in interchange congestion,

(b) physical capacity along the corridor, and

(c) agency/system-operator preference for incorporating queue-growth considerations [Fan and Asmussen, 1990].

Mathematical description of the area-wide coordination problem

Consider a unidirectional freeway with N on-ramps and M off-ramps where the demand (veh/hr) d_i at each on-ramp i, $i=1...N$ is provided by either

(1) the physical beginning of the freeway facility,

(2) a freeway-freeway connector, or

(3) a surface-street interchange.

We assume that demand $d_0=V_0$ provided at the beginning of the freeway cannot be controlled via ramp metering. The only freeway controls available are ramp metering rates (veh/hr) r_i $i=2...N$. By convention we assume that $r_1 = d_0$, (i.e. the freeway input or the first "ramp" is uncontrollable. Speed limits are assumed fixed in each freeway section, but need not be equal everywhere. Speed advisories, such as those that could be provided by variable message signs (VMS), are not considered in this algorithm.

The vehicular flows x_j in each freeway link j are determined by evaluating the route-proportional flows from each on-ramp to each off-ramp, such that

$$\sum_{i=1}^{j} A_{i,j} r_i = x_j \quad \forall j \qquad \textbf{Eqn. 5- 1}$$

where $A_{i,j}$ values have the special structure

$$0 \le A_{i,j} \le 1$$

$$\begin{aligned} A_{i,j} &= 0 \qquad i < j \\ A_{i,j} &\le A_{i,j-1} \quad i > j \end{aligned} \qquad \textbf{Eqn. 5- 2}$$

The matrix $A = \{A_{i,j}\}$ describes the proportion of the flow entering at ramp i that continues through link j en route to its destination, it will be referred to as the route-proportional matrix. The matrix A is assumed to be constant and known over the control period horizon, T. In a *steady-state* input-output description of the freeway system such that $x_j(k) = \bar{x}_j \quad \forall\ k \le T$, it is assumed that all demand entering at ramp i bound for off-ramp j will exit at off-ramp j during the time horizon T. Thus, we need only be concerned with the physical limit of freeway capacity

$$\sum_{i=1}^{j} A_{i,j} r_i \le CAP_j \quad \forall j \qquad \textbf{Eqn. 5- 3}$$

in each segment. The physical limit CAP_j is derived for each segment from the volume-density curve specific to that segment. The volume-density curve can be empirically derived (i.e. curve-fit) from observations or computed from the saturation flow rates, number of lanes, merge area restrictions, and other factors as detailed in established procedures [McShane and Roess, 1990]. Given the concerns noted in Chapter 1, it may be advantageous to set a capacity CAP_j for each link j in the optimization model that is slightly less than the *critical* maximum volume to ensure stable flow. As will be shown in Chapter 9, it is difficult to maintain flows at the critical value CAP_j without beginning a backward-traveling congestion wave, confirming the difficulties of the freeway management problem as presented in Chapter 1.

An additional necessary set of constraints for the area-wide coordination problem is a limitation on the minimum and maximum ramp metering rate, such that

$$r_{i,MIN} \le r_i \le r_{i,MAX} \quad , \quad i = 1 \ldots N \qquad \textbf{Eqn. 5- 4}$$

where $r_{i,MAX} = \min(d_i, s_i)$. Here, s_i is the saturation flow rate of the ramp i (a ramp could have more than one lane) and $r_{i,MIN}$ is the slowest rate acceptable to drivers, such as two vehicles per minute (120 veh/hr). The rate $r_{i,MIN}$ could be as low as zero if the ramp was allowed to be and/or capable of being fully closed.

In this optimization formulation, metering/closure of a ramp only creates a queue at the ramp and does *not* result in driver diversion. Diversion rates would be computed by an external processor (not discussed in this research) that updates the route-proportional matrix and demands given the control decisions, e.g. [Cremer and Keller, 1987; Ashok and Ben-Akiva, 1993; Madanat et al., 1995; Ding et al., 1997]. Chapter 7 discusses a system identification procedure based on statistical process control which could be used to aid diversion modeling by detecting changes to the flows x_j in each link deviating from their assumed nominal values $\bar{x}_j$.

Derivation of the objective function

Given the constraints detailed above, a popular objective function is to maximize the total inputs to the freeway

$$\max_{r \in R} \sum_{i=1}^{N} r_i . \qquad \textbf{Eqn. 5- 5}$$

This objective is derived from minimizing the total travel time in the freeway system, which is a typical operational goal of freeway control [Wattleworth and Berry, 1968; Messer, 1971; May, 1979; Papageorgiou, 1983]. Using this objective, the current freeway conditions $\rho_j(k)$, and on-ramp demands $d_i(k)$ must be continually monitored and compared to the assumed steady-state values $\overline{\rho_j}$ and d_i. When the values of $\rho_j(k)$ and $d_i(k)$ drift outside of a reasonable upper or lower bound, the problem must be redefined and re-optimized [Messer, 1971; May, 1979] as developed in Chapter 7.

Consideration of queue storage limits

The classical linear programming rate coordination problem is described by (5-3), (5-4), and (5-5) [Wattleworth and Berry, 1968]. This formulation does *not* consider the formation of ramp queues $q_i(k+1) = q_i(k) + \Delta T\left(d_i(k) - r_i(k)\right)$ at each on-ramp due to the

application of metering rates less than the offered demand. Thus, to reflect the physical limitations of ramp queuing areas, an additional set of constraints must be added such that

$$(d_i - r_i)T \le Q_i \quad \forall i \qquad \textbf{Eqn. 5- 6}$$

such that Q_i is the physical limit on the number of vehicles that can be stored on the ramp without causing spillback into the interchange (assuming some average vehicle length) and T is the optimization time horizon. Q_i would be based on the length of the ramp storage area and the average vehicle length. In operational practice, one may want to keep the capacity limitation Q_i slightly less than the physical limit of storage to provide an additional cushion for unexpected surges in demand. The constraints (5-6) limit the rate at which the queue is allowed to grow (based on a constant arrival rate) and *fill to capacity* during the optimization horizon T. Vehicles queued at the ramp at the beginning of the optimization period are included in the offered demand d_i such that

$$d_i = d_{ext} + \frac{q_i(0)}{T} \quad \forall i \qquad \textbf{Eqn. 5- 7}$$

converting the queued vehicles $q_i(0)$ into a flow rate (veh/hr) by assuming that all of the vehicles queued "demand" to be discharged during the time horizon T.

Inclusion of constraints (5-6) would indicate that, in the absence of re-optimization during the time horizon T, at $t = t_0 + T$, several queues *may* be filled to capacity. This would require, at least for a short time, $r = r_{i,MAX}$ (saturation flow rate) to clear the queue and to create ramp capacity. This clearing at the maximum rate could have significant detrimental effects to freeway conditions as the vehicles attempt to merge into traffic as a platoon. We address the problem of alternately filling queues to capacity and dissipating them at the saturation flow rate in two ways. First, we implement a *rolling-horizon* solution to the area-wide coordination problem (5-3, 5-4, 5-5, 5-6) with frequent estimates of the current constant demand d_i and queue lengths q_i. Thus, as the unfilled queue storage capacity begins to decrease as q_i approaches Q_i, the demand rate at the ramp increases and it becomes more likely that queue dissipation will occur. Second, by modifying the nominal rate $r_{i,N}$ in real-time such that $r_i(k) = r_{i,N} + \Delta r_i(k)$ by solving a predictive-cooperative optimization problem at each ramp for $\Delta r_i(k)$, we can take advantage of the opportunities to *dissipate queues* when

(a) the demand rate to the ramp meter is lower than expected and/or

(b) the freeway conditions are lighter than expected.

More details of the predictive-cooperative real-time optimization subproblems solved at each ramp are provided in Chapter 6.

Development of a multi-criteria objective function

Inclusion of constraints (5-6) into the linear programming problem formulation provides queue-growth management due to physical limitations of each ramp, but does *not* control queue-growth according to the prevailing *congestion levels* at each interchange. Such constraints (5-6) also do *not* guarantee that equitable decisions will be made as to where to hold vehicles in queues to provide freeway congestion relief. A quadratic optimization criterion

$$\min_{r \in R} \sum_{i=1}^{N} (d_i - r_i)^2 \qquad \textbf{Eqn. 5- 8}$$

was proposed by Yuan and Kreer to address the need to balance ramp queues at each ramp, such that $q_1 \cong q_2 \cong \ldots \cong q_{n-1} \cong q_n$ (rather than hold many vehicles at some ramps and none at others, such that $q_1 >> 0$, $q_2 >> 0$, $q_3 = \ldots = q_{n-1} = q_n = 0$ [Yuan and Kreer, 1971] which is a typical result of linear programming formulations such as (5-3), (5-4), and (5-5) where the objective results in optimal solutions at the extrema of the feasible region). We can thus use a combination of the two objectives (5-5) and (5-8) to obtain a compromise solution that addresses *both* freeway throughput and ramp queue management.

Thus, we would like to simultaneously minimize freeway total travel time (by maximizing on-ramp flow in steady-state) and balance ramp queues throughout the corridor. It would be imprudent to simply add the cost functions (5-5) and (5-8) together because the units are not the same (i.e. (veh/hr) and $(\text{veh/hr})^2$, respectively). As such, we use a simple technique to combine objectives with differing units by dividing each objective by the "ideal" cost and adding the dimensionless quantities. However, in (5-8) the optimal cost is *zero* when $r_i^* = d_i$, $i=1...N$, and thus we cannot divide by the ideal cost solution to obtain a dimensionless objective for (5-8).

Thus, we modify objective (5-8) from *minimizing* the distance from the ideal point $r_i^* = d_i$, $i=1...N$, to *maximizing* the distance from the *anti-ideal* point. The anti-ideal point is

the (also usually infeasible) solution $r_{i*} = r_{i,MIN}$ $i=1...N$ which creates the longest possible queues at each ramp, and thus, the worst possible value for (5-8). Hence our single objective now combines the distance for (5-5) from the ideal point $r_i^* = d_i$, $i=1...N$ and the distance for (5-8) from the anti-ideal point $r_i^* = r_{i,MIN}$ $i=1...N$ resulting in a compromise objective function

$$\max_{r \in R} \quad \sum_{i=1}^{N} r_i + \left(\sum_{i=1}^{N} d_i \right) \left(1 - \frac{\sum_{i=1}^{N} \left(d_i - r_i \right)^2}{\sum_{i=1}^{N} \left(d_i - r_{i,MIN} \right)^2} \right) \qquad \textbf{Eqn. 5- 9}$$

Even though the terms from (5-5) and (5-8) are now in equivalent units, the *relative* difference in the size of the two cost components for typical feasible choices of r_i will *still* influence in the importance attributed to each objective. We can provide decision-makers with a preference between the two components by including a weighting factor β such that (5-9) becomes

$$\max_{r \in R} \quad \sum_{i=1}^{N} r_i + \beta \left(\sum_{i=1}^{N} d_i \right) \left(1 - \frac{\sum_{i=1}^{N} \left(d_i - r_i \right)^2}{\sum_{i=1}^{N} \left(d_i - r_{i,MIN} \right)^2} \right). \qquad \textbf{Eqn. 5- 10}$$

Setting β large will increase the importance of balancing ramp queues and setting β small will decrease the importance on balancing queues and increase the importance of maximizing freeway throughput.

Setting costs according to interchange congestion level

Although the objective (5-14) includes considerations for queue growth, the mechanism to distinguish queue growth at one ramp over another is only the storage limitations in (5-6) and the freeway conditions surrounding each ramp location. To reflect the current congestion conditions at each interchange, we weight each of the components of objective (5-10) with a weighting factor c_i such that

$$c_i = \frac{\sum_{m=1}^{M} \frac{V_{m,i}}{C_{m,i}}}{\max_i \left(c_i \right)} \quad \forall i \qquad \textbf{Eqn. 5- 11}$$

where $C_{m,i}$ is the capacity of phase m at interchange i and $V_{m,i}$ is the offered volume for phase m at interchange i. Thus, the weighting factors c_i reflect the relative importance of vehicle storage on one ramp versus another according to the possible impacts on the surface streets if spillback should occur. This result was developed independently in this research, but found to have been proposed previously for the queue balancing objective (5-8) [Fan and Asmussen, 1990]. As the congestion levels change at each interchange, the cost coefficients c_i are updated to reflect the most current conditions. Such updates are enacted at least as often as the area-wide coordination optimization problem is re-solved. However, the scheme for updating c_i estimates is not in the scope of this research.

The weighting factors c_i could also be set by decision-makers/system-operators based on other considerations such as

(a) ad-hoc values set to discourage or encourage long-term flow changes at certain interchanges,
(b) average delay at each interchange,
(c) surface-street incident conditions, as well as
(d) virtually any other performance measure computable in real-time, and as suggested by agency preference [Powell, 1997].

In this research, we restrict derivation of the c_i cost coefficients to the congestion level as reflected in the time-varying V/C ratio.

Integration of surface-street flows in ramp demands

Use of the weighting factors c_i derived in real-time as the area-wide coordination optimization problem is re-solved is one way in which the area-wide coordination optimization problem incorporates the interchange conditions. In addition, the area-wide ramp metering coordination optimization problem incorporates interchange conditions in its decision-making by building the ramp demands d_i from the surface street flows using

$$d_i = p_{R,NB} d_{NB} + p_{L,SB}\left(1 - p_{R,SB}\right) d_{SB} + p_{T,EB} d_{EB} + \frac{q_i(0)}{T} \quad \forall i \qquad \textbf{Eqn. 5- 12}$$

where $q_i(0)$ is the queue length at the ramp when the optimization begins, $p_{R,NB}$, $p_{L,SB}$, and $p_{T,EB}$ are the current probabilities of turning right, left, and through, respectively at each of the approaches to the interchange feeding ramp i , and d_{NB}, d_{SB}, and d_{EB} are the demands on the northbound, southbound, and eastbound approaches to interchange i, respectively.

These definitions assume an Eastbound freeway simply for demonstration purposes, as illustrated in Figure 5-1.

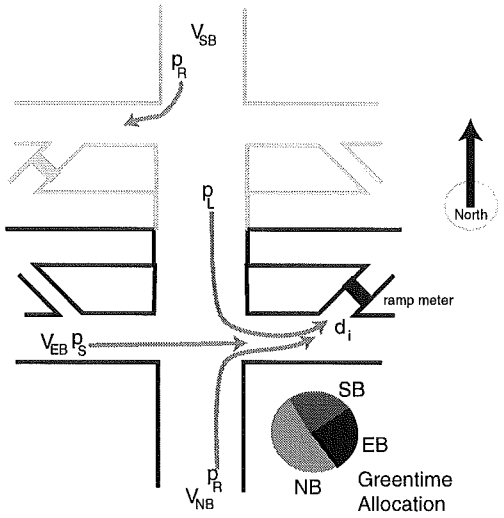


Figure 5- 1. Ramp meter demand sources

Using the turning probabilities $p_{R,NB}$, $p_{L,SB}$, and $p_{T,EB}$ and surface-street demands d_{NB}, d_{SB}, and d_{EB} to build the ramp demand d_i makes the assumption that these quantities are reasonably constant (and available) over the optimization time horizon.

It should be noted that although the method(s) chosen to derive the turning probabilities $p_{R,NB}$, $p_{L,SB}$, and $p_{T,EB}$ and estimate the real-time demands d_{NB}, d_{SB}, and d_{EB} will play a significant role in the validity and applicability of the resulting area-wide rate coordination, the coordination algorithm is *independent* of the method used. Slight inaccuracies in turning probability and demand rate estimation should be mitigated because the area-wide rates r_i are modified in real-time by the lower-layer optimization sub-problems. In the event that surface street demand and/or turning probability estimates are *significantly* different from current real-time conditions, the SPC-based anomaly detection module will identify this and begin a new iteration of the area-wide coordination optimization problem with appropriately modified parameters d_{NB}, d_{SB}, and d_{EB} and $p_{R,NB}$, $p_{L,SB}$, and $p_{T,EB}$. The operation of the SPC module is detailed in Chapter 7.

Quadratic objective function summary

To simplify notation of the objective function of the area-wide coordination optimization problem, (5-10) with the additions of (5-11) can be expressed as

$$\max_{r \in R} \quad \sum_{i=1}^{N} r_i + \beta\gamma\left(\sum_{i=1}^{N} c_i d_i - c_i\left(d_i - r_i\right)^2\right) \qquad \textbf{Eqn. 5- 13}$$

where

$$\gamma = \frac{\left(\sum_{i=1}^{N} d_i\right)}{\sum_{i=1}^{N} c_i\left(d_i - r_{i,MIN}\right)^2}. \qquad \textbf{Eqn. 5- 14}$$

Expanding the square and neglecting terms that do not contain the decision variables r_i we obtain the optimization problem

$$\max_{r \in R} \quad J = \sum_{i=1}^{N}\left(1 + 2\beta\gamma c_i d_i\right) r_i - \beta\gamma c_i r_i^2 \qquad \textbf{Eqn. 5- 15}$$

subject to the constraints (5-3), (5-4), and (5-6) where c_i is specified as in (5-11), d_i is specified as in (5-12), and all other parameters A, β, T, Q_i, $r_{i,MIN}$, $r_{i,MAX}$ and $q_i(0)$ are specified from external data. (5-15) is a quadratic objective and (5-3), (5-4) and (5-6) are linear constraints and thus the solution has a unique optimum when a feasible solution exists.

Resolving infeasibility

It is possible, however, that the formulation posed in (5-3), (5-4), (5-6), and (5-15) does *not* have a feasible solution. For example, an accident on a freeway link could reduce the capacity considerably in that section, requiring many more vehicles to be metered at upstream ramps than could be stored in the available ramp queues. In such a case, constraints from (5-3), (5-4), or (5-6) must be relaxed to render the problem feasible. Constraints (5-6) are the best candidate for relaxation, since we cannot increase the physical carrying capacity of a freeway section in (5-3) or change the maximum (minimum) possible metering rate, limited by the saturation flow rate (zero), in (5-4). Thus, we must opt to allow spillback for a short time into the interchanges by increasing the queue storage capacity in constraints (5-6) to accommodate the overflow in a system-equitable manner.

Let $z_i \geq 0 \quad i = 1...N$ be the extra capacity allocated at each ramp queue i to accommodate the flow at that ramp. In the same way that the queue storage is balanced according to the

interchange congestion cost c_i in a feasible problem, consider an allocation of the queue *overflow* in a similar manner. Thus, the constraints (5-6) are modified such that

$$(d_i - r_i)T - z_i \le Q_i \quad \forall i. \qquad \textbf{Eqn. 5- 16}$$

A penalty term is added to the objective function (5-15) incorporating the cost of allowing queue i to extend beyond its capacity Q_i over the time horizon T, such that the objective function now becomes

$$\max_{r \in R} \quad J = \sum_{i=1}^{N} (1 + 2\beta\gamma c_i d_i) r_i - \beta\gamma c_i r_i^2 - \beta_2 \gamma c_i z_i^2 \qquad \textbf{Eqn. 5- 17}$$

where β_2 is an appropriately chosen scaling constant. In particular, β_2 should be specified large enough, say β_2 = 100 β, to induce $z_i \cong 0$ for all solutions that are feasible *without* the inclusion of the additional capacity variables z_i , $i=1...N$. Choosing a "small" value of β_2 can result in a solution where some ramps are specified to spill-back, and others are allowed to flow unconstrained. This is much like the solution resulting from using an LP method, but with the c_i terms the congestion at each interchange is considered. It should be noted here that the objective function (5-17) has no physical meaning with the introduction of the penalty term $\beta_2 \gamma c_i z_i^2$ and is likely a negative quantity in the overcapacity situation. However, each of the components of (5-17) is suitably derived to benefit both freeway and surface-street system operation.

Area-wide coordination problem summary

The quadratic optimization problem is summarized as

$$\max_{r \in R} \quad \sum_{i=1}^{N} (1 + 2\beta\gamma c_i d_i) r_i - \beta\gamma c_i r_i^2 - \beta_2 \gamma c_i z_i^2 \qquad \textbf{Eqn. 5- 18}$$

subject to

$$\sum_{i=1}^{j} A_{i,j} r_j \le CAP_j \quad \forall j$$

$$(d_i - r_i)T - z_i \le Q_i \quad \forall i$$

$$r_{MIN} \le r_i \le r_{MAX} \quad \forall i$$

$$c_i = \frac{\sum_{m=1}^{M} \frac{v_{m,i}}{C_{m,i}}}{\max_i(c_i)} \quad \forall i \qquad \gamma = \frac{\left(\sum_{i=1}^{N} d_i\right)}{\sum_{i=1}^{N} c_i\left(d_i - r_{i,MIN}\right)^2}$$

$$d_i = \rho_{R,NB} d_{NB} + \rho_{L.SB}\left(1 - \rho_{R,SB}\right) d_{SB} + \rho_{T,EB} d_{EB} + \frac{q_i(0)}{T} \quad \forall i$$

which can be solved with any constrained nonlinear programming or specialized quadratic programming method, not detailed here. Note that the inclusion of the overcapacity variables z_i, $i=1...N$ ensures at least the feasible solution

$$r_i = r_{i,MIN}, \quad z_i = d_i T - r_{i,MIN} T - Q_i \quad i = 1...N \qquad \textbf{Eqn. 5- 19}$$

for reasonable (realistic) values of the freeway capacities CAP_j. In the evaluation results of this chapter and Chapter 9, problem (5-18) is solved using the QP barrier algorithm of the CPLEX math programming optimization software package [CPLEX, 1997]. For more details of the iteration details of barrier optimization algorithms, see [Bazaraa et al., 1993].

Operation under severe congestion

If any

$$CAP_j \le \sum_{i=0}^{j} A_{i,j} r_{i,MIN} \qquad \textbf{Eqn. 5- 20}$$

then there is a *severe* limitation of capacity (an *incident*) in that section and even (5-18) with the inclusion of the z_i variables will be infeasible. In this case, we can prescribe a heuristic solution such that $r_i = r_{i,MIN}$ for all ramps upstream of the congestion with condition (5-20) and $r_i = r_{i,MAX}$ downstream of the severely congested section.

In addition, the higher-layer processor(s) of the ramp metering control system should send information to the surface-street controllers regarding the incident location (severe limitation of capacity) and the prescribed emergency settings of $r_{i,MIN}$. In the presence of ATIS, such information could also be provided to travelers to increase the diversion effect away from the congested segment. As the congestion clears, the anomaly detection module will detect the favorable change to the state variables x_j and re-run the area-wide optimization for a new, feasible solution to the area-wide coordination problem (5-18).

Integration with predictive-cooperative real-time rate regulation layer

After solving (5-18) to optimality, we obtain the *nominal* ramp metering rates $\bar{r}_i$, volumes $\bar{V}_j$, and queue lengths at the end of the time horizon $q_{i,N}(K)$. From this, we derive the steady-state *set-point* densities $\bar{\rho}_j$, speeds $\bar{v}_j$ from the characteristic volume-density and speed-density curves of freeway flow. These set-points, the dual variables λ_κ and slack values ε_k of constraints (5-3), (5-4), and (5-6), are provided to the predictive-cooperative real-time rate regulation layer to derive the real-time flow measurements. More detail of how these issues are addressed is provided in Chapter 6.

Preliminary evaluation of the area-wide coordination problem on a small example

In this section, we describe an example problem and the solution results using the quadratic problem formulation (5-18) versus:

(1) a linear programming formulation without queue storage considerations

(2) the no-control case, and

(3) a policy to set ramp metering rates at 600 veh/hr, regardless of demand.

Consider a long (30 km) eastbound two-lane freeway with 5 controllable ramps at 5 km spacing, as shown in Figure 5-2. At the initial freeway entrance, there is considerable external demand. In this example, we assume each lane of the freeway can carry 2000 vehicles per hour.

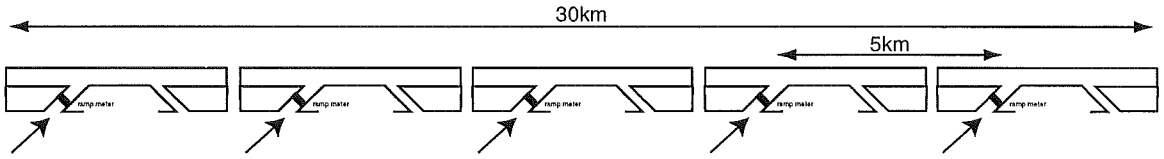


Figure 5- 2. Example problem

This example is similar to the example used in the work by Papageorgiou in the development of a large-scale hierarchical freeway control system [Papageorgiou, 1983]. The parameters and essential data for this example problem are presented in Tables 5-1 and 5-2. Recall from earlier discussion in this chapter that the parameters $A_{i,j}$ correspond to the proportion of vehicle flow from on-ramp i continuing through freeway section j.

from\to	begin	1	2	3	4	5
begin	1	0.95	0.9	0.85	0.8	0.75
1		1	0.95	0.9	0.85	0.8
2			1	0.95	0.9	0.85
3				1	0.95	0.9
4					1	0.95
5						1

Table 5- 1. Route-proportional matrix of example problem

MILOS requires information about the interchange flows that comprise the ramp demand, and needs performance information to derive the objective function cost coefficients. The data in Table 5-2 includes the volumes and turning probabilities at each interchange, that make up the demands to each ramp. Coupling these data with the green-time percentages (GT%), the *V/C* ratios of each interchange are computed and the scaled cost coefficients c_i are computed from the *V/C* ratios. These data have been selected to produce values for the demands at each ramp similar to the example problem of Papageorgiou [1983] and, at the same time, to produce differentiation in the congestion level at each interchange. As indicated by the *V/C* ratio at each location, interchanges 3 and 4 are more congested than 1, 2, and 5. As a result, the QP coordination algorithm should store less vehicles on ramps 3 and 4 (relative to demand at that ramp and the storage capability of the ramp) than at ramps 1, 2, and 5.

The values chosen in Table 5-2 do not represent real locations, but they are intended to be reasonable approximations of real behavior. The interchange parameters have also been selected to illustrate the ability of the quadratic programming formulation to react to differences in the congestion level at adjacent ramps. The quadratic programming problem (5-18) was solved using the data in Tables 5-1 and 5-2 with several values of β, the trade-off parameter between the two sub-objectives. The solutions, as listed in Table 5-6, indicate that the LP method restricts on-ramp volume at ramps 3, 4, and 5 *only*, where the freeway capacity is exceeded. The QP methods, since they take into account the queue storage restriction, restrict on-ramp volume at *all* ramps. The last line of Table 5-3 indicates when the freeway would become over-capacity if *no* ramp metering was implemented. Thus, in this example problem, if no ramp metering was implemented a backward-traveling congestion wave would be started in the section containing ramp 3.

On-Ramp data	begin	1	2	3	4	5
Northbd vol (veh/hr)		2200	1600	1800	1000	1100
Southbd vol (veh/hr)		2000	2000	1900	1000	900
Eastbd vol (veh/hr)		200	260	282	295	307
$P_{nb,r}$		0.25	0.35	0.15	0.25	0.25
$P_{sb,l}$		0.1	0.05	0.07	0.1	0.1
$P_{eb,t}$		0.05	0.07	0.05	0.15	0.05
Ramp demand d_i (veh/hr)	3000	684	610	355	355	342
Initial queue	-	0	0	0	0	0
Time horizon (hrs)		0.5	0.5	0.5	0.5	0.5
Ramp storage (veh)	-	40	30	40	40	50
$GT\%_{,EB}$	-	0.18	0.25	0.15	0.2	0.23
$GT\%_{,NB/SB}$	-	0.55	0.6	0.7	0.55	0.6
$GT\%_{,SB\ left}$	-	0.27	0.15	0.15	0.25	0.17
v/C ratio		1.67	1.55	1.74	1.24	1.15
scaled cost c_i		0.95	0.89	1	0.71	0.66

Table 5- 2. Ramp interchange data

	ramp 1	ramp 2	ramp 3	ramp 4	ramp 5
Demand	684	610	355	355	342
LP	684	610	320	275	275
QP, β=1, β_2=100β	642	555	295	275	253
QP, β=100, β_2=100β	644	550	275	275	271
QP, β=0.01, β_2= 100β	669	550	275	275	252
6-second cycle	600	600	355	355	342
freeway over capacity?	NO	NO	YES	YES	YES

Table 5- 3. Comparison of metering rate coordination methods

Influence of β

In this example, the differences between the QP solutions for radically different values of β are negligible because of the queue-growth constraints (5-6). In addition, setting $\beta_2 = 10000\beta$ in each case prescribes a solution that allows no queues to spillback regardless of their congestion level. It is possible with a lower setting for β_2 (relative to the value of β) that spillback could occur at a ramp with a relatively uncongested interchange. As such, care must be taken in setting the value for β_2 when queue balancing is de-emphasized.

β has the most influence on the resulting rate allocation when metering is required to avoid freeway congestion near interchanges where the congestion is also considerable. In such a situation, an LP approach that does *not* consider queue restrictions or interchange congestion may apply restrictive metering at locations where the most adverse impact on the surface-streets would result. The QP approach of (5-18) would enact metering upstream at interchanges where the congestion was (possibly) lower (i.e. $c_{i'} < c_i \mid i' < i$) and resulting ramp queues would have less effect on the surface-street congestion.

To further illustrate that the QP solution balances queues according to interchange congestion, consider Table 5-4. Table 5-4 compares the rates for each ramp solved by the QP method for $\beta=1$, $\beta_2=100\beta$. The queue storage size Q_i, demand rate d_i, $A_{i,j}$, and cost coefficient c_i of each interchange all interact to produce the metering rates r_i. Comparing the rate at ramp 5 from the QP and LP in Table 5-3 provides some evidence that more vehicles are held at uncongested ramps with the QP approach. Table 5-4 confirms this, since ramp 5 also has the largest ramp storage capacity Q_5.

	ramp 1	ramp 2	Ramp 3	ramp 4	ramp 5
demand (veh/hr)	684	610	355	355	342
Metering rate (veh/hr)	642	555	295	275	253
Vehicle storage size	40	30	40	40	50
Cost coefficient	0.95	0.89	1	0.71	0.66
Percentage holdback	6.1	9.0	16.9	22.5	26.0
Maximum queue, 30-min horizon	21	28	30	40	45

Table 5- 4. Rate comparison for β=1

Comparison of area-wide metering rate settings in macrosimulation

We now compare the area-wide coordination optimization algorithm based on quadratic programming (5-18) with the no-control and LP-control cases in the macrosimulation environment of Chapter 4. For the example problem of Figure 5-2, we use parameters of the macrosimulation similar to those used by Papageorgiou [1984] with the addition of the cut-off level ρ_c which phases in the flow limitation in (4-11) and (4-12). These parameters are listed in Table 5-5.

0.013	21.6	10	2	4	0.004 hr	0.8	100	123	80
τ	ν	κ	l	m	T	α	ρ_{Max}	υ_{Max}	ρ_c

Table 5- 5. Parameters for example problem

Each of the area-wide rate coordination solutions listed in Table 5-3 were tested using the macrosimulation environment. Each simulation was run with the initial conditions listed in Table 5-6. The initial speeds $v_j(0)$ were computed from the theoretical speed-density relationship $v_j=v_e(\rho_j)$ of (4-5) Each simulation was executed for one hour of simulated time using the constant route proportional matrix A given in Table 5-1 and the constant input demands d_i given in Table 5-2.

state variable / section	section 1	section 2	section 3	section 4	section 5
density (veh/km)	51	59	65	65	65
speed (km/hr)	69.7	68	70.5	70.	70.5
speed limit (km/hr)	123	111	98.4	98.4	98.4

Table 5- 6. Initial conditions for simulation run

Figure 5-3 depicts the time-space plots of density through the corridor over the duration of the simulation for the no-control, LP-control, and QP-control case ($\beta=1$, $\beta_2=1$), respectively. Figure 5-4 illustrates the differences in the queue growth using the LP method and the QP method. Queue growth for the no-control case is not shown in Figure 5-4 but it should be noted that a small queue of 6 vehicles develops and dissipates at the second on-ramp when the backward-traveling congestion wave passes the ramp around the 40-minute mark of the simulation. This queue buildup is due to the inclusion of the soft-limiter in (4-12) and would *not* have been represented in previous versions (prior to this research) of the macroscopic simulator.

Table 5-7 lists the system performance measures of total travel time, queue delay time, and throughput for each of the methods over the duration of the simulation. Throughput is computed by totaling the vehicles leaving the freeway over the duration of the simulation and dividing by the total number of vehicles entering the simulation. Because the simulation ends with vehicles still in the system, the throughput will not be close to unity. Notice, however, that the QP method provides approximately 2% more throughput than the LP method and 4% more throughput than no control during the transient. Although the LP method results in lower freeway travel time, the QP method has lower total queue time, even though more ramps are metered. As indicated in Figure 5-4, the LP method spills-back two ramps and builds very small queues at two ramps. The corresponding QP

solution does not result in queue spillback at any ramp, and results in queues that are more balanced throughout the corridor.

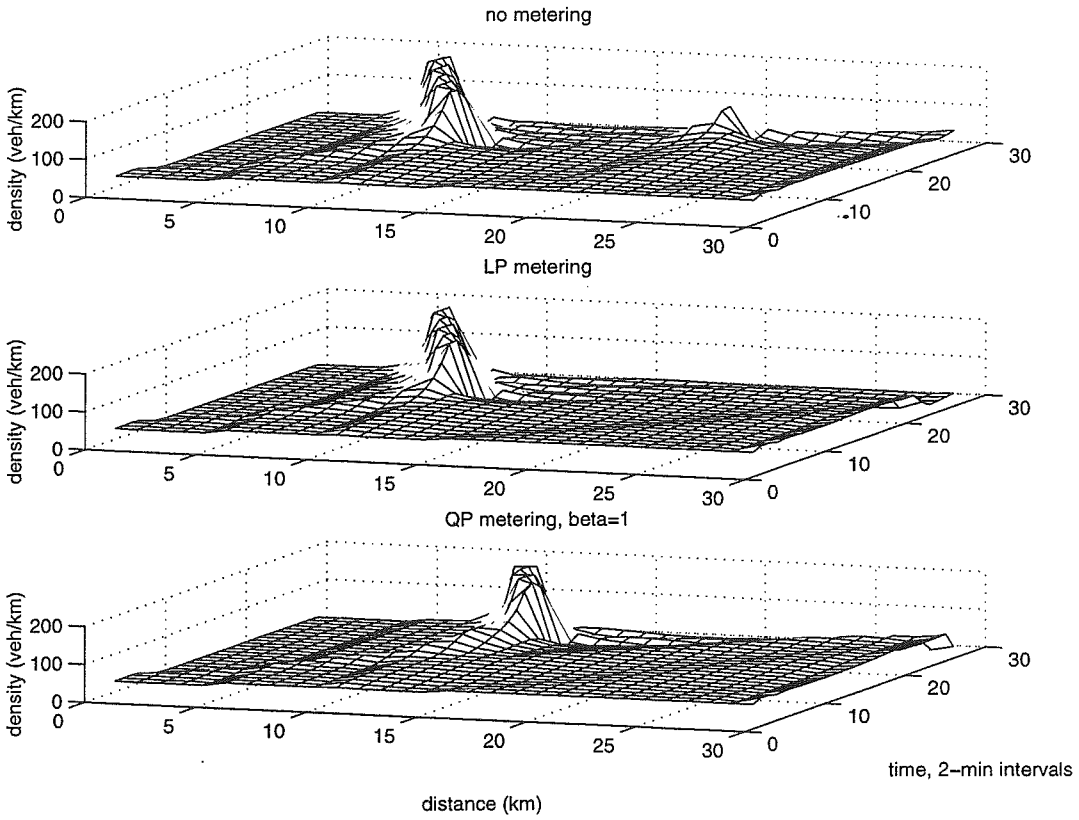


Figure 5- 3. Density evolution comparison

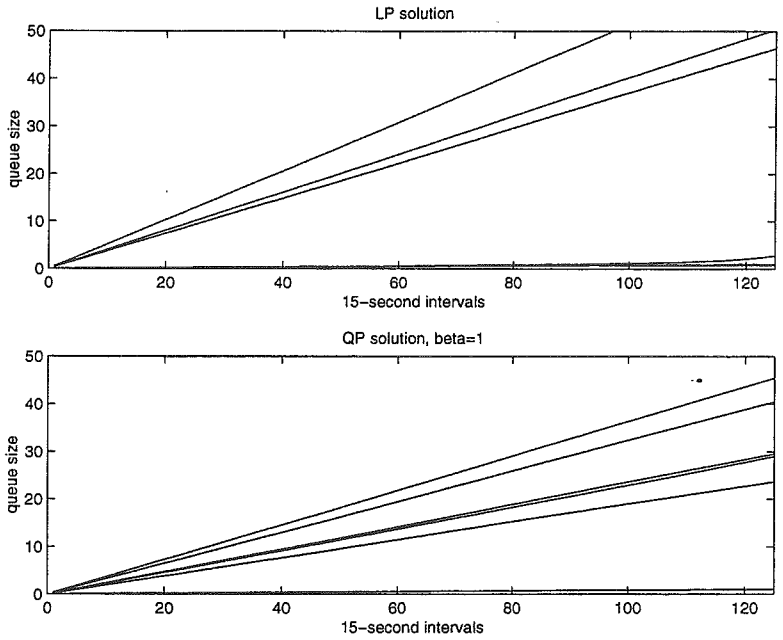


Figure 5- 4. Queue growth comparison

Method / Measure	Fwy travel time (veh-hr)	Queue time (veh-hr)	served load / offered load
No control	989	0.57	3048/5366 = 0.568
LP control	964	144	2976/5046 = 0.589
QP control	941	131	3077/5037 = 0.611

Table 5- 7. Preliminary method comparisons

Simulation test with extended queue dissipation

The previous comparison may be somewhat biased towards indicating the LP and QP methods are superior because the simulation ends with a significant number of vehicles still stored in ramp queues. Thus, another deterministic simulation was executed for a total time of two hours. The first 30 minutes was run with a set of low input volumes d_i, not requiring ramp metering, then 30 minutes of the high volumes from Table 5-8, and then an additional hour of simulation, at lower volumes as indicated in Table 5-8.

Table 5-9 indicates the initial conditions used in this experiment. This simulation would allow the queues that were built at the ramps by using the LP and QP metering methods to dissipate and result in a more *just* comparison of the LP, QP and no control situations. The hypothesis being that the discharged queues could possibly create a secondary congestion when released simultaneously (by any method), where in the no-control case, this

secondary congestion could not be built up because no such queues were created in minutes 31-60 of the experiment.

time period / input stream	external	ramp 1	ramp 2	Ramp 3	ramp 4	ramp 5
0 - 30 minutes	2500	484	410	255	255	242
31 - 60 minutes	3000	684	610	355	355	342
61 - 120 minutes	2500	484	410	255	255	242

Table 5- 8. Demand volumes for evaluation example 2

state variable / section	section 1	section 2	section 3	section 4	section 5
density (veh/km)	18	21	26	26	26
speed (km/hr)	101	101	95	95	95
speed limit (km/hr)	123	111	98.4	98.4	98.4

Table 5- 9. Initial conditions for evaluation example 2

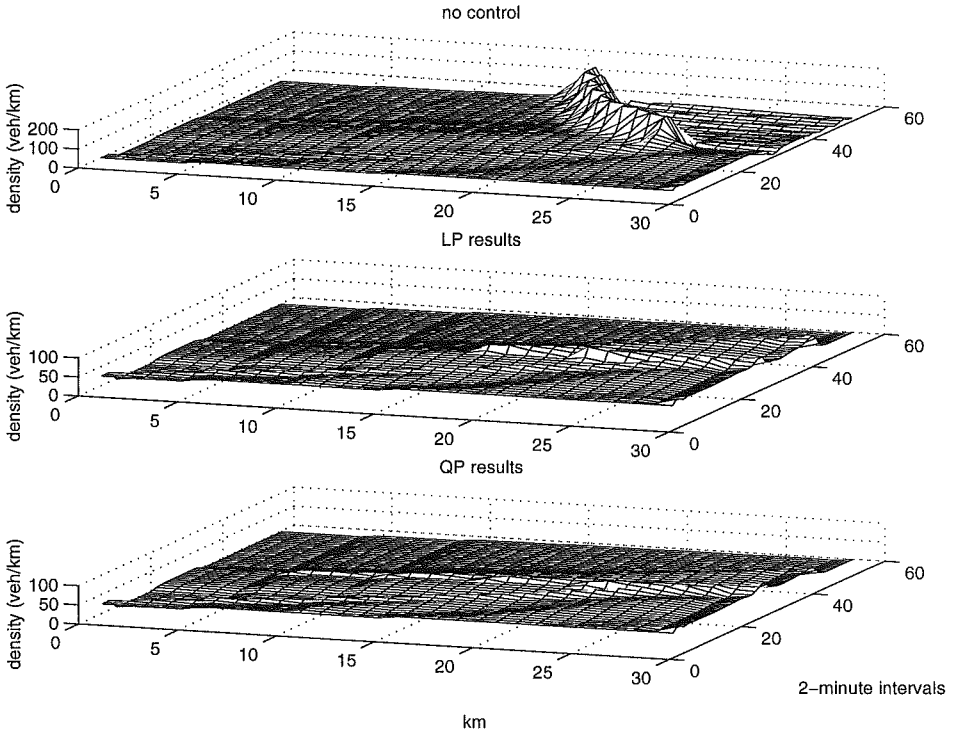


Figure 5- 5. Density evolution comparisons for evaluation example 2

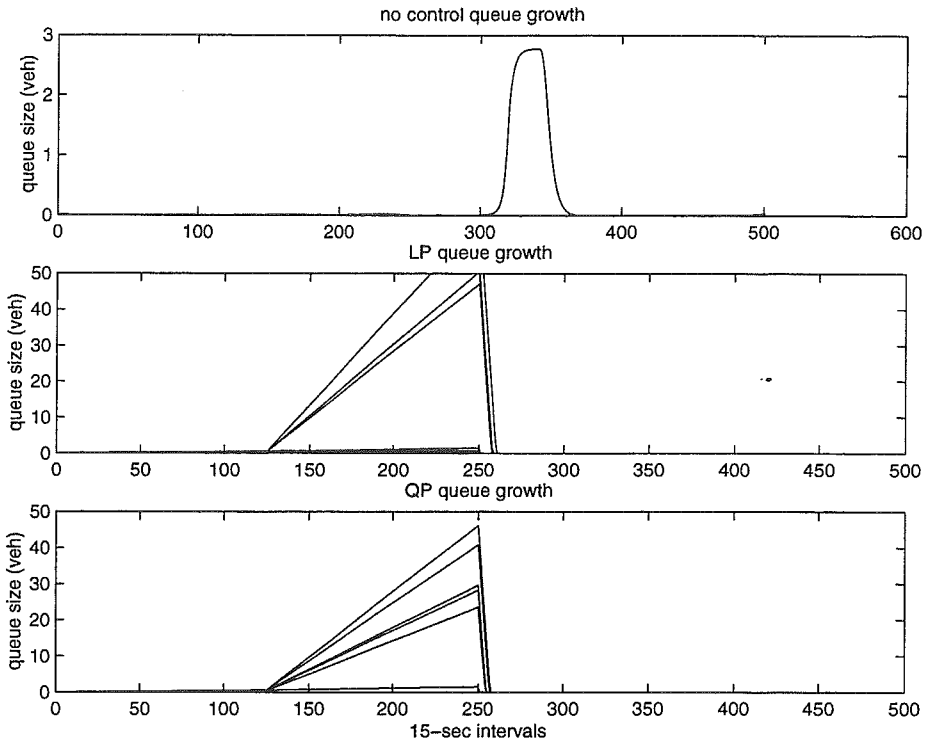


Figure 5- 6. Queue growth comparisons, evaluation example 2

Method / Measure	Fwy travel time	Queue time	offered load/served load
No control	2357	0.35	4847/9000 = 0.539
LP control	2262	44.9	5559/8917 = 0.623
QP control	2249	45.2	5559/8917 = 0.623

Table 5- 10. Performance comparisons, evaluation example 2

In this evaluation, the LP and QP methods have identical throughput, both 9% higher than the no-control case as listed in Table 5-10. The QP method results in slightly less freeway travel time than the LP method. With constant, deterministic arrival rates, this difference is considered statistically insignificant. As indicated in Figure 5-5, the backward-traveling congestion wave of the no-control situation is eliminated using both the LP and QP methods. Notice also in Figure 5-6, as in the previous example problem, that a queue of two vehicles is created and dissipates at the second on-ramp as the backward-traveling congestion wave passes the ramp area in the "no-control" case. This indicates the effect of adding the soft-limiter to queue-length computation in (4-12).

Summary

Overall, this simulation test showed that over a single time-horizon (i.e. without a rolling-horizon implementation) *queue-growth management* is the main benefit provided by the QP area-wide coordination algorithm over the LP approach. In an example problem, the performance of the freeway system under the two methods was virtually identical but two queue limitations were exceeded using when LP was used. The resulting benefits of using the QP approach *off of* the freeway must be inferred since we do not simulate the operation of the interchanges. We conjecture, however, that since the LP method spills-back more queues than the QP method, it is unlikely that the performance of the LP solution at the interchanges is superior to that of the QP result. In addition, as the congestion levels of the interchanges vary, the QP area-wide coordination optimization method reacts to these changes and should show even more powerful benefits. More extensive simulation testing, using stochastic input flows and the full rolling-horizon, two-layer optimization structure of MILOS is detailed in Chapter 9. Next, in Chapter 6, a predictive-cooperative real-time control layer will be described that realizes additional performance benefits (i.e. travel-time savings and queue dissipation) by modifying the nominal ramp metering rate according to the observed stochastic disturbances in the freeway and ramp demand flows.

Chapter 6: Predictive-cooperative real-time rate regulation algorithm

Introduction

The concept of a multi-layer controller structure and use of a regulation algorithm to further refine *nominal* ramp metering rates defined by a higher-level processor is not a new one [Payne and Isaksen, 1973; Papageorgiou, 1983; Payne et al., 1985]. The approach taken in this research is distinguished by several characteristics

(1) in contrast to the *reactive* linear-quadratic-regulator (LQR) approach based on automatic control theory, we develop a predictive, anticipatory approach that is proactive to *future* traffic states,

(2) queue management is *explicitly* considered in the real-time optimization formulation,

(3) a method is presented to *continue* real-time optimization even in the case where one or more segments of the freeway are congested, and

(4) information about the *dual* of the upper-layer area-wide coordination problem is used to guide real-time subproblem optimization.

Virtually no traffic-reactive ramp metering methods have been developed which *explicitly* consider queue management in the optimization procedure. There have been some preliminary attempts to manage queues at ramps, such as by synchronizing ramp metering rates with surface-street signal phases [Han and Reiss, 1995]. Another report lists "strategies" and "tactics" for integrated freeway/surface-street traffic management [Pooran et al., 1992].

Queue management has been historically addressed by traffic engineers using a *spillback* detector just before the surface-street interchange which create oscillations between high $r_{i,MAX}$ and low rates r_i with their switching behavior [ADOT, 1997]. Recently, a queue-control method has been presented that reduces the oscillating behavior by estimating occupancy on shorter time-scales using a first-order filter [Gordon, 1996]. This algorithm, however, does *not* derive/suggest the "higher" rate to change to when the queue length is to be reduced.

Adding queue management to state feedback control methods

To date, feedback control methods offer the most theoretically-founded solution to the problem of reacting to the inevitable stochastic irregularities in freeway traffic flow. These algorithms compensate for the stochastic disturbances and drive the measured state ρ_j towards a desired "nominal" state $\rho_{j,N}$. This nominal state is provided by an external system, algorithm, or decision-maker [Papageorgiou, 1983]. However, there are several drawbacks of applying linear feedback control methods.

First, when the disturbance is severe, linear feedback control methods may not drive the system back to the nominal state $\rho_j(k) \rightarrow \rho_{j,N}$ because the linear approximation to the system dynamics is no longer valid [Papageorgiou, 1983; Payne et al., 1985]. This can be addressed by turning "off" the local feedback controller in the presence of congestion and re-solving the upper-layer area-wide coordination optimization problem with a reduced-flow restriction in the congested section(s). When the system has returned to normal operation, the local feedback controller can then be turned back "on". We will show later that it is possible to continue locally traffic-responsive rate regulation even in the presence of congestion by using the predictive-cooperative real-time rate regulation algorithm developed in this chapter. The second main drawback of linear feedback control methods is the difficulty of such methods to consider queue management *explicitly* in the optimization procedure. These methods do not appropriately model the costs of changes to the queue and the matrix ricatti equation cannot be solved since in the system model the state-variable coefficient matrix $\mathbf{A}_c$ is singular.

Central concept of PC-RT rate regulation algorithm

The method developed in this chapter for queue management is termed the predictive-cooperative real-time rate regulation algorithm, hereafter referred to as PC-RT. The PC-RT rate regulation algorithm addresses the need to integrate the control effort of freeway control system with the concerns of the surface-street control system by:

(a) responding to statistically-significant short-term fluctuations in the stochastic demand flows to the ramp system (i.e. *both* the upstream freeway flow and the ramp demand),

(b) satisfying the coordination requirements of the area-wide problem solved at the upper layer,

(c) scheduling rates that are *pro-active* to possible future ramp and freeway demands, and

(d) reducing the possibility of and managing queue spillback at ramp entrances.

Point (a) reflects our assumption that, at the area-wide coordination layer, ramp demands and upstream freeway flows are considered constant over a short (i.e. 15-minute) time horizon. However, during that time horizon, it is well known that demand is *not* constant due to stochastic fluctuations. As such, any constant metering rate $r_i(k) = r_{i,N}\ \forall k$ will neglect these fluctuations and create queues when not necessary and/or release vehicles when disadvantageous to freeway conditions. It is important, however, at the real-time layer to distinguish between *negligible* statistical variation in the traffic stream and *significant* deviations in the flow. An algorithm for identifying these deviations based on statistical process control is discussed in Chapter 7.

Point (b) indicates that the PC-RT rate regulation algorithm should continue to satisfy the *area-wide* metering objectives by applying rates that do not deviate significantly from the nominal rates recommended by the upper-layer coordination optimization problem that is, $r_i(k) \in \left[r_{i,N} \pm \Delta r_{i,N}\right]$. One of the issues to be evaluated in this research is then, obviously, can a ramp metering control system be both coordinated on an area-wide basis, yet remain locally traffic-responsive?

To provide a preliminary response to this issue, consider the analogy of the *cycle, split, offset* concept from surface street control. In this concept, network coordination is maintained by using pre-timed signal controllers that operate on the same background cycle length and setting the splits and offsets to provide progression opportunities along arterials [McShane and Roess, 1990]. With the same cycle, split, offset paradigm, *semi-actuated* signal controllers can be used in the coordinated surface-street control system to provide some reactivity to local traffic conditions at an intersection while at the same time maintaining the progression opportunities of the coordinated system.

Finally, points (c) and (d) describe the *predictive* and *cooperative* components of the PC-RT rate regulation algorithm, respectively. The PC-RT rate regulation algorithm is designated as *predictive* to indicate that the control algorithm evaluates several *scenarios* for a short (i.e. five to seven minute) time horizon into the future of what *could* happen to the upstream freeway flow $V_0(k+1)$, $V_0(k+2)$,..., $V_0(k+N)$ and ramp demands $d_i(k+1)$, $d_i(k+2)$,..., $d_i(k+N)$. The PC-RT rate regulation algorithm is designated as *cooperative* to indicate the focus of the algorithm on *cooperating* with the surface-street signal controller to manage queue spillback by adapting the metering rate to react to the next few minutes of predicted flows from the interchange. Hence, the *cooperative* queue management method proposed in the PC-RT rate regulation algorithm does *not* default to the oscillating behavior resulting from using an occupancy threshold at a static "queue detector" to raise the metering rate to some pre-set "high" rate when the queue is sufficiently long.

Anticipated effects of PC-RT rate regulation algorithm

At the very least, the intended effect of implementing the PC-RT rate regulation algorithm with the solution from the QP area-wide coordination optimization problem should make the overall system (i.e. freeway and ramp queue) performance that is:

(a) *equivalent to* or *better than* the area-wide QP solution alone,

(b) *equivalent to* or *better than* feedback control methods [Papageorgiou, 1984, 1991; Payne et al., 1985] which do *not* consider queue management in their optimization procedures, and

(c) *equivalent to* or *better than* volume-occupancy traffic-responsive metering algorithms.

Thus, the intent of this research is to show that queue management can be considered in deriving metering rates that result in similar, if not better, freeway and surface-street performance than methods that do not consider queue management explicitly.

Basic function of the PC-RT rate regulation algorithm

The basic function of the PC-RT rate regulation algorithm is to exploit, at any time k, the excess local capacity $\rho_j(k) < \rho_{j,N}$ and $q_i(k) < q_{i,N}(k)$ in the freeway/ramp system by reacting in the following ways to the *fundamental* combinations of predicted ramp demand and predicted upstream freeway flow:

(1) *increase* the metering rate when the freeway density is *lower* than the nominal density and the ramp demand is *higher* than nominal,

(2) *decrease* the rate when the ramp demand is *lower* than nominal and freeway density is *higher* than nominal

(3) *increase* the rate when ramp demand is *lower* than nominal and freeway density is *lower* than nominal

(4) *increase* or *decrease* the metering rate according to a *trade-off* solution when ramp demand is *higher* than nominal and freeway density is *higher* than nominal.

How much to decrease or increase the rate $r_i(k)$ from the nominal setting $r_{i,N}$ is specified by formulation of a linear programming optimization problem (LP). This LP is formulated with a linearized description of the macroscopic freeway flow equations (from Chapter 4) about the nominal *equilibrium state* ($\rho_{j,N}$, $v_{j,N}$ $r_{i,N}$) and a linear description of queue growth about the nominal queue-growth *trajectory* $q_{i,N}(k)$. The cost function of this LP optimization problem is a weighted sum of travel-time savings in each section of the freeway and on the ramp approaches. The weights of each state-variable are derived from the dual multipliers λ_k and constraint slack ε_k values of the solution to the upper-layer area-wide QP optimization problem. In this manner, a trade-off between travel-time savings on the ramp and on the freeway is based on the *current* interchange conditions (i.e. how important it is to manage spillback at *this* ramp) and the conditions in critical freeway sections.

The PC-RT rate regulation algorithm can be described as a three-step process:

(1) Given that a *significant* deviation from the upstream freeway or ramp demand nominal flow is detected, *predict* several possible subsequent flows to the ramp and the upstream freeway segment,

(2) Given these predicted possible future *scenarios*, solve an LP optimization problem for each predicted scenario that reduces queuing time on the ramp and/or reduces the possibility for congestion on the freeway over the next few minutes, and

(3) In the next optimization interval, collect the "actual" upstream freeway flow and ramp demand, compare the actual flow to the predicted scenarios, and apply the appropriate metering rate for the scenario that best matches the actual flow.

A *rolling-horizon* framework is used in the three-step process listed above. Thus, the PC-RT optimization problems are solved for a 5 to 7 minute predictive time-horizon, but the metering rate is only applied for the first 1-2 minutes of the time horizon before the problem is possibly re-evaluated due to the stochastic fluctuations.

Reasonable and important assumptions

Some reasonable and important assumptions are needed to successfully formulate and apply meaningful optimization results at the predictive-cooperative, real-time rate regulation layer:

(1) The traffic flows to ramps and on the freeway have a stochastic component that can be *identified* and *separated* from structural changes in the underlying process. For the case of ramp demands, it is assumed that this stochastic effect can be separated from the flow-rate changes induced by the traffic signal at the interchange.

(2) At least one-minute upstream measurements of flows to the ramp meter are available from the surface street detector system and are reliable enough for use in the PC-RT rate regulation and area-wide coordination optimization problems.

(3) The linear approximation of the nonlinear macroscopic flow equations (Chapter 4) can be used to accurately predict freeway flow dynamics over 1-7

minute time scales. Correspondingly, route-proportional flow rates and other parameters of the linearized model are assumed to remain constant over the real-time optimization time horizon.

(4) The flows and metering rates solved for at the area-wide coordination layer of approximation are realizable as an equilibrium-state, in the absence of stochastic variation.

(5) The flows resulting from solution of the area-wide coordination problem are *desirable* settings from a system-wide perspective. That is to say that the solution of the area-wide problem provides benefits that *individual* solution of the local control problems, with independently-derived nominal settings (e.g. critical flows that do not vary by time-of-day), could not provide. Thus, the goal of keeping the real-time rates "close to" the area-wide nominal settings is a sound objective.

Linearization about an equilibrium state

Having an equilibrium solution and a set of desired nominal rates that address network-wide concerns from the area-wide coordination problem, we can begin to simplify the difficult optimal control problem to one that can be solved in real-time. First, as developed in previous work [Papageorgiou, 1983; Payne et al., 1985], consider the differential equation model for the freeway density dynamics $\dot{\rho} = f(\rho, v, r, d, t)$. ρ is considered the *primary* dynamic variable in freeway flow modeling since, as shown by empirical data, speeds are approximately constant until the density approaches the critical value ρ_c. As such, the dynamic equation for the speed v is replaced by the equilibrium speed-density function $v_e(\rho)$ in the freeway flow dynamics for the development of the PC-RT optimization subproblems.

Hence, we approximate the nonlinear function $f(\rho, v, r, d, t)$ by a linearization (first-order Taylor-series expansion) about the equilibrium point (ρ_N, v_N, r_N)

$$\hat{\dot{\rho}}_j = f_j(\rho_N, r_N) + \sum_{j=1}^{J} \left.\frac{\partial f_i(\rho,r)}{\partial \rho_i}\right|_{\rho_N, r_N} (\rho_j - \rho_{j,N}) + \sum_{j=1}^{J} \left.\frac{\partial f_j(\rho,r)}{\partial r_j}\right|_{\rho_N, r_N} (r_j - r_{j,N}) \quad \textbf{Eqn. 6- 1}$$

where J indicates the number of segments in the freeway model. Each segment of the freeway may not contain an off-ramp, hence many $r_j = 0$. We include those terms for completeness only. A linear system description results for the deviation from the nominal state

$$\Delta\dot{\rho}_j = \mathbf{A}^* \Delta\rho_j + \mathbf{B}^* \Delta r_j \quad \textbf{Eqn. 6- 2}$$

such that

$$\begin{aligned} \Delta\rho_j &= \rho_j - \rho_N \\ \Delta r_j &= r_j - r_N \\ \Delta\dot{\rho}_j &= \dot{\rho}_j - \dot{\rho}_N \\ \dot{\rho}_N &= f(\rho_N, r_N) \end{aligned} \quad \textbf{Eqn. 6- 3}$$

and

$$\mathbf{A}^* = \begin{bmatrix} \left.\frac{\partial f_1(\rho,r)}{\partial \rho_1}\right|_{\rho_N, r_N} & \cdots & \left.\frac{\partial f_1(\rho,r)}{\partial \rho_J}\right|_{\rho_N, r_N} \\ \cdots & \cdots & \cdots \\ \left.\frac{\partial f_J(\rho,r)}{\partial \rho_1}\right|_{\rho_N, r_N} & \cdots & \left.\frac{\partial f_J(\rho,r)}{\partial \rho_J}\right|_{\rho_N, r_N} \end{bmatrix}, \quad \mathbf{B}^* = \begin{bmatrix} \left.\frac{\partial f_1(\rho,r)}{\partial r_1}\right|_{\rho_N, r_N} & \cdots & \left.\frac{\partial f_1(\rho,r)}{\partial r_J}\right|_{\rho_N, r_N} \\ \cdots & \cdots & \cdots \\ \left.\frac{\partial f_J(\rho,r)}{\partial r_1}\right|_{\rho_N, r_N} & \cdots & \left.\frac{\partial f_J(\rho,r)}{\partial r_J}\right|_{\rho_N, r_N} \end{bmatrix} \quad \textbf{Eqn. 6- 4.}$$

with $f_j(\rho, \upsilon, r)$ defined as

$$f_j(\rho, \upsilon, r) = V_{j,IN} - V_{j,OUT} + r_j - s_j \quad \textbf{Eqn. 6- 5}$$

and

$$\begin{aligned} V_{j,IN} &= \alpha \upsilon_{j-1} \rho_{j-1} + (1-\alpha) \upsilon_j \rho_j \\ V_{j,OUT} &= \alpha \upsilon_j \rho_j + (1-\alpha) \upsilon_{j+1} \rho_{j+1} \\ s_j &= \theta_j \upsilon_j \rho_j \end{aligned} \quad \textbf{Eqn. 6- 6}$$

as developed in Chapter 4. Collecting like terms in (6-5) after substituting the relationships of (6-6) we get

$$f_j(\rho,v,r) = \alpha v_{j-1}\rho_{j-1} + \{(1-2\alpha-\theta)\}v_j\rho_j - (1-\alpha)v_{j+1}\rho_{j+1} + r_j \qquad \textbf{Eqn. 6- 7}$$

This formulation uses the *turning-percentage* representation of the off-ramp rate, indicating that, as the density and speed in the section change, so does the number of vehicles exiting the section. An alternative to this proportional off-ramp rate model is to use the (fixed) off-ramp rate specified by the nominal solution of the upper-layer area-wide coordination problem.

Elimination of the dynamic speed equation

The complex nonlinear dynamic equation (4-3) for v_j is replaced by solving for the equilibrium speed $v_{j,N}$ from the nominal density $\rho_{j,N}$ from (4-5)

$$v_{j,N} = v_e(\rho_{j,N}) = v_f\left(1-\left(\frac{\rho_{j,N}}{\rho_{j,MAX}}\right)^l\right)^m \qquad \textbf{Eqn. 6- 8}$$

and thus the matrices A^* and B^* can hence be written as

$$\mathbf{A}^* = \begin{bmatrix} a_{1,1} & a_{21} & 0 & \ldots & 0 \\ a_{1,2} & a_{2,2} & a_{3,2} & & \ldots \\ 0 & & \ldots & & 0 \\ \ldots & & a_{M-2,M-1} & a_{M-1,M-1} & a_{M,M-1} \\ 0 & \ldots & 0 & a_{M-1,M} & a_{M,M} \end{bmatrix} \qquad \textbf{Eqn. 6- 9}$$

$$\mathbf{B}^* = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ldots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \qquad \textbf{Eqn. 6- 10}$$

where elements of A^* are defined as

$$a_{j,j} = \left.\frac{\partial f_j(\rho,r)}{\partial \rho_j}\right|_{\rho_N, r_N} = \{(1-2\alpha-\theta)\}\left[v_e(\rho_{j,N}) + v_e'(\rho_{j,N})\rho_{j,N}\right]$$

$$a_{j,j-1} = \left.\frac{\partial f_j(\rho,r)}{\partial \rho_{j-1}}\right|_{\rho_N, r_N} = \alpha\left[v_e(\rho_{j-1,N}) + v_e'(\rho_{j-1,N})\rho_{j-1,N}\right] \qquad \textbf{Eqn. 6- 11}$$

$$a_{j,j+1} = \left.\frac{\partial f_j(\rho,r)}{\partial \rho_{j+1}}\right|_{\rho_N, r_N} = -(1-\alpha)\left[v_e(\rho_{j+1,N}) + v_e'(\rho_{j+1,N})\rho_{j+1,N}\right]$$

where $v_e'(\rho_{j,N})$ is the derivative of $v_e(\rho_{j,N})$ at the nominal density $\rho_{j,N}$. Recall that α is calibrated to the freeway location and θ is the percentage of volume exiting at the off-ramp. B^* can be specified as a diagonal unity matrix because $r_j=0$ for sections that do not have an on-ramp.

The replacement of the nonlinear model with its linearization results in significant reduction of the model order (by elimination of the v state-variables for section speeds) without significant loss of descriptive accuracy around the nominal state point (ρ_N, v_N, r_N) as developed in previous research [Papageorgiou, 1983; Payne et al., 1985].

By using the (simple) definition of the derivative

$$\frac{\rho(t+\Delta t) - \rho(t)}{\Delta t} \qquad \textbf{Eqn. 6- 12}$$

the continuous nonlinear dynamic system around the equilibrium state is converted to the difference equation

$$\Delta\rho(t+\Delta t) = \Delta\rho(t) + \Delta t\left(\mathbf{A}^*\Delta\rho + \mathbf{B}^*\Delta r_{on}\right) \qquad \textbf{Eqn. 6- 13}$$

for simulation on a digital computer and formulation of the PC-RT rate regulation problems as an LP.

The range of descriptive accuracy of the linear approximation to the nonlinear model has not been studied empirically, but analytical studies have indicated that the equilibrium point (ρ_N, v_N, r_N) is *not* globally asymptotically stable [Papageorgiou, 1983; Zhang and

Ritchie, 1996]. That is, a significant disturbance can drive the system away from the equilibrium point (ρ_N, v_N, r_N) and into a congested "equilibrium" state. This limitation of the region of attraction of the equilibrium point (ρ_N, v_N, r_N) is used as our first set of constraints for the PC-RT optimization problem. Thus, the state variables $\Delta\rho_j$ must remain within the limits of modeling accuracy and applicability such that

$$\max[\rho_{j,MIN}, -\rho_{j,N}] \le \Delta\rho_j \le \min[\rho_{crit} - \rho_{j,N}, \rho_{j,MAX}]. \qquad \textbf{Eqn. 6- 14}$$

The lower bound of constraint 6-14 results from the fact that the density must be non-negative, but the model may not be applicable for densities anywhere near zero (when $\rho_{j,MIN} >> 0$). The upper bound of (6-14) results from the system description being invalid for density values greater than the critical value ρ_{crit}, but may still only be applicable up to some point less than the critical point $\rho_{j,N} + \rho_{j,MAX} << \rho_{crit}$. In this research, however, we use bounds determined by engineering judgment. Identification of analytical form for these modeling limitations may be a subject of future research.

Similarly, from a modeling perspective the metering rates Δr_j are restricted to the set

$$r_{j,MIN} - r_{j,N} \le \Delta r_j \le \hat{r}_{j,MAX} - r_{j,N} \qquad \textbf{Eqn. 6- 15}$$

such that $r_{j,MIN}$ and $r_{j,MAX}$ may be set to some value higher than (lower than) the absolute lowest (highest) metering rate to maintain the model's accuracy to real system behavior. The constraints 6-15 are also needed to satisfy the coordination requirements of the upper-layer QP problem. Hence, the real-time rates r_j should remain close to the nominal rate. Currently, the $r_{j,MIN}$ and $r_{j,MAX}$ bounds are set proportional to the nominal metering rate, such that

$$\begin{aligned} r_{j,MIN} &= \omega r_{j,N} \\ r_{j,MAX} &= \frac{r_{j,N}}{\omega} \end{aligned} \qquad \textbf{Eqn. 6- 16}$$

where $\omega \in (0,1)$ is a parameter to be chosen by the system operator or based on the attraction region of the equilibrium point (ρ_N, v_N, r_N).

Structure of the PC-RT objective function

In the PC-RT rate regulation problem, we are interested in adding the objective of managing queue lengths (and minimizing them when possible) to the standard objective of maintaining smooth freeway flow and minimizing total travel time. We use a *linear* objective of the total "additional" travel time in the system as the overall objective of the PC-RT rate regulation problem, since our variables are now defined as the *deviation* from the nominal state.

This implies that, with the state at the nominal point (ρ_N, υ_N, r_N), a certain total travel time $\sum_{k=1}^{K}\left[\sum_{j=1}^{N}\Gamma_j\rho_{j,N}(k)+\sum_{i=1}^{M}q_{i,N}(k)\right]$, where Γ_j is the length of segment j and the real ramp locations are indexed by i, would be realized for the system over the time horizon. However, because of stochastic variation in the ramp demands $d_i(k)$ and upstream freeway demand $V_{0,IN}(k)$, the travel time will be higher or lower than the expected travel time according to this stochasticity. Our objective in the PC-RT rate regulation problem is to minimize a weighted combination of any *additional* travel time incurred by stochastic variation, and, at the same time, exploit the stochastic fluctuations to our advantage by dissipating queues at appropriate times. Hence, we define the objective of the PC-RT problem as

$$\min\sum_{k=1}^{K}\left[\sum_{j=1}^{J}c_j\Gamma_j\Delta\rho_j(k)+\sum_{i=1}^{M}c_{i,q}\Delta q_i(k)\right] \qquad \textbf{Eqn. 6- 17}$$

where c_j is a weighting cost coefficient of freeway section j, $\Delta\rho_j(k)$ is the deviation from the nominal point $\rho_{j,N}$ of freeway section j at time k, $c_{i,q}$ is a weighting coefficient of queue i, and $\Delta q_i(k)$ is the deviation from the nominal queue length $q_{i,N}(k)$ of the queue at ramp i at time k. Now we denote the total number of ramps in the problem as M. Derivation of the cost coefficients c_j and $c_{i,q}$ will be detailed in a later section.

The first component of the cost function (6-17) reflects the desire to keep the freeway from being forced into congestion by the stochastic variation in the flow rate. Hence, at the expense of throughput (i.e. higher densities), we would like to keep the density on the freeway at or below the nominal point $\rho_{j,N}$ (i.e.$\Delta\rho_j(k) < 0$) specified by the upper-layer area-wide coordination problem. This reduction in density is achieved by holding (even more) vehicles back on the ramp, which is in direct conflict with the concerns of the surface-street control system to keep the ramp queue from spilling back into the adjacent interchange.

Queue growth modeling

The second part of the PC-RT objective function (6-17) reflects the queue-management concerns of the surface-street system where $c_{i,q}$ is the cost coefficient reflecting the *importance* of deviations of the queue length $q_i(k)$ at queue i from the *nominal* queue $q_{i,N}(k)$ length at time k. We now define the concept of a "nominal queue". Consider that, at the area-wide optimization layer of control, we specify ramp metering rates $r_{i,N}$ that are based on a constant arrival rate d_i at each ramp. Applying these constant rates $r_{i,N}$ with constant demands d_i results in a (known) queue length $q_i(k)$ when the initial queue length $q_i(0)$ is known.

Hence, $q_{i,N}(k)$ could be considered the "acceptable" queue as a limit of how long we would be willing to allow the queue to grow to at each time instant k over the time horizon K. Thus, the deterministic queue growth function $q_{i,N}(k)$ becomes the nominal *trajectory* of the new state variable

$$\Delta q_i(k) = q_i(k) - q_{i,N}(k) \qquad \textbf{Eqn. 6- 18}$$

with the evolution equation

$$\Delta q_i(k+1) = \Delta q_i(k) + \Delta T\left[\Delta d_i(k) - \Delta r_i(k)\right] \qquad \textbf{Eqn. 6- 19}$$

such that $\Delta d_i(k)$ is defined as the deviation of the current ramp demand from the constant demand assumed at the area-wide optimization layer

$$\Delta d_i(k) = d_i(k) - d_{i,N} \qquad \textbf{Eqn. 6- 20}$$

and $\Delta r_i(k)$ is defined as the deviation of the current metering rate $r_i(k)$ from the constant rate $r_{i,N}$ specified from the area-wide optimization problem such that

$$\Delta r_i(k) = r_i(k) - r_{i,N}. \qquad \textbf{Eqn. 6- 21}$$

Coupling the definition (6-18) of the new variable $\Delta q_i(k)$ with the objective of minimizing this variable, it should be apparent that the PC-RT objective (6-17) reflects minimizing the *additional* travel time in the system. Hence, the goal of the PC-RT controller is to *avoid* additional congestion in the freeway system while taking advantage of *opportunities* to release vehicles from ramp queues. This situation in illustrated in Figure 6-1.

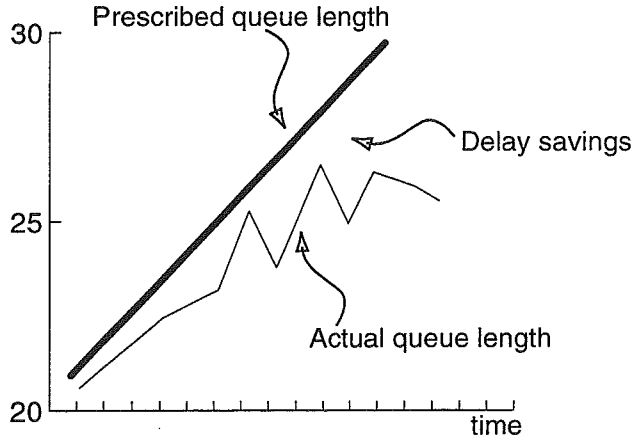


Figure 6- 1. Prescribed maximum queue growth rate

As shown in Figure 6-1, we allow the queue to grow from $q_i(0)$, at most, on average, as fast as the deterministic rate $d_{i,N} - r_{i,N}$ specified in the solution to the area-wide coordination problem. This is illustrated in Figure 6-1 as the thick dark line. Thus, at the end of the short-term time-horizon of the PC-RT optimization, the queue length should be no greater than the length specified by applying the constant rate $r_{i,N}$. This does not mean that faster rates of queue growth are not allowed, but, if faster rates of queue growth should occur, a corresponding increase in the metering rate must be enacted to return the average queue length at or below the length specified by $q_i(k+1) = q_i(k) + \Delta T\left(d_{i,N} - r_{i,N}\right)$. Thus, we have the constraints

$$\Delta q_{i,MIN}(k) \le \Delta q_i(k) \le \Delta q_{i,MAX}(k) \quad \forall\, i,k \qquad \textbf{Eqn. 6- 22}$$

for the PC-RT problem, where $\Delta q_{i,MIN}(k)$ is defined as

$$\Delta q_{i,MIN}(k) = -q_{i,N}(k) \quad \forall\, i,k \qquad \textbf{Eqn. 6- 23}$$

indicating that the queue cannot become negative, and $\Delta q_{i,MAX}(k)$) is defined as

$$\begin{aligned} &\Delta q_{i,MAX}(k) = q_{i,MAX} - q_{i,N}(k) \quad \forall\, i,k \\ &q_{i,MAX} = q_{i,storage} + z_i \end{aligned} \qquad \textbf{Eqn. 6-24}$$

where z_i is the value of the "overflow" variable for ramp i from the solution of the area-wide coordination problem. Notice that $q_{i,MIN}(k)$ and $q_{i,MAX}(k)$ are functions of time according to the nominal growth rate $q_{i,N}(k)$.

Control variable modeling

Similar to the definition of the $\Delta q_i(k)$ variables, we define the control variable $\Delta r_i(k)$ as in(6-21). For realistic representation of the problem, we cannot expect to be able to make changes to the metering rate $r_i(k)$ on time scales of 5-15 seconds, as frequently as the macroscopic model equations are re-evaluated, Thus the constraints

$$\begin{aligned} &\Delta r_i(k) = \Delta r_i(k+1) = \cdots \Delta r_i(k+t) \qquad \forall\, i \\ &\Delta r_i(k+t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+2t) \\ &\vdots \\ &\Delta r_i(k+(Z-1)t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+Zt) \end{aligned} \qquad \textbf{Eqn. 6- 25}$$

are added to the formulation to specify that the ramp metering rate $r_i(k)$ can only be changed once each t re-evaluation steps, where t is the number of re-evaluation steps per minute and Z is the number of minutes in the time horizon K.

Derivation of the PC-RT cost coefficients from the QP solution

The final aspect of the mathematical description of the PC-RT rate regulation optimization problem is the derivation of the weighting coefficients c_j and $c_{i,q}$ in the objective function (6-17). These weighting coefficients indicate the relative importance

of the state variables $\Delta q_i(k)$ and $\Delta \rho_j(k)$ and can be used to integrate the solution results of the area-wide coordination problem with the local PC-RT optimizations. To illustrate this, consider when the upper-layer area-wide coordination optimization problem is solved, a set of dual multipliers λ_ζ and slacks ε_ζ are obtained for the constraints $g_\zeta(r,q) \geq 0 \quad \forall \zeta$ in the problem. When a constraint is *tight*, (i.e. $g_\zeta(r,q) = 0$), the dual multiplier is nonzero $\lambda_\zeta \neq 0$ and the slack is zero $\varepsilon_\zeta = 0$. When a constraint has *slack*, (i.e. $g_\zeta(r,q) > 0$) then $\lambda_\zeta = 0$ and $\varepsilon_\zeta \neq 0$. The dual multiplier of a constraint indicates the *price* the decisionmaker/modeler would pay to obtain an additional unit of that resource. In other words, the dual multiplier indicates, in terms of the units of the objective cost function being optimized, how much cost would be incurred if the RHS of the constraint was increased by one unit. Therefore, when a constraint has slack, $\varepsilon_\zeta \neq 0$, the stakeholder should *not* be willing to "pay" (or incur cost) to obtain an additional unit of that resource.

Computational procedure to obtain cost coefficients

Consider that the area-wide QP is a maximization problem. If all constraints are converted to $g_\zeta(r,q) \leq 0$ constraints, then $\lambda_\zeta \geq 0 \quad \forall \zeta$. In the following derivation, we do not distinguish between c_j and $c_{i,q}$, terming each cost parameter generically c_ζ. Let λ_{min} by the smallest non-zero dual price such that

$$\lambda_{MIN} = \min_{\zeta} \left\{ \lambda_\zeta \mid \lambda_\zeta > 0 \right\} \qquad \textbf{Eqn. 6- 24}$$

and let λ_{MAX} be the largest non-zero dual price such that

$$\lambda_{MAX} = \max_{\zeta} \left\{ \lambda_\zeta \mid \lambda_\zeta > 0 \right\}. \qquad \textbf{Eqn. 6- 25}$$

Constraints with $\lambda_\zeta \neq 0$ are then assigned the cost coefficient

$$c_\zeta = \frac{\lambda_\zeta}{\lambda_{MAX}} + 1 \qquad \textbf{Eqn. 6- 26}$$

to scale the cost coefficient relative to the maximum cost. Constraints with zero dual prices $\lambda_\zeta = 0$ are assigned the cost coefficient

$$c_\zeta = \frac{\lambda_{MIN}}{\lambda_{MAX}}\left(\frac{cap_\zeta - \varepsilon_\zeta}{cap_\zeta}\right)+1 \qquad \textbf{Eqn. 6- 27}$$

where cap_ζ is the capacity (right-hand-side) of constraint ζ and ε_ζ is the slack of constraint ζ. In this manner, the constraints with zero dual prices are linearly scaled to the cost coefficient of the constraint with λ_{MIN}. In the event that *no* constraints have non-zero dual multipliers, $\left\{\lambda_\zeta \mid \lambda_\zeta > 0\right\} = \varnothing$, then the cost coefficients are assigned according to

$$c_\zeta = \left(\frac{cap_\zeta - \varepsilon_\zeta}{1-\varepsilon_{MIN}}\right)+1 \qquad \textbf{Eqn. 6- 28}$$

where ε_{MIN} is the lowest slack ratio (i.e. the closest constraint to being tight as a percentage of its capacity) such that

$$\varepsilon_{MIN} = \min_\zeta\left\{\frac{cap_\zeta - \varepsilon_\zeta}{cap_\zeta}\right\}. \qquad \textbf{Eqn. 6- 29}$$

This pricing method could be considered a "trickle-down" approach to reflect the congestion conditions at the interchange and the area-wide coordination priorities in each of the local PC-RT optimization problems. Thus, the combination of the dual multipliers λ_ζ and the slack values ε_ζ incorporates information from all of the following: cost coefficients of the area-wide QP, queue storage size at each ramp, demand to each ramp, and capacity of each freeway segment.

Modification to the linearization procedure for unstable conditions

When a section of the freeway is in the *unstable* regime of the characteristic equation $V_j = \rho_j \cdot v_j$ such that $\rho_j > \rho_{j,crit}$, the linearized system (6-13) does not effectively describe the system dynamics. Recall from Chapter 1, Figure 1-1, the shape of the volume-density

characteristic. Note that in Figure 1-1, the axes are reversed from the development to follow. When the density ρ_j in a section is overcritical, a backward-traveling (i.e. upstream-traveling) congestion wave is created with wave speed equal to the slope of the volume-density curve at the congested point [Newell, 1993]. Linearizing about the stable equilibrium point $\rho_{j,N}$ results in approximating the *upstream* traveling wave with a *downstream* traveling wave as shown in Figure 6-2. Here, the grey points indicate the current state, nominal state, and approximated state. Notice that the uppermost point is the estimate of the current state by linearizing the system about the nominal state and the predicted volume is significantly higher than the volume-density characteristic function, and unrealistic.

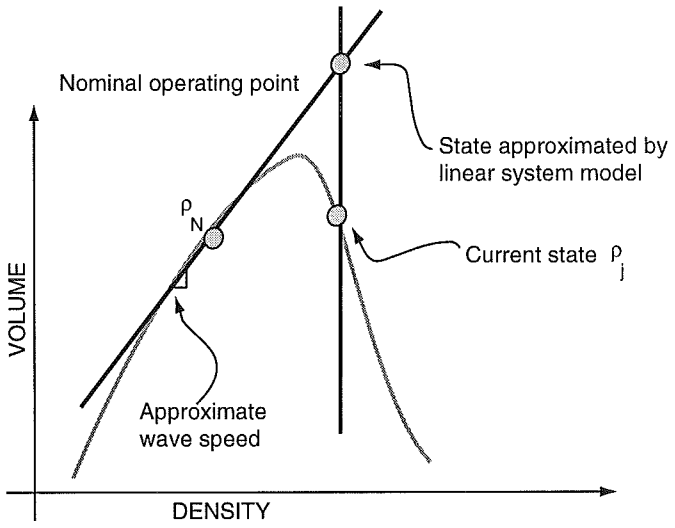


Figure 6- 2. Example of incorrect wave-speed model for congested section

One approach to remedy this modeling inaccuracy is to re-define the nominal point ρ_N in the congested section at the current *congested* density $\tilde{\rho}_N = \rho_j$ and solve the PC-RT control problem linearized about ρ_j. As shown in previous research, this model results in containment of the congestion to that segment when appropriately strong feedback control is applied [Payne et al., 1985]. However, the congestion cannot be eliminated if the cost function (i.e. LQR quadratic feedback rules) penalizes *both* positive and negative deviations from the nominal point. With such a cost function, the feedback rule increases

upstream metering rates $r_{i-t}(k) \geq r_{i-t,N} \mid t>0$ when the density decreases in the congested downstream segment $\rho_j(k) \leq \tilde{\rho}_{j,N}$ (and thus $V_j(k) > \tilde{V}_{j,N}$). Therefore the downstream segment is kept in the congested regime $\rho_j(k) \rightarrow \tilde{\rho}_{j,N} \mid \tilde{\rho}_{j,N} \geq \rho_{crit}$. Obviously this is not an acceptable solution to the congestion problem.

Congestion *can* be eliminated in the oversaturated segment by re-solving the upper-layer area-wide coordination problem with a reduced maximum flow rate $\tilde{V}_{j,N} < V_{j,MAX}$ in the congested section. Simply put, the congestion in that segment cannot be eliminated unless $V_{j-1}(k) < V_{j+1}(k)$ for the time period until $\rho_j(k) < \rho_{j,crit}$ in the congested section [Papageorgiou, 1983]. This results in nominal points for upstream segments that must be below the maximum volume level of the oversaturated segment, that is $\tilde{V}_{j-t,N} \leq \tilde{V}_{j,N} \mid t>0$, since in our "steady-state" assumption all vehicles that enter the system will exit the system during the time horizon T. Otherwise, if $\tilde{V}_{j-t}(k) \geq \tilde{V}_{j,N} \mid t>0$, vehicles will be stored in section j, and possibly sections $j-t \mid t>0$.

An example of this case is illustrated in Figure 6-3. Here, a small five-section freeway is illustrated with segments A-E. Segment E is congested, (i.e. $\rho_E(k) > \rho_{E,crit}$) at point E in Figure 6-3. Correspondingly, in the solution to the area-wide coordination problem, the flow rates at the uncongested (i.e. $\rho_j(k) \leq \rho_{j,crit}$) segments A, B, C, and D must be less than the flow rate at E, indicated by points A, B, C, and D in Figure 6-3.

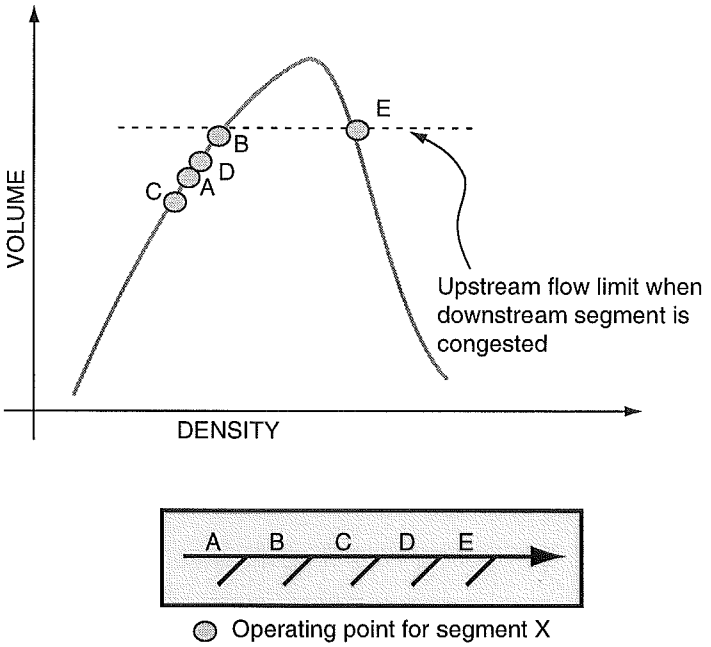


Figure 6- 3. Overcapacity segment results in upstream area-wide flow limitations

The condition illustrated in Figure 6-4 can result in congestion clearing by turning "off" any traffic-responsive control and using the nominal metering rates $r_{i,N}$ from the area-wide *capacity-limited* coordination problem. However, it seems apparent that some additional performance benefit may be realized by using a real-time problem that is appropriately structured in the congested section. Namely, an optimization problem formulation that does not result in forcing the state to the undesirable congested point.

Hence, to consider the application of the PC-RT rate regulation optimization problem to the congested section, recall that (6-16) is a *linear* cost function. Thus, the cost function already reflects the fact that $\rho_j(k+1) > \rho_{j,N}$ should be penalized when $\rho_j(k) > \rho_{j,crit}$ but $\rho_j(k+1) < \rho_{j,N}$ should not. Second, the *nominal* density $\rho_{j,N}$ in the over-capacity section is re-defined as the *critical* density $\rho_{j,crit}$ at which the maximum flow $V_{j,MAX}$ occurs. The system is linearized, however, at the currently congested point $\rho_j(k)$. The resulting approximation is illustrated in Figure 6-4.

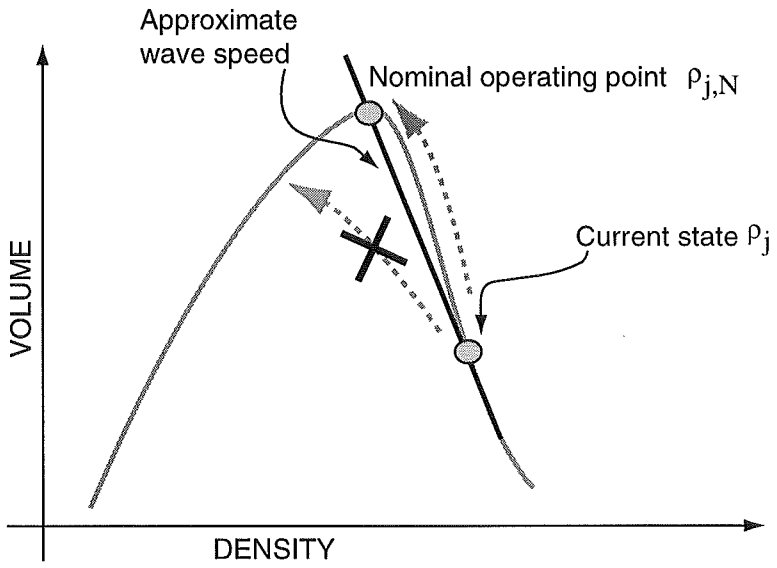


Figure 6- 4. Alternative model for the over-capacity situation

This model assumes that, in a dynamical sense, the combination of flow and density in a congested section must travel *continuously* along the fundamental curve from the congested point $\rho_j(k)$ to the maximum flow point $\rho_{j,crit} = \rho_{j,N}$, as shown in Figure 6-5, assuming infinitesimally small time intervals, Δt. The density does not jump from point to point discontinuously as uncongested flow resumes, as illustrated in Figure 6-4 by the "X-ed" arrow from the congested regime to the uncongested regime. The assumption of continuous flow change is further substantiated by Figures 1-2 and 1-3 from the Phoenix area freeway indicating that the transition from congested to uncongested flow is continuous during the rush-hour period.

Hence, for the specification of the PC-RT optimization problem in the congested section, constraints (6-14) must be modified in the case where $\rho_j > \rho_{j,crit}$, since we must periodically re-linearize the model in the congested segment at the current point as the state $\rho_j(k)$ approaches the critical value $\rho_{j,crit}$. The state constraints are thus given as

$$\max[\rho_{j,MIN},\ \rho_{j,crit} - \rho_j] \leq \Delta\rho_j(k) \leq \min[\rho_{j,jam} - \rho_j,\ \rho_{j,MAX}] \quad \forall j \qquad \textbf{Eqn. 6- 30}$$

replacing constraints (6-14) for the congested section only. The control constraints are unchanged from the previous definition of (6-15). This periodic re-linearization and re-solution of the area-wide QP, and therefore re-formulation of the PC-RT optimization problem, is illustrated in Figure 6-5. In this figure, the topmost diagram shows the initial condition when the freeway segment is severely congested. Hence, the first approximation to the (backward-traveling) wave speed is shown as the line between the nominal point and the current point. After a few minutes, the constraints on the upstream flow rates will begin to lower the congested density and, correspondingly the volume exiting the congested segment increases. Thus, in the middle diagram, the wave-speed is re-approximated and the area-wide coordination problem is re-solved with the new (higher) volume in the congested segment. Similarly, as the density continues to decrease and the volume continues to increase in the congested segment, the PC-RT and area-wide coordination problems are re-formulated and re-solved again, as shown in the final diagram at the bottom of Figure 6-5.

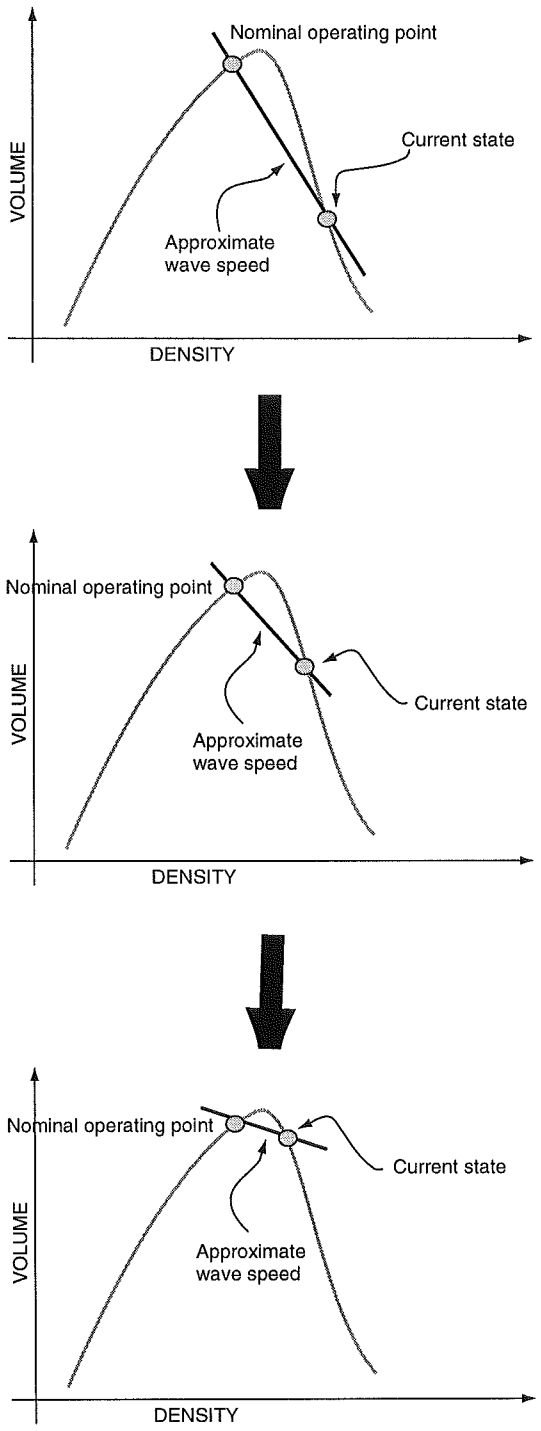


Figure 6- 5. Re-linearization for PC-RT and periodic solution of the QP

Summary of PC-RT mathematical formulation

To summarize the development of the PC-RT optimization problem to this point, we have specified the following single-objective *monolithic* control problem for the entire freeway system to save additional travel time where possible

$$\min \sum_{k=1}^{K}\left[\sum_{j=1}^{J} c_j \Gamma_j \Delta\rho_j(k) + \sum_{i=1}^{M} c_{i,q}\Delta q_i(k)\right] \qquad \textbf{Eqn. 6- 31}$$

st

$$\Delta\rho(k+\Delta T) = \Delta\rho(k) + \Delta T\left(\mathbf{A}^*\Delta\rho + \mathbf{B}^*\Delta r\right) \quad \forall k$$

$$\Delta q_i(k+1) = \Delta q_i(k) + \Delta T\left[\Delta d_i(k) - \Delta r_i(k)\right] \quad \forall i,k$$

if $\rho_j(k) \le \rho_{j,crit} \quad \forall j$

$$\max[\rho_{j,MIN}, -\rho_{j,N}] \le \Delta\rho_j(k) \le \min[\rho_{j,crit} - \rho_{j,N}, \rho_{j,MAX}] \quad \forall k$$

otherwise

$$\max[\rho_{j,MIN}, \rho_{j,crit} - \rho_j] \le \Delta\rho_j(k) \le \min[\rho_{j,jam} - \rho_j, \rho_{j,MAX}] \quad \forall k$$

$$r_{i,MIN} - r_{i,N} \le \Delta r_i(k) \le \hat{r}_{i,MAX} - r_{i,N} \quad \forall i,k$$

$$\Delta q_{i,MIN}(k) \le \Delta q_i(k) \le \Delta q_{i,MAX}(k) \quad \forall\, i,k$$

$$\Delta q_{i,MIN}(k) = -q_{i,N}(k) \quad \forall\, i,k$$

$$\Delta q_{i,MAX}(k) = q_{i,MAX} - q_{i,N}(k) \quad \forall\, i,k$$

$$q_{i,MAX} = Q_i + z_i \quad \forall i$$

$$\Delta r_i(k) = \Delta r_i(k+1) = \cdots \Delta r_i(k+t) \qquad \forall\, i$$

$$\Delta r_i(k+t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+2t)$$

$$\vdots$$

$$\Delta r_i(k+(Z-1)t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+Zt)$$

This formulation is a linear programming problem with unrestricted-sign variables $\Delta r_i(k)$, $\Delta q_i(k)$, and $\Delta\rho_j(k)$.

This LP has $J*K$ freeway variables $\Delta\rho_j(k)$ and $M*K$ queue variables $\Delta q_i(k)$, $3*M*K$ queue constraints and $4*J*K$ freeway constraints. The PC-RT rate regulation optimization problem is thus appreciably larger than the upper-layer area-wide QP as a *single* optimization problem, but not insurmountably large, in the "large-scale" sense, given

modern computing power and solution algorithms. The difficulty, however, of solving this monolithic optimization problem (and the primary reason for the hierarchical decomposition of the freeway management problem) is that the solution depends on the predictions of (1) $\Delta d_i(k)$ and (2) $\Delta V_0(k)$ to the freeway system at the ramp meter and upstream freeway input, respectively, over the short-term time horizon $K\Delta T$. The following development establishes the reasoning behind the decomposition of the monolithic control problem into subproblems.

Difficulty in solving the monolithic PC-RT optimization problem

Consider first that, over the time horizon $K\Delta T$ of the PC-RT optimization problem, the nominal inputs, as assumed in the area-wide coordination problem, $d_{i,N}(k) = \overline{d}_{i,N}$ and $V_{0,IN}(k) = \overline{V}_{0,IN}$ remain constant. At $k=0$ we measure the current d_i and $V_{0,IN}$ and find that $V_{0,IN}(k) \neq V_{0,IN}$ and/or $d_i(k) \neq d_{i,N}$. At this point we try to predict what *could* happen in the next 5-7 minutes to $d_i(k)$ and $V_{0,IN}(k)$ and the corresponding effect on the total system.

Consider each ramp having three "fundamental" predicted scenarios, for $d_i(k)$;

(1) increases $d_i(k) > d_i(k-1)$,

(2) decreases $d_i(k) < d_i(k-1)$, or

(3) remains constant $d_i(k) = d_i(k-1)$.

Similarly the upstream freeway input $V_{0,IN}(k)$ can

(1) increase further $V_{0,IN}(k) > V_{0,IN}(k-1)$,

(2) decrease $V_{0,IN}(k) < V_{0,IN}(k-1)$, or

(3) remain constant $V_{0,IN}(k) = V_{0,IN}(k-1)$.

Thus, in eqn. 6-44 we need to evaluate 3^{M+1} LPs to evaluate the optimal rate combinations for each *possible* combination of predicted inputs at the M ramps and 1 freeway upstream input. After solving 3^{M+1} LPs and measuring the actual demands at each ramp in the next observation period, we must search (and therefore also store) a $M+1$-dimensional table to extract the appropriate set of metering rates.

For a typical problem size, say 10 ramps, evaluating just three predicted scenarios at each ramp necessitates solving 3^{11} = 177,147 LPs each minute to obtain the new metering rate for each possible combination of predicted demands and be guaranteed that the "optimal" metering rate is applied in the next minute. Thus, our approach is to use a sub-optimal procedure for the solution of the PC-RT rate regulation problem and *decompose* the monolithic problem into *M* smaller problems. For example, in an area-wide coordination problem with 10 ramps, we evaluate just (3*3)*M = 90 LPs and search *M 2-dimensional* rate tables for the appropriate rate for the next observation period. Such a scheme is more applicable for real-time control, hopefully with little degradation in performance.

An argument to suggest that the performance would *not* degrade considerably is simply to consider the fact that the predictions being made are *approximate*, and thus the inaccuracy of the predictions likely outweighs any benefits of real-time solutions between the ramps by solving many, many inaccurate monolithic PC-RT rate regulation optimization problems. Further, because of the short-time horizon (5-7 minutes) and continual re-optimization in a rolling-horizon fashion (1-2 minutes), the inaccuracies of predictions should be suitably mitigated. The next section details how the monolithic PC-RT rate regulation optimization problem is decomposed into sub-problems.

Decomposition of full optimization problem into subproblems

For continuous linear systems, many methods exist to decompose a large-scale system into subsystems to satisfy various objectives of reliability, bounded stability, or strength of interaction between subsystems [Isaksen and Payne, 1973; Sandell et al., 1978; Goldstein and Kumar, 1982; Papageorgiou and Mayr, 1982]. For the PC-RT rate regulation optimization problem, our decomposition relies on the concept of the "strength of interaction" between subsystems as well as the requirement to keep the optimization problem *computationally tractable* in real-time. That is, all computations must be completed in less than one-minute for all ramps in the system.

Hence, although we are guided by the analytical approaches of previous work, we implement a straightforward decomposition principle of including *one* ramp per

subsystem, with enough upstream freeway to satisfy the predicted travel-time requirement and enough downstream freeway to allow for anticipatory "reactivity" to downstream congestion. In general, from the topology of most freeway networks (i.e. most ramps are spaced approximately 1 mile apart), our decomposition scheme results in the following *overlapping* subsystem description [Kahng et al., 1984; Haimes et al., 1990] shown in Figure 6-6.

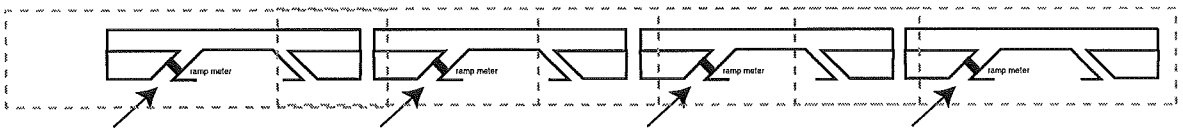


Figure 6- 6. Typical overlapping subsystem decomposition

Hence, we *decouple* (6-33) into M individual-ramp subproblems of the same form

$$\min \sum_{k=1}^{K} \left[\sum_{m=j-t}^{j+1} c_m \Gamma_m \Delta\rho_m(k) + c_{i,q} \Delta q_i(k) \right] \qquad \textbf{Eqn. 6- 32}$$

st

$$\Delta\rho_m(k+\Delta T) = \Delta\rho_m(k) + \Delta T\left(\mathbf{A}^{*j}\Delta\rho_m(k) + \mathbf{B}^{*j} IO_j(k)\right) \quad \forall m \in I^j, \forall k$$

$$IO_j(k) = [V_{j-B,IN}(k), 0, 0, \ldots, 0, r_i(k), V_{j+1,out}(k)]$$

$$\Delta q_i(k+1) = \Delta q_i(k) + \Delta T\left[\Delta d_i(k) - \Delta r_i(k)\right] \quad \forall k$$

if $\rho_m(k) \le \rho_{m,crit} \quad \forall m \in I^j$

$$\max[\rho_{m,MIN}, -\rho_{m,N}] \le \Delta\rho_m(k) \le \min[\rho_{m,crit} - \rho_{m,N}, \rho_{m,MAX}] \quad \forall k$$

otherwise

$$\max[\rho_{m,MIN}, \rho_{m,crit} - \rho_m] \le \Delta\rho_m(k) \le \min[\rho_{m,jam} - \rho_m, \rho_{m,MAX}] \quad \forall k$$

$$r_{MIN} - r_{i,N} \le \Delta r_i(k) \le \hat{r}_{MAX} - r_{i,N} \quad \forall k$$

$$\Delta q_{i,MIN}(k) \le \Delta q_i(k) \le \Delta q_{i,MAX}(k) \quad \forall k$$

$$\Delta q_{i,MIN}(k) = -q_{i,N}(k) \quad \forall k$$

$$\Delta q_{i,MAX}(k) = q_{i,MAX} - q_{i,N}(k) \quad \forall k$$

$$q_{i,MAX} = Q_i + z_i$$

$$\Delta r_i(k) = \Delta r_i(k+1) = \cdots \Delta r_i(k+t) \qquad \forall\, i$$
$$\Delta r_i(k+t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+2t)$$
$$\vdots$$
$$\Delta r_i(k+(Z-1)t+1) = \Delta r_i(k+t+2) = \cdots \Delta r_i(k+Zt)$$

such that A^{*j} is defined as

$$\mathbf{A}^* = \begin{bmatrix} a_{j-B,j-B} & a_{j-B+1,j-B} & 0 & \cdots & 0 \\ a_{j-B,j-B+1} & a_{j-B+1,j-B+1} & a_{3,j-B+1} & & \vdots \\ 0 & & \cdots & & 0 \\ \vdots & & a_{j-1,j} & a_{j,j} & a_{j+1,j} \\ 0 & \cdots & 0 & a_{j,j+1} & a_{j+1,j+1} \end{bmatrix} \qquad \textbf{Eqn. 6- 33}$$

and B^{*j} is

$$\mathbf{B}^* = \begin{bmatrix} 1 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 1 \\ & & & & & 1 \end{bmatrix} \qquad \textbf{Eqn. 6- 34}$$

using a numbering convention where the most upstream density section in the freeway is designated as ρ_1. $IO_j(k)$ is a vector of the inputs and outputs to this ramp subsystem, such as the upstream freeway input $V_{j\text{-}B,IN}(k)$, ramp input $r_i(k)$ and off-ramp output $s_i(k)$, and downstream freeway output flow $V_{j+1,OUT}(k)$. The example shown in (6-34) indicates no off-ramp in the section with the on-ramp, and no other on-ramps in the sub-system.

According to the modeling convention of the macroscopic freeway model of Chapter 4, each freeway *section j* is composed of num_j *segments* and thus the total number of segment in the freeway model is J^*num_j. The next modeling convention is that each on-ramp or off-ramp is found in the *first* segment of a new section, such that the changes in the system are found at the 1st, num_1+1, $num_1 + num_2 + 1$,...., $\sum_{j=1}^{N-1} num_{j-1} + 1$ segments of the whole system. Thus, the set I^j is defined as

$$I^j = [j-B, j-(B-1), \ldots, j, j+1] \qquad \textbf{Eqn. 6- 35}$$

where j is the segment containing the i^{th} ramp ($num_1 + num_2 + \ldots + num_{(j-1)} + 1$) and B is the minimum between the number of segments upstream that constitute more than one-minute travel time, but less than two minutes, or the number of segments until another ramp is encountered. The choice of one-minute travel time is from the specification that the metering rate only be changed in one-minute intervals. Note also that as the travel-time in each segment of the system changes, as during incident conditions, the size definitions of each sub-problem may change.

This criterion for B also ensures that only *one* segment with an on-ramp occurs in each sub-problem and thus the metering rate decisions made by each PC-RT subproblem are independent. It may be necessary, however, in some real-world freeway systems where there are several ramps less than one-minute of travel-time apart, to consider their real-time optimization together using a different decomposition scheme. This possibility is not explored in this research.

We also assume, in the specification of the subproblems, that the volume immediately downstream $V_{j+1,OUT}$ of the subsystem remains at the nominal flow rate $\overline{V}_{j+1,OUT,N}$ during the PC-RT optimization horizon. This is a simplifying assumption since the downstream flow could *also* be considered a location for detection of flow anomalies $V_{j+1,OUT}(k) >> V_{j+1,OUT,N}$ *or* $V_{j+1,OUT}(k) << V_{j+1,OUT,N}$. With this addition, however, the number of optimization problems would increase three-fold, and hence we would need to solve for 3X3X3 = 27 prediction scenarios. This possibility may be studied in future research.

Scenario prediction

To provide the *pro-active* capability of the PC-RT optimization layer we need to predict the possible changes to the ramp demands and the upstream freeway flow for each sub-problem. When a significant deviation to the flow rate is detected one minute upstream of the ramp section $V_{j-B}(k) \notin \left[V_{j-B,N} \pm \Delta V_{j-B,N}\right]$ on the freeway and/or at the ramp meter

entrance $d_i(k) \notin \left[d_{i,N} \pm \Delta d_{i,N}\right]$ (also one-minute upstream of the interchange flows), a *set* of possible future scenarios is created, each with a corresponding LP optimization problem. Since the detected flow variation is one minute upstream, we do not need to select the appropriate rate immediately. Thus, the procedure is as follows

(a) Detect the anomaly $d_i(k) \notin \left[d_{i,N} \pm \Delta d_{i,N}\right]$, and/or $V_{j-B}(k) \notin \left[V_{j-B,N} \pm \Delta V_{j-B,N}\right]$,

(b) create the set of predictions and solve for travel-time minimizing metering rates with the PC-RT subproblem decomposition and store in a table,

(c) in the next minute, sample the flow rate $V_{j-B}(k+1)$ and $d_i(k+1)$ again one-minute upstream and evaluate which predicted scenario was closest to being realized,

(d) apply the appropriate metering rate to "react" to the realized system state from the stored table of possible rates for ramp *i, and*

(e) repeat the entire process for the current measurement.

Hence, the predictions do not necessarily need to be entirely accurate, but they do need to reflect both

(a) the *range* of conditions possible over the short-term time horizon of 5-7 minutes, and

(b) the *possible dynamics* over the short-term time horizon of 5-7 minutes.

In addition, at this level of decoupling, we neglect the fact that, for sub-problems in the *interior* of the system, we have more upstream information than just one-minute of travel-time. Neglecting this information does not seem egregious since the uncertainty of the entire flow process and the fact that our predictions are approximate (at best) will probably outweigh any benefits of using the additional upstream flow information for flow predictions in the interior of the freeway.

As previously stated, to represent the range of possible future states three *fundamental* prediction scenarios are created for the two input flows to each sub-problem: *increasing* flow linearly, *decreasing* flow linearly, and *remaining* at the current level, as illustrated is

Figure 6-7. The three predicted scenarios for the upstream freeway segment are labeled A_{fy}, B_{fy}, and C_{fy} and A_{rd}, B_{rd}, C_{rd} identify the three predicted demand scenarios to the ramp meter.

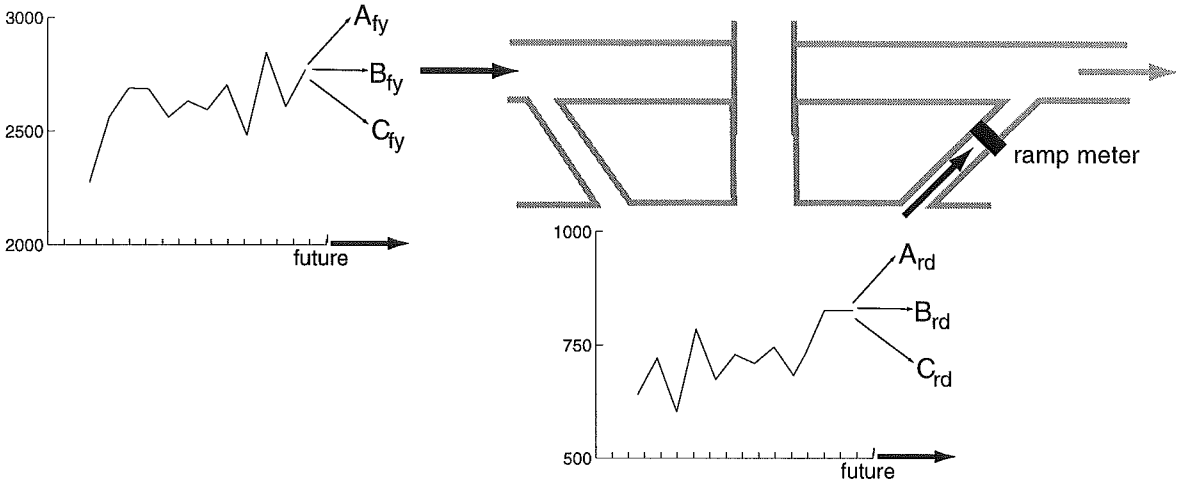


Figure 6- 7. Predicted trends for a given subproblem

Scenario prediction example

Recall that we plan the PC-RT optimization of a sub-problem *only* when an anomaly $d_i(k) \notin \left[d_{i,N} \pm \Delta d_{i,N}\right]$ and/or $V_{j-B}(k) \notin \left[V_{j-B,N} \pm \Delta V_{j-B,N}\right]$ is detected at one or both of the input streams, so the "remaining at current level" prediction could or could not be an anomalous condition. For example, consider a subsystem where the nominal demand to the ramp meter is 600 veh/hr and a demand of 750 veh/hr is detected one minute upstream. At the same subsystem the nominal demand to the upstream freeway segment is 1700 veh/hr-lane and a flow of 1850 veh/hr-lane is detected. Assume that the volume level at the upstream freeway input $V_{j-B,IN}(k) = 1850$ constitutes an anomalous flow condition for this subsystem. When this is detected, three predictions are created for the possible ramp flows and three predictions are created for the upstream freeway volume over the next five to seven minutes, for a total of *nine* possible scenarios. We do not

require that both the freeway and ramp demands be anomalous to begin PC-RT optimization.

Consider also, for example, that ramp demands can change, on average, (after smoothing out short-term detector errors and removing the cyclical effect of the interchange traffic signal) approximately 15 veh/hr per minute, either linearly positive or negative, such that

$$d_i(k+1) = d_i(k) \pm 15. \qquad \textbf{Eqn. 6- 36}$$

In addition, assume that on average (after smoothing) freeway volume can change approximately 60 veh/hr per minute linearly, such that

$$V_{j-B,IN}(k+1) = V_{j-B,IN}(k) \pm 60. \qquad \textbf{Eqn. 6- 37}$$

Hence, the *fundamental* predicted time-series of flows to the ramp meter for the next five minutes are given in Table 6-1 and the *fundamental* predicted time-series of flows at the upstream freeway segment for the next five minutes are given in Table 6-2.

Prediction	Minute 1	Minute 2	Minute 3	Minute 4	Minute 5
Increasing	765	780	795	810	825
Constant	750	750	750	750	750
Decreasing	735	720	705	690	675

Table 6- 1. Ramp meter demand predictions

Prediction	Minute 1	Minute 2	Minute 3	Minute 4	Minute 5
Increasing	1910	1970	2030	2090	2150
Constant	1850	1850	1850	1850	1850
Decreasing	1810	1750	1690	1630	1570

Table 6- 2. Upstream freeway demand predictions

With these predictions and the initial conditions of the subsystem (i.e. $q_i(0)$, $\rho_{j,N}$, $v_{j,N}$, $r_{i,N}$, $d_i(0)$, and $V_{j\text{-}B,IN}(0)$), nine optimization problems of the form 6-45 are solved for the metering rates over the next five minutes that minimize additional travel time in the subsystem.

Construction of scenario rate tables

The result of solving the nine optimization problems is a 3X3 table of metering rate changes from the current metering rate, as illustrated in Table 6-3. Table 6-3 indicates the metering rate modification to be applied depending on the scenario that is actually detected in the next minute at both of the input streams to the subsystem. Note also that, at each minute, the table entries will change as the initial conditions and flow rates change. Thus, for further clarification, the row and column headings *constant* in Table 6-3 do not indicate that the flow remains at the *nominal* ramp demand and/or upstream freeway flow, but rather could mean that some very high ramp demand and/or upstream freeway flow will *remain* very high for the next five to seven minutes. Thus, each time the PC-RT problem is re-solved for a subsystem, a new rate table is constructed that *only* applies to the current combination of local conditions $q_i(0)$, $\rho_{j,N}$, $v_{j,N}$, $d_i(0)$, $V_{j\text{-}B,IN}(0)$ and nominal metering rate $r_{i,N}$.

freeway demand / **meter demand**	**increasing**	**constant**	**decreasing**
increasing	0	+50	+100
constant	-20	0	+20
decreasing	-50	-25	0

Table 6- 3. Rate table for next minute

Infeasible PC-RT scenarios

It may be the case, however, that some predicted flow trends result in infeasible PC-RT optimization problems. Say, for example, that the freeway density is very near the critical density $\rho_j(k) \approx \rho_{j,crit}$, and the prediction of this scenario is for increasing

upstream freeway volume $V_{j-B,IN}(k+1) = V_{j-B,IN}(k) + 60$. It may not be possible with limited control of the local ramp meter alone to keep the freeway from becoming congested $\rho_j(k) \leq \rho_{j,crit} \quad \forall j,k$ even by reducing the metering rate all the way to its minimum allowable rate $\Delta r_i = r_{i,MIN} - r_{i,N} \quad \forall k$. Recall that this minimum allowable rate is *not* the same as the *absolute* minimum allowable rate. At the same time, it may also not be possible to keep the ramp queue from spilling back (or spilling back further, if spillback was already planned by the area-wide QP) into the adjacent interchange. If either or both of these events are true, the PC-RT scenario is infeasible and the optimization routine should provide a recommendation to the upper-layer processor(s) to re-solve for new nominal metering rates if this combination of freeway and ramp flows is realized in the next minute. The existence of infeasible scenarios is demonstrated in Table 6-4.

This strategy of re-solving the area-wide coordination problem when a PC-RT problem is infeasible addresses the local nature of the PC-RT optimization problems and the hierarchical structure of the MILOS control system. First, the local PC-RT controllers do *not* know the relative congestion levels of the interchanges in the corridor and thus if the PC-RT control algorithm is allowed to make the decision to spill-back the ramp, it could well be spilling-back the queue at the *most* congested interchange in the network. In this way, the PC-RT subproblems remain sensitive to the priorities specified at the area-wide coordination layer, but in a *decentralized* manner. That is to say, for example, that a subproblem with *large* weighting coefficients c_j for the freeway variables $\Delta\rho_j(k)$ and a relatively *small* weighting coefficient $c_{i,q}$ for the ramp queue changes $\Delta q_i(k)$ does not "know" that this combination of weighting coefficients is based on the relative costs of the *entire* area-wide problem. Instead, the subproblem is provided the information that, for its part of the problem, changes to the freeway flow $\Delta\rho_j(k)$ are much more expensive to system operation than changes to the ramp queue growth $\Delta q_i(k)$. Hence, re-solving the area-wide coordination problem allows us to distribute the spill-back amongst several

upstream ramps that have lower interchange congestion levels than the current ramp, better than re-solution of the PC-RT problem with less restrictive $r_{i,MIN}$ and $r_{i,MAX}$ settings.

The second reason not to allow the PC-RT algorithm too much ability to modify the metering rate (and therefore to keep a given PC-RT scenario *feasible*) is that this flexibility could likely result in oscillatory behavior of alternately spilling-back the ramp queue (with a restrictive rate) and then "dumping" the queue at a very high metering rate. This oscillatory behavior is particularly unacceptable to practicing traffic engineers since it is viewed as defeating the purpose of ramp metering altogether [TAC, 1998].

freeway demand / meter demand	increasing	constant	decreasing
increasing	resolve QP	+50	+100
constant	resolve QP	0	+20
decreasing	resolve QP	-25	0

Table 6- 4. Rate table with infeasible optimization problems

Summary

The PC-RT optimization subproblem layer of the MILOS hierarchical control system is a pro-active approach to planning real-time, traffic-responsive ramp metering rates that considers surface-street conditions in its optimization. Queue management is explicitly considered in the optimization formulation to realize "additional" travel-time savings at each ramp by taking advantage of opportunities to reduce the queue when the freeway is under-utilized and hold back additional vehicles when there is queue capacity and the freeway has a short-term surge in demand. The PC-RT optimization problem structure is tightly integrated with the solution of the area-wide coordination problem since the cost coefficients are derived from the dual multipliers and slack values of the constraints in the area-wide coordination problem. PC-RT optimization runs at each single-ramp subsystem are scheduled when a significant difference between either the upstream freeway flow rate, or the ramp demand rate (or both) is detected. A monitoring function

is required to detect such statistically-significant flow fluctuations. The theoretical foundation and typical operation of this process monitoring function, based on the concepts of statistical process control (SPC), is the subject of the next chapter.

Chapter 7: SPC-based anomaly detection

Introduction

One of the fundamental attributes of MILOS is the periodic re-optimization of the area-wide coordination problem and the PC-RT rate regulation sub-problems at the appropriate times to adjust the ramp metering rates to the current freeway and surface-street conditions. Our approach in this research is to schedule a re-optimization of a subproblem only when the network conditions *warrant* a re-optimization. Thus, a subproblem is only re-solved when the current relevant state variable measurement(s) vary significantly from the *nominal* predicted state, such that $V_j(k) \neq V_{j,N}$ and/or $d_i(k) \neq d_{i,N}$ as developed previously in Chapter 6.

We denote a significant deviation of a state variable from its nominal setting $V_j(k) \neq V_{j,N}$ as an *anomaly*, not to be confused with an *incident*. An *incident* could be defined as an anomaly that causes congestion and/or reduction of capacity in a given segment(s). The method developed in this chapter could ultimately be used for incident detection, but it is intended for more general identification of flow changes. The central questions to be addressed in this chapter are;

(1) How is an anomalous condition *defined*?

(2) How can an anomalous condition be *detected*?

We answer these questions by applying the statistical process control concept of *control limits* for tracking structural changes of a dependent-variable time-series. We will show that effective operation of a two-level optimization structure, such as the combination of an area-wide coordination problem and a set of predictive-cooperative real-time metering problems, is highly dependent upon the accurate identification of such deviations of the appropriate state-variables.

Overview of Statistical Process Control

Statistical process control (SPC) is based on comparing the sample mean $\bar{x}(k)$ and/or the sample variance $s^2(k)$ of sub-sample k to the long run average $\bar{\bar{x}}$ for evidence of *structural* changes to the underlying process. The last N sub-samples ($[\bar{x}(k\text{-}N),\ldots, \bar{x}(k)]$, $[s^2(k\text{-}N),\ldots s^2(k)]$) are used in SPC to identify short-term trends in the process parameters. A typical SPC control chart for the mean $\bar{x}$ is shown in Figure 7-1. Note in Figure 7-1 the

upper control limit and *lower control limit*, which bracket the long run average $\bar{\bar{x}}$ and form the acceptable interval of the sample mean $\bar{x}(k)$.

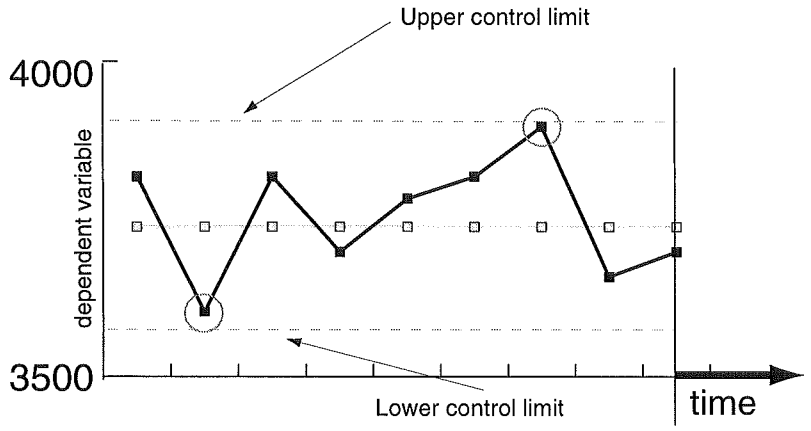


Figure 7- 1. Typical SPC control chart

Other information in the time-series of the quality characteristic $\bar{x}$ can be detected using control charts, such as a "jump" from one level to another such that

$$\bar{x}(k) >> \bar{x}(k-1) \quad or \quad \bar{x}(k) << \bar{x}(k-1). \qquad \textbf{Eqn. 7- 1}$$

The goal of using SPC is to *identify* such trends or changes to the operating point using the statistics $\bar{x}$ and s^2 of the process and apply the necessary corrections to redirect the process back to the target value x^*.

The process under control, however, has undeniably random components to the time-series $(x_1,x_2,\ldots,x_n)$ that results from exogenous factors that may or may not be influenced by the controllable parameters $(v_1, v_2, \ldots, v_n)$. Thus, some variability in the quality characteristic is to be expected and can be tolerated, but only that variability that results in the process average $\bar{x}$ remaining between the control limits. Thus, there are two levels of control limits; $[UCL_x, LCL_x]$ which indicate whether or not a given part x is useable or not, and $[UCL_{\bar{x}}, LCL_{\bar{x}}]$ which indicate if the process is under "control" or not and the time-series is remaining "close to" the desired target value x^*. Hence, the dependent-variable specifications $[UCL_x, LCL_x]$ must be *significantly* separated from the process control limits $[UCL_{\bar{x}}, LCL_{\bar{x}}]$ for reasonable application of SPC, as shown in Figure 7-2.

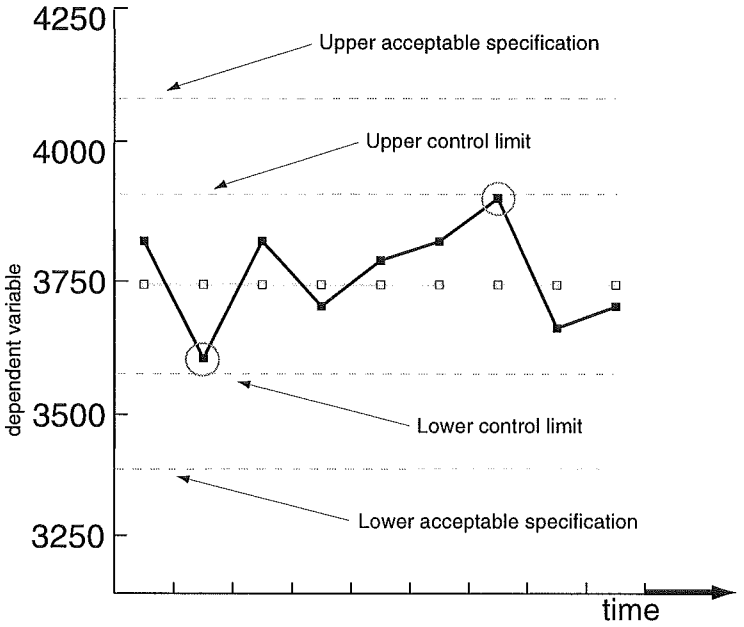


Figure 7- 2. SPC limits and part specifications

The final important characteristic of SPC is that, while data is taken one-by-one, the time-dependence of the quality characteristic measurements x within a sample is neglected. When drawing a sub-sample $[x_{s1},\ldots x_{st}]$ of t parts from the time-series fragment $[x_{k\text{-}N},\ldots.x_k]$ of N parts (a sample)from time k-N to k, the measurements are considered independent and identically distributed so that the sample statistics $\bar{x}(k)$ and $s^2(k)$ (or R_k) can be computed. As shown in Figure 7-3, the gray points indicate the samples drawn from the population of each production run and the dark black line indicates the $\bar{x}(k)$ time-series.

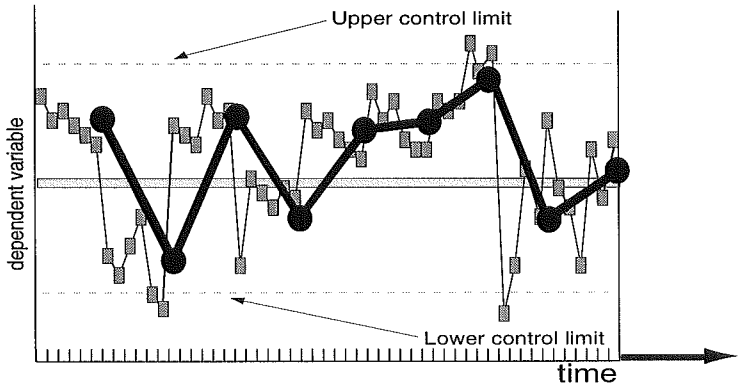


Figure 7- 3. SPC chart showing sampled time-series

Relationship of SPC concepts to freeway control

SPC can be used as a process estimation and tracking tool in the hierarchical freeway control problem for two primary reasons:

(1) traffic detector data processes have many similarities to the characteristics of production process time-series controlled with SPC methods, and

(2) the SPC method as a process estimator, exhibits the *two-level* structure necessary for use with a *two-level* control algorithm, such as MILOS.

Typically, detector data is collected as counts over a passage detector or as the occupancy level of a presence detector. The volume level is derived from that observations, say by converting vehicles/min to vehicles/hour by averaging across all lanes at a station to smooth lane-to-lane fluctuations. The smallest reported interval of detector measurements is typically not less than 20-seconds. The last three 20-second observations are averaged again to derive one-minute volumes, and so on, as shown in Figure 7-4. Other filtering schemes have been proposed to dampen sharp fluctuations and reduce variability of the measurements [Okutani, 1987; Coifman, 1996].

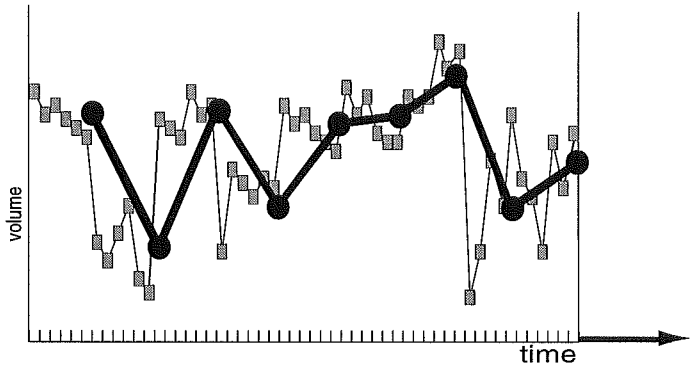


Figure 7- 4. Detector time-series and underlying detection history

The example detector data time-series in Figure 7-4 is deliberately shown to have the same profile as the SPC process chart of Figure 7-3 to indicate the similarity of the two concepts. In the detector output chart of Figure 7-4 we are not attempting to "control" the output at each point like in a traditional SPC application, but we can use the similarities of the two time-series processes to derive [UCL_x, LCL_x] and [$UCL_{\bar{x}}$, $LCL_{\bar{x}}$] control limits for the detector data stream. These limits will be used to indicate when our underlying process, e.g. the *volume* or *occupancy* of the detector group, has changed *significantly*.

An example of this identification and re-estimation process for a longer time-series is shown in Figure 7-5. Figure 7-5 illustrates the intended use of the SPC technique to identify:

(a) statistically-significant random fluctuations in the detector data time-series (an *anomaly*) and

(b) the transition from one "approximately constant" demand/volume target level x^* to the next (a specific type of *anomaly*, but not necessarily an *incident*).

The five dotted lines in each of the three sections of Figure 7-5 indicate the assumed "constant" target level x^* and the [UCL_x, LCL_x] and [$UCL_{\bar{x}}$, $LCL_{\bar{x}}$] control limits in that section. When the process mean $\bar{x}$ (heaviest line) moves outside the [$UCL_{\bar{x}}$, $LCL_{\bar{x}}$] region for two consecutive intervals, the target level x^* is re-evaluated and new [UCL_x, LCL_x] and [$UCL_{\bar{x}}$, $LCL_{\bar{x}}$] control limits are computed.

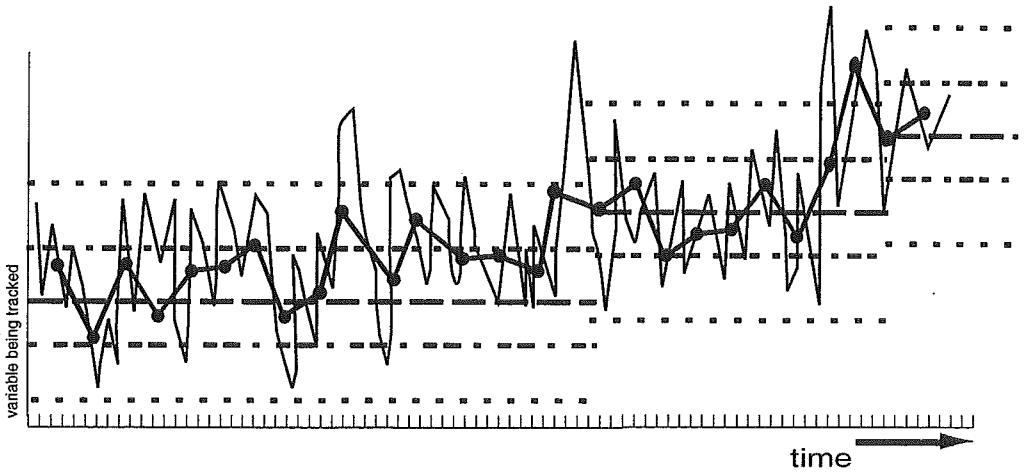


Figure 7- 5. Re-evaluation of "approximately constant" demand level

Justification of approximately-constant demand

It is well known in traffic engineering that the volume increases and then decreases during the peak periods on commuter freeways. Thus, the assumption of "approximately constant" demand in short time periods seems unrealistic during the peak-period, since the trend is not being considered in the estimation and control procedures. However, it appears natural to approximate the underlying process as being a process with discrete *jumps* between constant demand/volume levels to use the SPC concepts for demand tracking in conjunction with two-level optimization approach of MILOS. This "jump-process" assumption is illustrated in Figure 7-6.

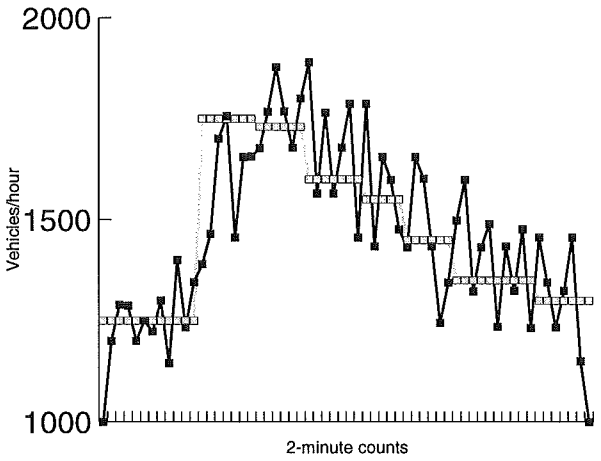


Figure 7- 6. Jumps between approximately-constant demand levels

The jump-process assumption depicted in Figure 7-6 integrates well into the MILOS hierarchical control method since the area-wide coordination problem is solved and re-solved for the nominal metering rates $r_{i,N}$ with "constant" demand inputs $d_{i,N}$, represented as the approximately-constant levels x^* in Figure 7-5. Then the *nominal* metering rate $r_{i,N}$ is modified in real-time with the PC-RT rate regulation module to adapt to the statistical variation of the actual flows $d_i(k)$, monitored and identified by the SPC anomaly detection module. Thus, in the MILOS hierarchy, SPC is used to fulfill two objectives:

(1) to identify when a PC-RT re-optimization is required at each ramp, and

(2) to identify when the area-wide coordination problem should be re-solved.

The objectives (1) and (2) map directly to the establishment of [$\mathrm{UCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$] and [$\mathrm{UCL}_{\bar{x}\,\mathrm{QP}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{QP}}$] *control limits* to denote the association with the PC-RT and QP optimization problems, respectively. If the monitored variable $d_i(k)$ and/or $V_{j\text{-}B}(k)$ exceeds the *inner* set of control limits [$\mathrm{UCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$], PC-RT re-optimization is scheduled. If either $d_i(k)$ and/or $V_{j\text{-}B}(k)$ exceeds the *outer* control limits [$\mathrm{UCL}_{\bar{x}\,\mathrm{QP}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{QP}}$] in a trend, a re-solve of the area-wide coordination problem is scheduled using a new estimate of the process "target value" x^* for the violating variable. We now present the computational procedure for deriving the control limits [$\mathrm{UCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{PC\text{-}RT}}$] and [$\mathrm{UCL}_{\bar{x}\,\mathrm{QP}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{QP}}$] and a heuristic forecasting technique for specifying the new process target value(s).

SPC Computational Procedure

SPC control limits are based on the concept from statistics of *confidence intervals*. Confidence intervals determine the trust-region of an estimate $\bar{x}$ for the true mean μ of a distribution when the samples $(x_1,x_2,....,x_n)$ come from a time-series. A confidence interval

$$\left[\mu - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\quad \mu + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] \qquad \textbf{Eqn. 7- 2}$$

indicates, with a probability of $1\text{-}\alpha$, that the average $\bar{x}$ of a random sample of size n will fall within the limits. $Z_{\alpha/2}$ is a parameter derived from assuming the process is normally distributed, the confidence level α, and the sample size n. The true mean and variance μ and σ are typically not known, and thus they are replaced with their estimates, $\bar{\bar{X}}$ and $\bar{R}$, respectively. $\bar{\bar{X}}$ is defined as

$$\bar{\bar{X}} = \frac{1}{K}\sum_{i=1}^{K}\bar{X}_i \qquad \textbf{Eqn. 7- 3}$$

where $\bar{X}_i$ is the mean of sub-sample i of the process

$$\bar{X}_i = \frac{1}{N}\sum_{j=1}^{N}x_{i,j} \qquad \textbf{Eqn. 7- 4}$$

and $\bar{R}$ is defined as

$$\bar{R} = \frac{1}{K}\sum_{i=1}^{K}R_i, \qquad \textbf{Eqn. 7- 5}$$

such that R_i is defined as

$$R_i = \max_j\left[x_{i,j}\right] - \min_j\left[x_{i,j}\right]. \qquad \textbf{Eqn. 7- 6}$$

$\bar{R}$ is typically substituted for s^2 in estimating σ in SPC. $\bar{R}$ estimates σ such that

$$\hat{\sigma} = \frac{\bar{R}}{d_2} \qquad \textbf{Eqn. 7- 7}$$

where d_2 is a tabulated constant based on the sample size. The confidence interval for μ is then given by

$$\left[\overline{\overline{X}} - \overline{R}\left(\frac{3}{d_2\sqrt{n}}\right), \quad \overline{\overline{X}} + \overline{R}\left(\frac{3}{d_2\sqrt{n}}\right)\right]. \qquad \textbf{Eqn. 7- 8}$$

The factor of 3 is derived from the fact that 3 standard deviations from the mean (in each direction) typically encompasses 99.9% (α = 0.001) for of all samples from a normal distribution. Redefining the quantity

$$A_2 = \frac{3}{d_2\sqrt{n}} \qquad \textbf{Eqn. 7- 9}$$

results in the confidence interval

$$\left[\overline{\overline{X}} - A_2\overline{R}, \quad \overline{\overline{X}} + A_2\overline{R}\right] \qquad \textbf{Eqn. 7- 10}$$

where A_2 is tabulated based on sample size n, and $\alpha = 0.001$. In this application of SPC to anomaly detection, we require *two* sets of limits, inner and outer, for real-time control and area-wide coordination, respectively. The UCL and LCL derived by 3σ levels are invariably the *outer* process limits that indicate, when exceeded, that the process is no longer in the same approximately-constant regime. These limits will be denoted [$\mathrm{UCL}_{\overline{x}\,\mathrm{QP}}$, $\mathrm{LCL}_{\overline{x}\,\mathrm{QP}}$]. The *inner* set of control limits, denoted [$\mathrm{UCL}_{\overline{x}\,\mathrm{PC\text{-}RT}}$, $\mathrm{LCL}_{\overline{x}\,\mathrm{PC\text{-}RT}}$], for scheduling PC-RT optimization at a ramp are derived from the *outer* control limits such that

$$\left[\overline{\overline{X}} - A'_2\overline{R}, \quad \overline{\overline{X}} + A'_2\overline{R}\right] \qquad \textbf{Eqn. 7- 11}$$

defines [$\mathrm{UCL}_{\overline{x}\,\mathrm{PC\text{-}RT}}$, $\mathrm{LCL}_{\overline{x}\,\mathrm{PC\text{-}RT}}$] where

$$A'_2 = \theta A_2 \qquad \textbf{Eqn. 7- 12}$$

and θ is a positive constant $0<\theta<<1$ chosen by the user of the control system based on engineering judgment. The parameter θ could be updated by the control system based on performance, but it is not readily apparent how such updates could be made without extensive further analysis of detector data time-series. This is a topic for further research.

Of course, after some trial and error, a value of θ which produces "acceptable performance" in simulation experiments should be chosen as the initial estimate of the inner control limit specification $A'_2 = \theta A_2$. In this context, "acceptable performance" of the MILOS hierarchical control system would indicate that the PC-RT rate regulation

optimizations do not run *too often*, nor do they run *too infrequently* given a certain level of stochasticity in the ramp demand $d_i(k)$ and freeway flow $V_{j\text{-}B}(k)$ measurements.

Transition to a new $\overline{\overline{X}}$ level

The performance of MILOS as a two-level optimization problem is highly dependent upon:

(1) how the transition from one approximately-constant level $V_{j\text{-}B,N}$ or $d_{i,N}$ to the next is detected, and

(2) exactly what new approximately constant level $V_{j\text{-}B,N}$ or $d_{i,N}$ is specified for the next time period.

Taking direction from the SPC literature, we look for the establishment of a *trend* before scheduling re-solution of the area-wide coordination problem. We define a trend, $T_{j,k}$ in freeway section k of segment j, or $T_{r,i}$, a trend at ramp i, as two or more out-of-outer control limit (i.e. [$\text{UCL}_{\overline{x}\,\text{QP}}$, $\text{LCL}_{\overline{x}\,\text{QP}}$]) measurements. As will be further described in Chapter 8 and used in the simulation experiment of Chapter 9, we let $T_{j,k}$ and $T_{r,i}$ take the values (0, 1, 2) to specify the number of consecutive out-of-control measurements in a given section or ramp demand.

The second issue of what demand level should be the next approximately-constant estimate $V_{j\text{-}B,N}$ or $d_{i,N}$ is considerably more complicated. In this application, the estimate of the process mean $\overline{\overline{X}} = \hat{V}_{j-B,N}$ or $\overline{\overline{X}} = \hat{d}_{i,N}$ must change quickly when a new constant level is "detected" to re-establish the control limits [$\text{UCL}_{\overline{x}\,\text{PC-RT}}$, $\text{LCL}_{\overline{x}\,\text{PC-RT}}$] and [$\text{UCL}_{\overline{x}\,\text{QP}}$, $\text{LCL}_{\overline{x}\,\text{QP}}$]. The variability estimates $A_2\overline{R}$ and $A'_2\overline{R}$ can be tracked with the simple averaging technique of (7-5) but $\overline{\overline{X}}$ *cannot* be computed with the averaging technique of (7-3). In traditional SPC, the process target value x^* is not changing over time. In this problem, however, the process target value is being periodically re-estimated. Hence, in each approximately constant demand regime t, consider the definition of a *short-term* process mean $\hat{\overline{\overline{X}}}^t$ defined as

$$\hat{\overline{\overline{X}}}^t = \hat{\overline{\overline{X}}}^{t-1} \pm A_2\overline{R} + f\left(\overline{R},T\right) \qquad \textbf{Eqn. 7- 13}$$

where $\hat{\overline{\overline{X}}}^{t-1}$ is the demand estimate in the *previous* regime and $f\left(\overline{R},T\right)$ is some function of the variance of the process $\overline{R}$ and the time-horizon T of the area-wide coordination

problem. The outer process control limits [$UCL_{\overline{x}\,QP}$, $LCL_{\overline{x}\,QP}$] in each approximately-constant regime are thus defined as

$$\left[\hat{\hat{\overline{X}}}^{t} - A_2\overline{R}, \quad \hat{\hat{\overline{X}}}^{t} + A_2\overline{R}\right] \qquad \textbf{Eqn. 7- 14}$$

and the inner process control limits [$UCL_{\overline{x}\,PC\text{-}RT}$, $LCL_{\overline{x}\,PC\text{-}RT}$] as

$$\left[\hat{\hat{\overline{X}}}^{t} - A'_2\overline{R}, \quad \hat{\hat{\overline{X}}}^{t} + A'_2\overline{R}\right]. \qquad \textbf{Eqn. 7- 15}$$

Hence, the estimate of the mean of the approximately-constant demand $\hat{\hat{\overline{X}}}^{t}$ in interval t is the *outer control limit* (either UCL or LCL, depending on the direction of change from the previous estimate $\hat{\hat{\overline{X}}}^{t-1}$) plus or minus a correction factor $f(\overline{R},T)$.

This choice for the new estimate $\hat{\hat{\overline{X}}}^{t}$ is ad-hoc, but based on operational concerns. Namely, if a new approximately constant level $\hat{\hat{\overline{X}}}^{t}$ is chosen *too close* to $\hat{\hat{\overline{X}}}^{t-1}$, updates to the area-wide coordination problem may be much more (or possibly much less) frequent than their intended application period (10-20 minutes). Similarly, if $\hat{\hat{\overline{X}}}^{t}$ is chosen *too distant* from $\hat{\hat{\overline{X}}}^{t-1}$, there may be oscillations between approximately-constant levels during upward or downward trends in the demand level $d_i(k)$ or $V_{j\text{-}B}(k)$. Thus, in both cases, the area-wide coordination problem (and/or PC-RT optimizations) may be re-solved at more frequent or less frequent intervals depending on the value of $f(\overline{R},T)$. Of course, at this stage of the research, there is no acceptable definition of what constitutes the *correct* ("optimal" is specifically avoided here) number of PC-RT and/or area-wide optimizations during a given interval. For further discussion, see [Gettman, 1998].

Without real-world field performance data from the application of policies for $f(\overline{R},T)$ it is impossible to speculate on which procedure to obtain $\hat{\hat{\overline{X}}}^{t}$ is superior. In this research, we avoid specifying a functional form for $f(\overline{R},T)$ and set its value to a constant based on preliminary testing. Further analysis and tuning of the SPC anomaly detection procedure may be a topic for future research.

Other issues in SPC-based anomaly detection

Recall that, as the SPC anomaly detection module "triggers" the re-solution of the area-wide coordination problem when [$UCL_{\bar{x}\,QP}$, $LCL_{\bar{x}\,QP}$] is exceeded, the remaining estimates $\hat{\bar{\bar{X}}}^t$ are derived from the solution to the QP, such that

$$\hat{\bar{\bar{X}}}_j = \sum_{i=1}^{j} A_{i,j} r_i \qquad \textbf{Eqn. 7- 16}$$

where $A_{i,j}$ is the proportion of flow from ramp i continuing through section j. However, if the anomaly $V_j(k) \neq V_{j,N}$ is detected in the *interior* of the freeway such that [$UCL_{\bar{x}\,QP}$,$LCL_{\bar{x}\,QP}$] is exceeded, one of three possible conditions is indicated:

(1) an upstream *ramp meter* was increased or decreased $r_i(t) >> r_{i,N}$ or $r_i(t) << r_{i,N}$ significantly enough to change the flow rate $V_j(k)$,

(2) one or more upstream *route-proportional rates* $A_{i,j}$ changed significantly, causing eqn. 7-16 to be inaccurate, or

(3) an incident is occurring near this section.

In addition, there could be cases where some combination of the above conditions occur simultaneously. Similarly, two or more conditions could cancel each other out to produce *no* appreciable flow change (e.g. $A_{0,j}(t) >> A_{0,j}(t\text{-}1)$ and $V_0(t) << V_0(t\text{-}1)$ results in $V_j(t) \cong V_j(t-1)$). As such, the problem of SPC anomaly detection in the *interior* of the freeway, for checking for breach of the [$UCL_{\bar{x}\,QP}$, $LCL_{\bar{x}\,QP}$] control limits, is not only to *identify* the new volume level $\hat{\bar{\bar{X}}}^t$ estimate, but also how, on an area-wide basis, the volume level is synthesized by changes to the route-proportional flow rate(s).

The PC-RT rate regulation subproblems and area-wide coordination optimization problems, however, *do* need to know the source(s) of the anomalies so that an accurate model can be constructed. In particular, we will assume that when [$UCL_{\bar{x}\,PC\text{-}RT}$, $LCL_{\bar{x}\,PC\text{-}RT}$] is breached, the route-proportional flow rates have not changed for a given subproblem. Thus, in the modeling procedure of Chapter 6, it is assumed that the upstream flow $V_{j-B}(t)$ changes, but the turning-probability parameter θ does not over the predicted time horizon. These assumptions are used in the evaluation of MILOS in the simulation experiment in Chapter 9.

Summary

A central characteristic of the MILOS ramp metering control system is that a PC-RT subproblem is only re-solved when the current relevant state variable measurement(s) $V_{j-B}(t)$ and $d_i(t)$ vary significantly from the *nominal* predicted state $V_{j-B,N}$ and $d_{i,N}$. Similarly, the area-wide coordination problem is re-solved when even stronger fluctuations are detected to the ramp demands and upstream freeway flows. Such statistically-significant deviation is termed an *anomaly*. An anomaly-detection procedure is developed based on the concept of *control limits* from statistical process control. The effective operation of the two-level MILOS hierarchical control system is highly dependent upon the accurate identification of such deviations $V_j(k) \neq V_{j,N}$ and/or $d_i(k) \neq d_{i,N}$.

Two sets of control limits are established by the SPC-anomaly detection module, the PC-RT re-optimization limits [$\mathrm{UCL}_{\overline{x}\,\text{PC-RT}}$, $\mathrm{LCL}_{\overline{x}\,\text{PC-RT}}$] and the area-wide coordination limits [$\mathrm{UCL}_{\overline{x}\,\text{QP}}$, $\mathrm{LCL}_{\overline{x}\,\text{QP}}$]. The PC-RT control limits (the *inner* set of limits) is derived from the area-wide coordination control limits (the *outer* set of control limits) by a simple percentage argument, chosen by engineering judgment. [$\mathrm{UCL}_{\overline{x}\,\text{QP}}$, $\mathrm{LCL}_{\overline{x}\,\text{QP}}$] are defined from the variability statistics of the demand process, measured by the average range $\overline{R}$ of the detector samples, and the number of samples n in each detection interval.

When [$\mathrm{UCL}_{\overline{x}\,\text{PC-RT}}$, $\mathrm{LCL}_{\overline{x}\,\text{PC-RT}}$] is exceeded, PC-RT optimization is scheduled immediately to solve for pro-active metering rates that mitigate the flow fluctuation, as detailed in Chapter 6. When a *trend* of points exceeding [$\mathrm{UCL}_{\overline{x}\,\text{QP}}$, $\mathrm{LCL}_{\overline{x}\,\text{QP}}$] is detected, the area-wide coordination problem is re-solved with new estimates of the flow rates in each section according to the estimation formula $\hat{\overline{\overline{X}}}^t = \hat{\overline{\overline{X}}}^{t-1} \pm A_2\overline{R} + f(\overline{R},T)$. We assume any two consecutive points outside of [$\mathrm{UCL}_{\overline{x}\,\text{QP}}$, $\mathrm{LCL}_{\overline{x}\,\text{QP}}$], such that $T_{j,k}=2$ or $T_{r,i}=2$, to be a trend. Upon detecting a flow rate change requiring the area-wide coordination problem to be re-solved, the route-proportional matrix parameters $A_{i,j}$ are assumed to be re-estimated exactly. The next chapter describes the structure of MILOS in terms of software modules and details the algorithmic operation of MILOS at each time-step. Chapter 9 then describes a simulation experiment that evaluates the performance of MILOS versus other ramp metering methods.

Chapter 8: MILOS software implementation

Introduction

In the last three chapters, the major components of MILOS have been described in detail. In this chapter, we summarize the algorithmic operation of the MILOS control strategy as an integrated *system* of components and describe the software implementation of the control *layers*. MILOS is composed of four hierarchically embedded, interactive subsystems;

(1) Subnetwork identification ,
(2) SPC anomaly detection,
(3) Area-wide coordination, and
(4) PC-RT rate regulation ,

developed by decomposing the large-scale freeway control problem into subproblems of varying spatial and temporal influence. The subnetwork identification layer is not implemented in the software and is left as a topic for further development. Hence, we only consider a single area-wide coordination problem in this software implementation and the simulation experiments described in Chapter 9.

The software is implemented in the MATLAB programming language as a collection of modules, as per standard software engineering practice. The modules are arranged such that each module performs a limited, specific task in the MILOS hierarchy. Sub-modules are defined where appropriate to perform repetitive and/or detailed tasks within a given module. The procedural flow-chart of modules is depicted in Figure 8-1.

In each iteration of MILOS, the main steps are characterized by the cyclical operation of SPC anomaly detection, application of the pro-active rates obtained in the previous planning cycle, area-wide coordination (if needed), and PC-RT rate regulation optimization (if needed). Note that, as indicated in Figure 8-1, the PC-RT optimization module sends its rate tables to the parameter database and *not* directly to the macrosimulation. This reflects the *pro-active* nature of MILOS since the current rates solved from each PC-RT subproblem will not be utilized until the next optimization interval. In contrast, if the area-wide coordination problem needs to be re-solved, the new nominal rates $r_{i,N}$ are applied in

this time-step and PC-RT optimization is disabled for this optimization interval for this ramp subsystem.

In the macrosimulation model, the *simulation* time-step is typically on the order of a few seconds while the *optimization* time-step is typically on the order of one-minute. Thus, before the density, speed, demand, etc. measurements are passed back to the parameter database, several simulation time-steps would have been completed.

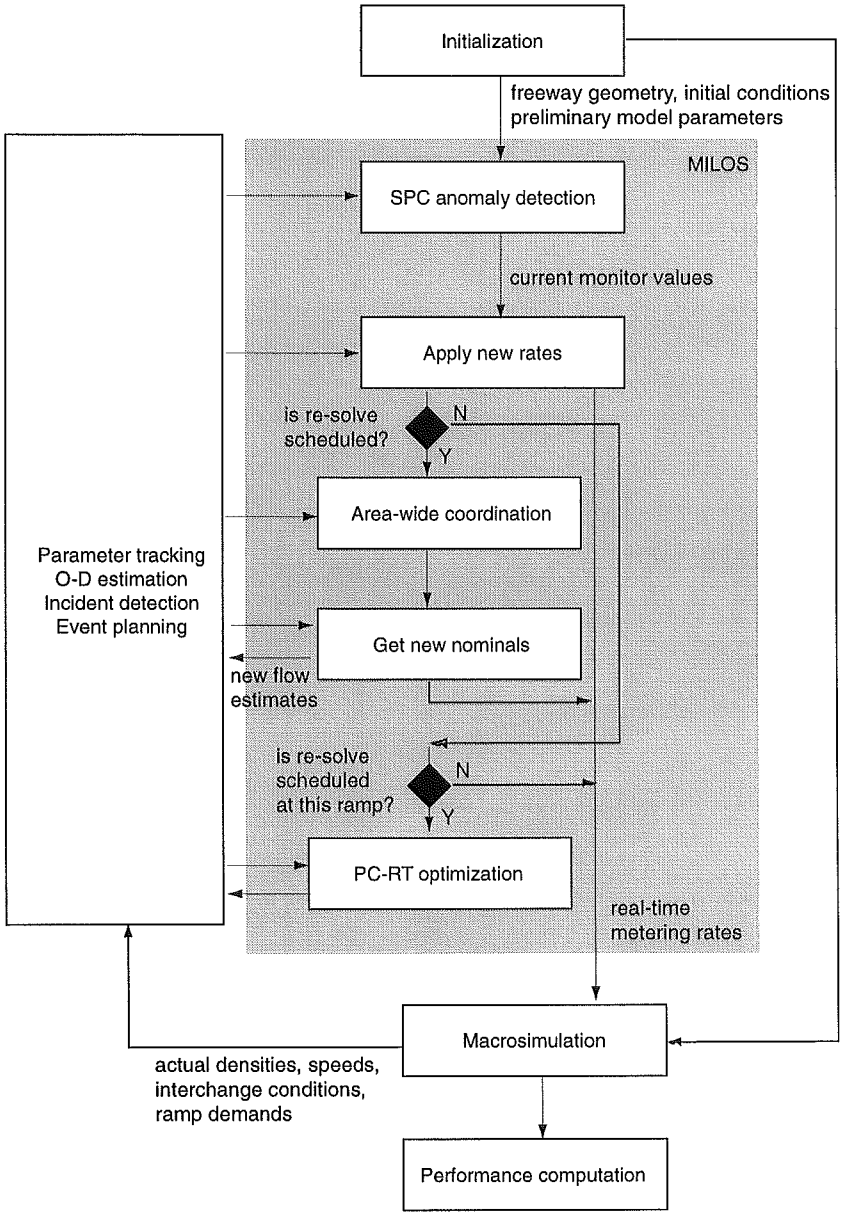


Figure 8- 1. MILOS operational flow chart

Initialization module

At system start-up, MILOS is initialized with the given freeway geometric information on link lengths, on-ramp and off-ramp locations, ramp queue capacities and initial lengths, etc. and the data structures for the network model are synthesized. The network model is based on the convention that the freeway is composed of *segments*. Each segment has exactly one change to the freeway geometry, such as an off-ramp, an on-ramp, both an off and on-ramp, a lane drop, or freeway connector [Papageorgiou, 1983]. Each *segment j* is composed of *k sections*, of approximately equal length. Each geometric change to the network (on-ramp, lane-drop, etc.) is assumed to occur in the *first* section of the segment. The standard section length is based on the modeling time-step and the minimum distance in the given freeway network between two adjacent geometric changes. The remainder of the initialization phase sets the initial estimates of time-varying parameters, initializes data structures, and "loads" the simulation with the initial freeway traffic states and ramp queues in each segment.

SPC-based anomaly detection module

The procedural operation of the SPC anomaly detection module is shown in Figure 8-2. In a given MILOS iteration, SPC is the "first" module to execute in the loop to identify locations where

(a) PC-RT optimization needs to be run, and

(b) if the area-wide coordination problem needs to be re-solved

in this iteration. Before checking for re-optimization intervals, if new approximately-constant flow estimates $V_{j,N}$ and/or $d_{i,N}$ were defined in the previous iteration of SPC, new [$UCL_{\bar{x}\,PC\text{-}RT}$, $LCL_{\bar{x}\,PC\text{-}RT}$] and [$UCL_{\bar{x}\,QP}$, $LCL_{\bar{x}\,QP}$] limits are defined for the new value of $V_{j,N}$ and/or $d_{i,N}$ in each segment. Otherwise, the module proceeds directly to the identification of the detector locations closest to being one-minute upstream of the ramp meter. Operationally, we track changes to $V_{j,N}$ for each section of each segment in the network, in case the travel time changes significantly. Of course, this flexibility would require numerous closely-spaced detector stations. Most of the time, however, the same locations will be used to measure $V_{j-B}(k)$, B sections one-minute upstream of each ramp meter. Recall from Chapter 6 that the monitoring location is approximately one-minute upstream of the ramp meter to provide the pro-active aspect of the PC-RT rate regulation optimization procedure.

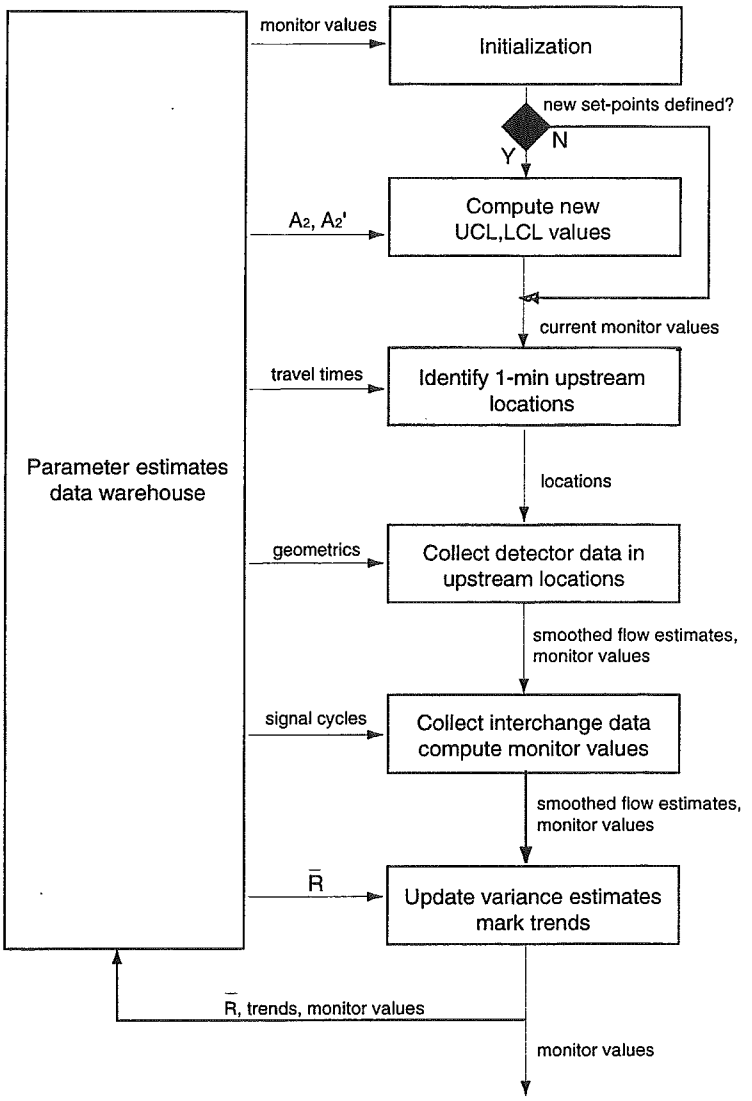


Figure 8- 2. SPC anomaly detection flow chart

Next, the appropriate flow data from the *interchange* detector placements is processed. A "processing" method to derive ramp demands from interchange flows was presented briefly in Chapter 4. However, this method does not remove the fluctuations of the second-by-second meter demand due to the traffic signal or address the estimation of turning probabilities at the interchange. Deriving a reliable method to obtain smoothed one-minute prediction of ramp demands is a topic for future work.

Both upstream freeway flow and ramp demand estimates are then compared against their respective [$UCL_{\bar{x}\,PC\text{-}RT}$, $LCL_{\bar{x}\,PC\text{-}RT}$] and [$UCL_{\bar{x}\,QP}$, $LCL_{\bar{x}\,QP}$] limits to signal the need to solve PC-RT optimization and/or area-wide coordination problems. A simple classification scheme is used to label each section of the control chart for each detector station

Applying new rates

After running SPC anomaly detection, the monitor values $M_{j,k}$ and $M_{r,i}$ ("1", "2", ..., "5") and the trend designations $T_{j,k}$ and $T_{r,i}$ (0, 1, 2) are passed to a routine that sets the current real-time rate $r_i(k)$. The predictions made for each PC-RT optimization scenario are based on the current (and possibly anomalous) flow rate at the upstream freeway segment and ramp demand. As such, the rate table for each ramp is valid only for that set of initial conditions.

Area-wide coordination module

If an out-of-specification trend has been detected, the area-wide coordination module is executed. The flow-chart structure of the module is illustrated in Figure 8-3. First, the appropriate data is collected from the database, including the new estimate of the route-proportional matrix A, new nominal ramp demand levels $d_{i,N}$, and segment capacities CAP_j. Recall that if a segment has an incident, the capacity CAP_j is periodically re-evaluated as the incident clears and area-wide coordination is re-scheduled. After the initial data are collected, the cost coefficients for weighting queues at each ramp are computed from the current interchange congestion conditions. Next, the available excess queue capacity Q_i is computed for the time horizon T for the current estimate of each queue length $q_i(0)$. Finally, the objective function and constraints of the quadratic programming problem are written to a file and solved with the CPLEX barrier optimization routine. CPLEX then exports the solution target variables $V_{j,N}$, control variables $r_{i,N}$, dual multipliers λ_k, and constraint slacks ε_k back to the database. The nominal rates $r_{i,N}$ are then immediately applied, bypassing any PC-RT optimizations scheduled.

If the area-wide coordination problem is re-solved in this iteration of MILOS, new nominal densities $\rho_{j,N}$ and speeds $v_{j,N}$ are derived from the nominal flow rates $V_{j,N}$. Recall that the

nominal speeds and densities are required to linearize the system dynamic equations for the definition of the PC-RT rate regulation subproblems. These nominal values $\rho_{j,N}$, $v_{j,N}$ are derived from the analytical volume-density and density-speed functions calibrated for each freeway section.

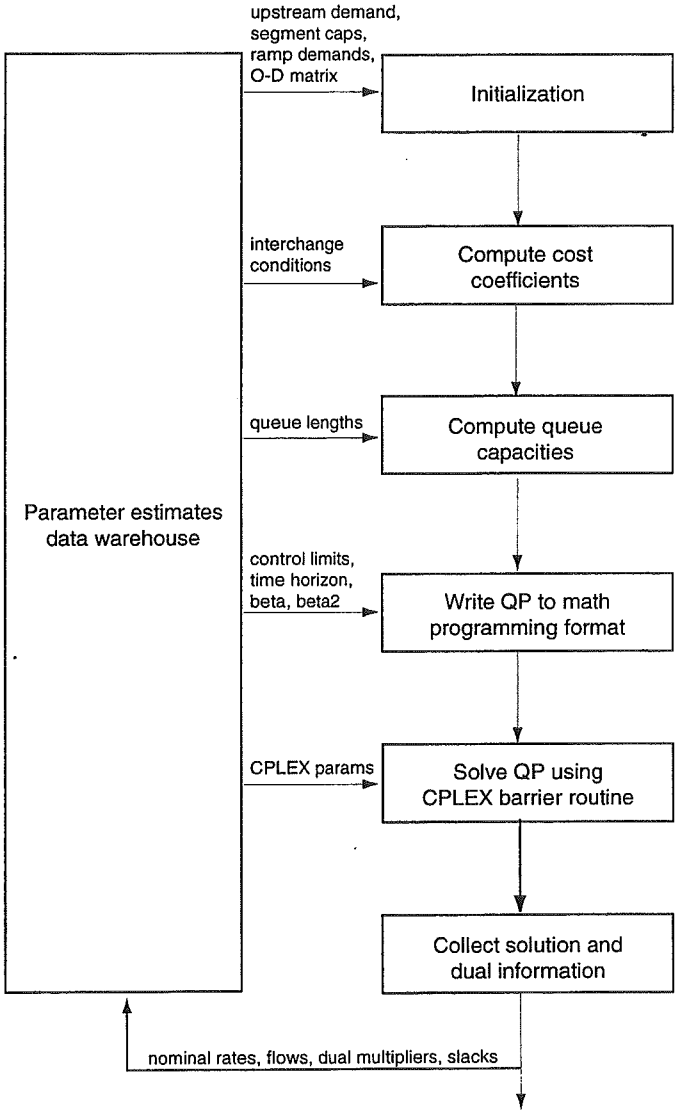


Figure 8- 3. Area-wide coordination flow chart

PC-RT optimization module

Recall from Figure 8-1 that if the area-wide coordination problem was re-solved in this iteration, any scheduled PC-RT optimizations are delayed. Hence, if area-wide

coordination was not executed, PC-RT optimizations are executed directly after the current rates are applied. The iterative operation of the PC-RT optimization module is illustrated in Figure 8-4. First, the monitor values $M_{r,i}$ and $M_{j,k}$ are collected from the database, indicating which ramp locations require PC-RT re-optimization. Any ramp subsystem that has an anomalous monitor value will be scheduled for PC-RT optimization. Next, the cost coefficients are computed from the dual variables and slack values from the area-wide coordination problem solution. Then, for each ramp that PC-RT optimization is scheduled for, the sub-network of the subproblem is defined to include enough sections to satisfy the one-minute of upstream travel time requirement and include the ramp section and one section downstream of the ramp.

After the subnetwork is defined, the predicted demand $V_{j\text{-}B}(k)$, $d_i(k)$ and queue growth $q_i(k)$ scenarios are created for each combination of "fundamental" flow predictions. Next, the freeway flow equations are linearized at the nominal point, or at the congested point if the section is congested, and the and the resulting linear difference equations are written to an LP format. The objective function, queue growth dynamic equations, and constraints are then written in the LP format, and the LP optimization problem is solved using CPLEX. If the problem is infeasible, "zero" is written for the rate table entry of that scenario, and if that scenario is realized at the next iteration, the nominal rate $r_{i,N}$ is applied. Typically, problems are only infeasible if the current state is already outside of [$\mathrm{UCL}_{\bar{x}\,\mathrm{QP}}$, $\mathrm{LCL}_{\bar{x}\,\mathrm{QP}}$] and the prediction is for increasing flow. In this case, it may be impossible for the local ramp meter, with limited ability to modify the rate, to contain the congestion from occurring. When an infeasible PC-RT scenario is realized, the area-wide coordination problem should be re-solved. If the PC-RT scenario is indeed feasible, CPLEX exports the rate modification for the first minute of the time-horizon to the corresponding location in the rate table.

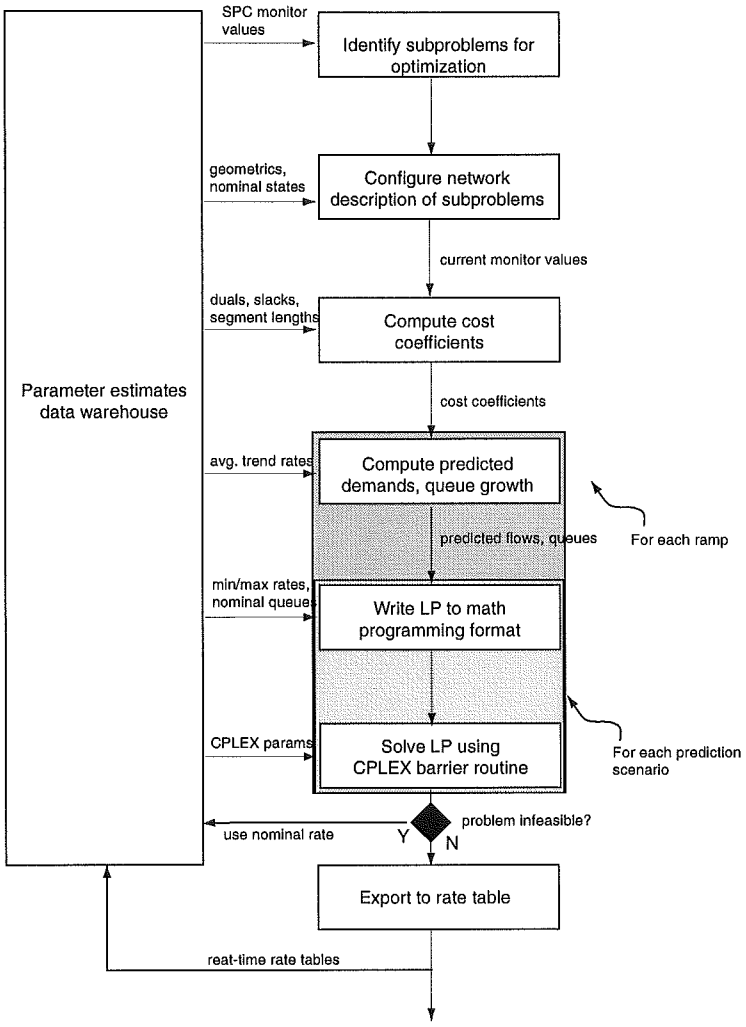


Figure 8- 4. PC-RT optimization flow chart

Example of MILOS operation

The typical operation of the MILOS algorithm is illustrated in Figures 8-5 and 8-6. Figure 8-5 illustrates the time-series of the upstream freeway flow and the ramp demand for a given subsystem. At the current time, the SPC module detects that the upstream freeway flow has gone below $LCL_{PC\text{-}RT}$ and the ramp demand has exceeded $UCL_{PC\text{-}RT}$. Thus, at this ramp location, PC-RT rate regulation re-optimization is scheduled. Note here that, although both conditions were exceeded in this example, only one condition needs to be exceeded to trigger PC-RT re-optimization.

The predicted future flow scenarios at the upstream freeway and ramp queue are shown as lines A_{fy}, B_{fy}, C_{fy}, and A_{rd}, B_{rd}, C_{rd}, respectively. For each of the nine combinations of upstream flow and ramp demand, an LP optimization problem is formulated and solved to obtain travel-time minimizing ramp metering rates for the next five minutes. As specified in Chapter 6, these rates $r_i(k)$, ..., $r_i(k+5)$, are not allowed to deviate too far from the nominal rate $r_{i,N}$ specified by the area-wide coordination problem.

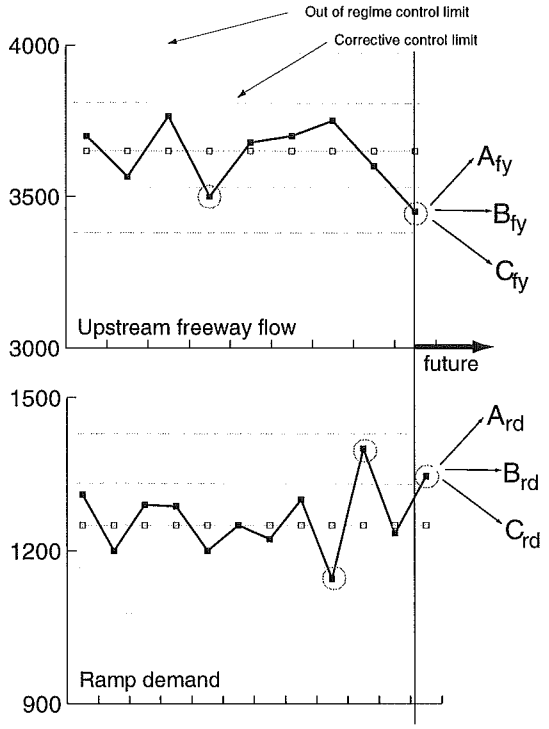


Figure 8- 5. MILOS operational example

Consider the case where, in the next minute, scenario C_{fy} and A_{rd} was realized, and thus the rate modification for "decreasing" freeway flow and "increasing" ramp demand was applied from row 1, column 3 of the rate table of the type of Table 8-3 for this ramp subsystem. This scenario (C_{fy} and A_{rd}) also results in a upstream freeway flow measurement that is less than LCL_{QP} and a ramp demand measurement greater than UCL_{QP}. Thus, if the next measurements are also outside of the corresponding outer limits (and thus a *trend* has been detected) the area-wide coordination is scheduled for an immediate re-solve. Figure 8-6 indicates the resulting situation three minutes after the re-solve of the area-wide coordination problem. Here, both the upstream freeway flow and ramp demand are again

exceeding the newly-defined PC-RT control limits and thus at this point, PC-RT re-optimization is scheduled. The process of SPC anomaly detection, area-wide coordination, and PC-RT rate regulation optimization repeats in this manner continually throughout the day.

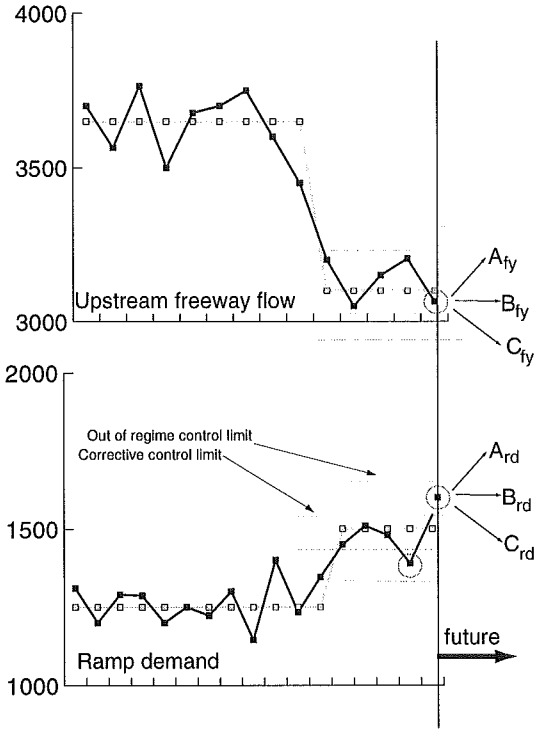


Figure 8- 6. MILOS operational example, continued

Summary

The software implementing the MILOS control strategy is an integrated system of modules. Each software module performs a limited, specific task in the MILOS hierarchy. Sub-modules are defined where appropriate to perform repetitive, detailed tasks within a given module. The main components of the software represent the major algorithmic layers of the MILOS strategy, namely:

(a) the PC-RT rate regulation optimization module,

(b) the area-wide coordination optimization module, and

(c) the SPC anomaly detection and optimization scheduling module.

The main operation of MILOS is characterized by the cycle of SPC anomaly detection, application of the pro-active rates optimized during the previous planning cycle, area-wide coordination (if needed), and, PC-RT optimization when the area-wide coordination

problem was *not* re-solved. Route-proportional matrix A estimation and ramp demand flow smoothing $r_i(k)$, as well as other parameter estimation tasks, are performed external to MILOS. An important characteristic of MILOS is that the software implementation can be executed in real-time to calculate analytically-based ramp metering rates. The next chapter details the evaluation of the MILOS ramp metering control algorithm in a simulation experiment on a small freeway in the Phoenix, AZ metropolitan area. The simulation experiment compares the operation of MILOS versus LP optimization re-solved at fixed intervals, locally traffic-responsive metering under evaluation at ADOT, and the no metering case for a typical operational rush-hour scenario.

Chapter 9: Simulation Experiments

Introduction

In Chapter 5, we briefly showed that the quadratic programming formulation of the area-wide coordination problem can improve upon linear programming approaches. The main performance benefits at the area-wide layer are realized in the ramp metering coordination problem by distributing queue growth to the appropriate ramp queues, where possible. These results were presented for a single *deterministic* simulation of the macroscopic equations (Chapter 4) for a 30km test network resembling the test problem used by Papageorgiou [1983]. In a deterministic simulation, application of the full MILOS hierarchical control system would show little performance improvement over more "traditional" methods since it is designed to take advantage of the stochastic component of freeway flow measurements and ramp demands. Hence, a more detailed, realistic set of simulation experiments with stochastic demand flows is constructed to evaluate the performance of MILOS versus several alternative ramp metering methods.

Structure of the simulation experiment

Our intent in this simulation experiment is to demonstrate the two-level MILOS scheme by comparing its performance to several metering algorithms on a realistic test network in the Phoenix, AZ metropolitan area. The ramp metering algorithms that will be compared with include:

(1) No ramp metering,

(2) Traffic-responsive volume/speed metering with queue management,

(3) LP metering using no weighting coefficients but with queue length constraints, re-solved at 5-minute intervals based on current demand rates, and

(4) MILOS two-level optimization method.

Method (2), as shown in Table 9-1, is being considered by ADOT as a traffic-responsive ramp metering policy. This table relates the metering rate at a given ramp to the freeway

volume or speed just upstream of the metering location. Functionally, after the volume and speed are collected from the upstream detectors, the values are compared with Table 9-1 beginning from the top and proceeding to the bottom. If the mainline measured volume is *less than* the given the threshold in column 2 *or* if the mainline measured speed is *greater than* the threshold given in column 3, the corresponding metering rate is applied. The mainline speed and volume measurements are collected and the metering rate is adjusted each minute. In this way, the policy is traffic-responsive.

Metering rate (veh/hr)	Mainline volume threshold (veh/hr/lane)	Mainline speed threshold (miles/hr)
900	480	60
720	720	57
600	1080	54
480	1560	46
360	1860	30
240	1980	10

Table 9- 1. Traffic-responsive ramp metering rates and thresholds

In any method, unrealistic queue lengths will grow at the ramp approaches since no diversion behavior is included in the macroscopic model. If the queue maximum length of a ramp approach is exceeded, method (2) will neglect the measured upstream volume and speed and enact the highest possible metering rate to attempt to flush the queue. This highest possible rate is set at 1450 veh/hr/lane. When the queue length again drops below the maximum length, Table 9-1 is used to derive the metering rate. In both LP and MILOS metering, the maximum metering rate is set to 1450 veh/hr.

The simulation experiment is divided into three parts:

(1) a test case with 20 minutes of medium-volume traffic, 20 minutes of high-volume traffic, and finally 40 minutes of a medium-volume traffic that may or may not need to be metered heavily, depending on the metering decisions made during the previous 20 minutes.

(2) a test case for a typical 3-hour rush-hour peak period, where during the second hour of operation, flow breaks down considerably if no metering is done. In this test case, the input flow rates and route proportional matrices change in 20 minute intervals.

(3) a test case for a typical 3-hour rush-hour peak period with an incident, of 30-minutes duration, occurring after the first hour of the simulation.

In the first test case, the time of the demand and route-proportional matrix changes will be known to all algorithms (although some algorithms do not use this information). Thus, the LP and QP area-wide coordination algorithms will know exactly when to solve the LP/QP the first time. Route-proportional matrices for LP/QP problems are assumed known perfectly. In the LP case, the area-wide coordination problem is re-solved in 5-minute increments using the most recent flow and speed data. If the freeway becomes overcapacity, both LP and MILOS are allowed to re-solve with restricted capacity in the congested section(s). Thus, both LP and MILOS use (perfect) information from an incident detection algorithm to schedule re-optimization with restricted freeway capacities.

In the second and third test-cases, MILOS will be given a new, perfectly-estimated route-proportional matrix and the value of the new target demand at the beginning of each transition. Thus, MILOS must use the SPC-based anomaly detection scheme to identify the scheduling of the PC-RT optimizations but not the transition to a new demand level. If, during the course of the simulation run, a new target demand level is identified (erroneously, or due to upstream congestion), MILOS will modify the target demand according to the SPC-based anomaly detection scheme with $f(\overline{R},T)=0$. However, when the demands change at the transition points, an external processor will provide these demands perfectly. Further testing of the full capability of the SPC algorithm is a topic of future work.

In each test, 5 iterations will be run for each metering algorithm with identical initial conditions. For each iteration, the same random number seeds will be used for each metering algorithm, so that any bias induced by an "unfair" draw of random numbers for one run of a particular metering algorithm is eliminated. The average value and standard deviation of several performance indices will be computed and a comparison of the resulting distributions will be presented. At this first-cut level of analysis, each performance distribution will be assumed to be normal for plotting purposes, although this may not be the case. In addition, plots of the freeway density, ramp queues, metering rates, and total vehicles in the system will be presented for the same simulation run of each algorithm for each of the three test cases.

Calibration of macroscopic model to SR202 CORSIM output

Before the test cases can be executed in the macroscopic simulation environment of Chapter 4, the macroscopic model should be calibrated to the real-world location that the model is assumed to represent. In the absence of actual freeway and on-ramp detector data from the street, a CORSIM simulation model of the test location, State Route 202, in Phoenix, AZ, was obtained from Catalina Engineering for the evaluation test. Thus, the simulation experiment was conducted using the *macroscopic* simulation calibrated to reflect the performance of the *microscopic* simulation for a specific scenario. The SR202 simulation model includes 7 miles of freeway, 4 controllable on-ramps and 1 freeway-freeway on-ramp, and 7 off-ramps with the configuration as illustrated in Figure 9-1. The freeway is divided into 11 sections of the following lengths in feet [2600, 2000, 1500, 2300, 5200, 4200, 3000, 2150, 5400, 2500, 1500] and lanes [5, 2, 2, 3, 3, 3, 2, 3, 3, 3, 3]. The speed limit was identical in each segment, with a free-flow speed of 104 km/hr and maximum density of 110 veh/km-lane. The number of ramp lanes for each segment was [5, 0, 0, 2, 1, 1, 0, 1, 0, 1, 0] and the queue capacity of each on ramp was assumed to be [0, 0, 0, 80, 50, 50, 0, 60, 0, 40, 0] thus a segment with zero ramp lanes does not have an

on-ramp. The off-ramps were each one-lane and located in the following segments [0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0].

The performance of the macroscopic and microscopic models at locations A-E will be used for calibration. Note that between locations C and D a bottleneck occurs in the model, where a merge area and a lane drop occur in the segment indicated by "CORSIM bottleneck" in Figure 9-1.

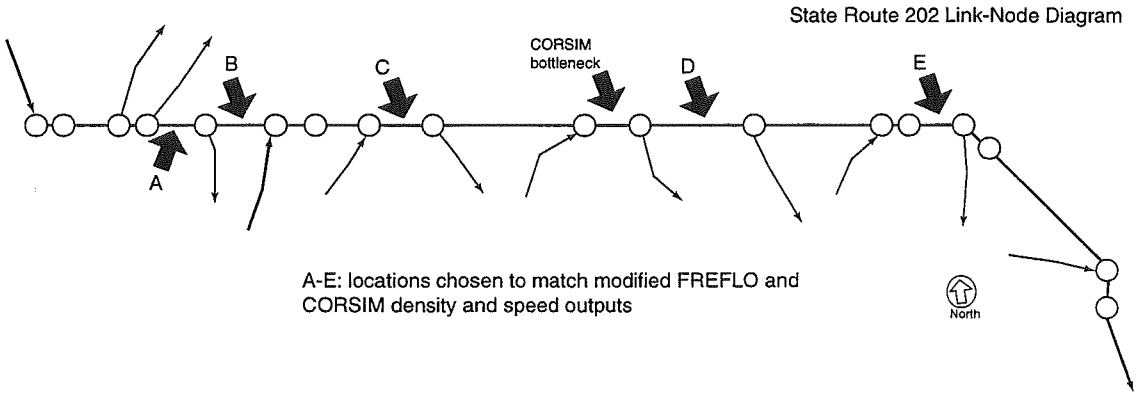


Figure 9- 1. State Route 202 CORSIM link-node diagram

The scenario used to calibrate the performance of the macroscopic model to the CORSIM performance was similar to test case (1) of the simulation experiment. In this calibration test, three levels of input volumes and route-proportional matrices were used to produce a "square-wave of traffic in the middle 30-minutes of the simulation that requires ramp metering. Hence, in the middle section of demand, some congestion should be observed in the absence of ramp metering. This is necessary to calibrate the macroscopic model to the full range of traffic situations. The route-proportional matrices used in the calibration test are shown in Table 9-2 and the input volumes are listed in Table 9-3. Table 9-4 indicates the initial conditions resulting after "equilibrium" was obtained in the CORSIM simulation. Hence, when the macroscopic simulation was executed, the identical initial conditions were used. The sections with '*' beside the density values indicate segments

which had significantly different initial values in one or more sub-sections of the segment.

from \ to	beg.	1	2	3	4	5	6	7	8	9
begin	1	0.69	0.614	0.614	0.52	0.475	0.304	0.304	0.267	0.267
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0.81	0.725	0.448	0.448	0.392	0.392
4	0	0	0	0	0.75	0.688	0.405	0.405	0.354	0.354
5	0	0	0	0	0	0.819	0.446	0.446	0.385	0.385
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0.802	0.802
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
from \ to	**beg.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
begin	1	0.71	0.66	0.66	0.59	0.535	0.417	0.417	0.375	0.375
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0.86	0.767	0.586	0.586	0.525	0.525
4	0	0	0	0	0.82	0.72	0.543	0.543	0.486	0.486
5	0	0	0	0	0	0.807	0.569	0.569	0.505	0.505
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0.833	0.833
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1
from \ to	**beg.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
begin	1	0.72	0.677	0.677	0.62	0.548	0.4	0.4	0.36	0.36
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0.88	0.771	0.548	0.548	0.491	0.491
4	0	0	0	0	0.83	0.727	0.51	0.51	0.456	0.456
5	0	0	0	0	0	0.796	0.513	0.513	0.455	0.455
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0.833	0.833
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1

Table 9- 2. Route proportional matrices

Time period / input stream	External	Ramp 1	Ramp 2	Ramp 3	Ramp 4	Ramp 5
0 - 20 minutes	4644	2268	463	600	552	480
21 - 40 minutes	5232	2676	540	480	396	708
41 - 60 minutes	5436	2184	540	588	660	732

Table 9- 3. Input volumes

state variable / section	sect 1	sect2	sect 3	sect 4	sect 5	sect 6	sect 7	sect 8	sect 9	sect 10
density (veh/km)	108.5*	53.6	48.4	70.2*	66.6*	86.1*	40.8	51*	35.4	38.1
speed (km/hr)	43	59.6	61.3	59.5	60.7	54.1	60.4	59.3	62	57.6
speed limit (km/hr)	104	104	104	104	104	104	104	104	104	104
link length (m)	2600	2000	1500	2300	5200	4200	3000	2150	5400	2500
number of lanes	5	2	2	3	3	3	2	3	3	3
on-ramp?	external			2 lanes	1 lane	1 lane		1 lane		1 lane

Table 9- 4. Initial conditions and parameter values for State Route 202

Data was collected for the observation locations A-E as indicated in Figure 9-1 for one-minute intervals throughout the one-hour simulation. Figure 9-2 indicates the match between the macroscopic flow model and the exponentially-smoothed CORSIM output volume and speed such that

$$F(t) = \alpha x(t) + (1-\alpha)F(t-1); \quad F(0) = x(0) \qquad \textbf{Eqn. 9- 1}$$

with $\alpha = 0.7$ at each of the observation locations A-E, from top to bottom, respectively. In eqn. 9-1, *F(t)* is the value (either volume or speed) plotted for the CORSIM output in Figure 9-2 at time *t* and *x(t)* is the observed average volume or speed at minute *t* of the simulation. The left column of Figure 9-2 is density (veh/km-lane) and the right column is speed (km/hr). The smooth curves are the output from the macroscopic model for the closest parameter match of the macro-model to the CORSIM data and the more variable curves are the output from the CORSIM simulation. Note here that in section D the macroscopic model predicts lower density and higher speed after t=40 minutes, when the CORSIM model becomes congested in that segment. No set of macroscopic simulation parameters could be found to reproduce this phenomenon using deterministic input rates in the macroscopic simulation.

Some questions still exist whether or not the CORSIM behavior for freeway modeling should be considered "realistic", and hence, further calibration of the macroscopic model to CORSIM was not considered. However, making the input demands normally-

distributed random variables with mean values as given in Table 9-3, (and using the same parameters from the calibration test shown in Figure 9-2), the simulation results shown in Figure 9-3 were obtained. Note that in Figure 9-3, the behavior of the macroscopic model in section D now more closely tracks the CORSIM output.

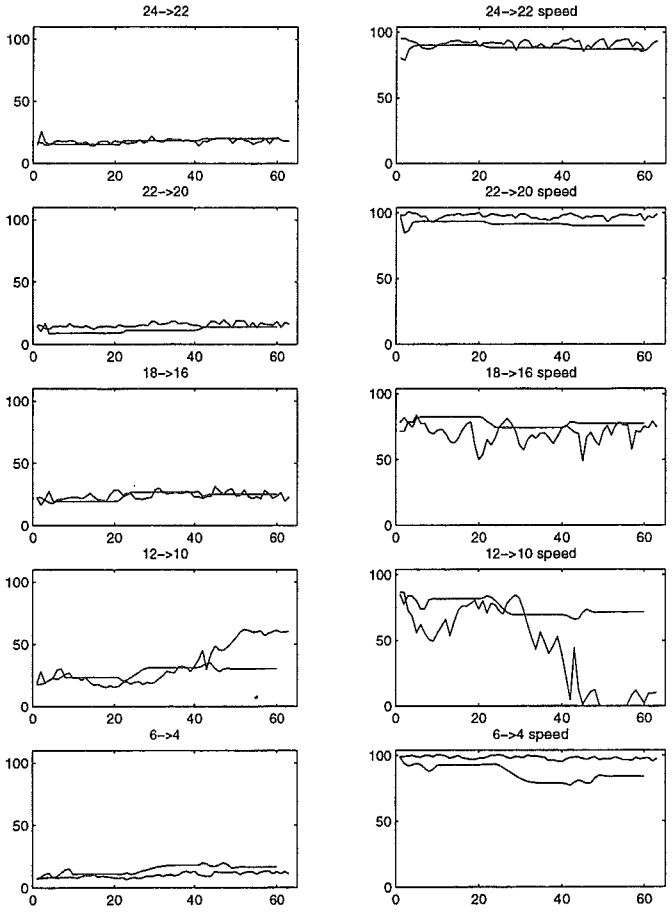


Figure 9- 2. Comparison of density and speed measurements

The performance of the two simulations were judged to be close enough so that the macroscopic flow model could be used to evaluate ramp metering strategies, and reasonable conclusions could be drawn about the resulting performance of those systems in a CORSIM simulation of the SR202. The resulting calibrated parameter table for the

macroscopic model is given in Table 9-4. Calibration of the macroscopic model to real data for SR202 (and/or other real freeway systems) is a topic of future work.

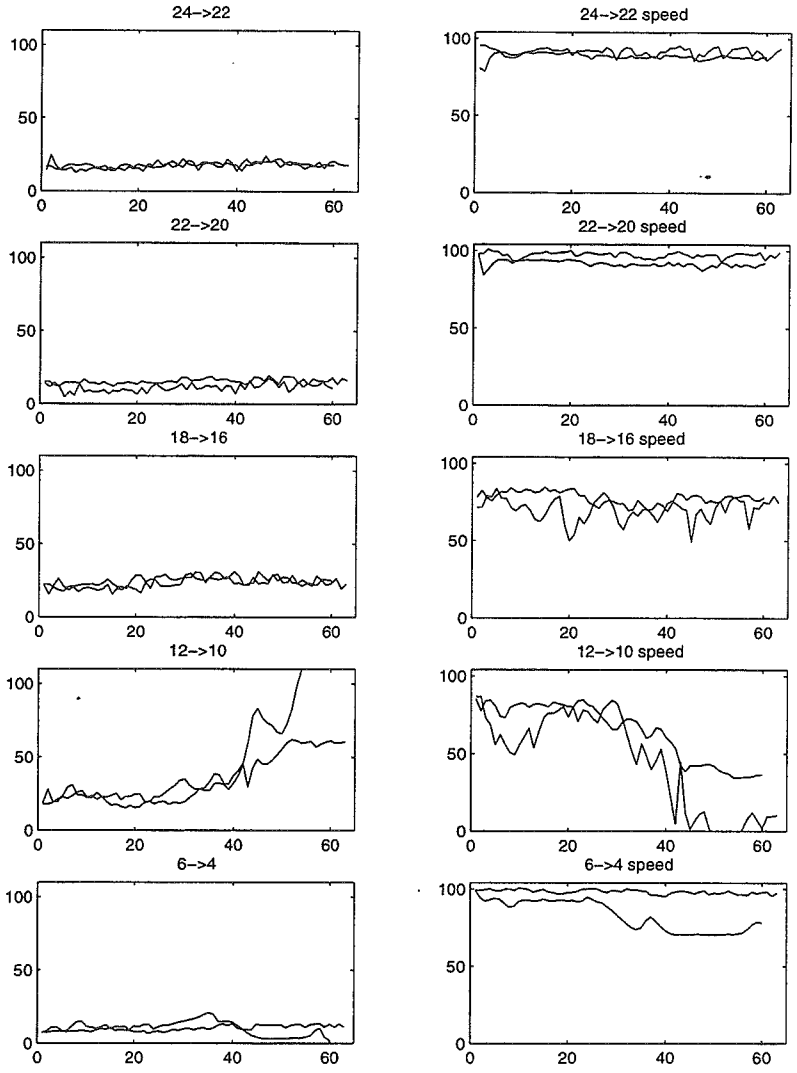


Figure 9- 3. SR202 comparisons, with stochastic input flows

0.01	0.75	10	3	18	0.0014 hr	0.65	110	104	0.9
τ	ν	κ	l	m	T	α	ρ_{max}	v_{max}	ρ_c

Table 9- 5. SR202 macroscopic simulation parameters

Test case #1

In the first test case, a 120-minute simulation was run with 20 minutes of medium-volume traffic, 20 minutes of high-volume traffic needing to be metered to maintain stable freeway flow, and finally 40 minutes of a second combination of medium-volume that may *or may not* need to be metered heavily, depending on the metering decisions made during the previous 20 minutes. After the first 80-minutes of realistic flows, the volume was reduced significantly to allow, if needed, the system to return to a "steady-state" condition such that the ramp queues return to essentially zero and the freeway density, throughout the corridor, returns to a stable state. The initial conditions for this, and all other test cases, were the same as given in Table 9-4. Table 9-6 lists the average input demands at each ramp approach during the three time periods. In each test case, the variance of the demand distribution was proportional to the number of lanes, such that

$$D_i(t) = \mu_i + 75\xi_i N(0,1) \qquad \textbf{Eqn. 9- 1}$$

Where μ_i is the average demand of the i^{th} ramp listed in Table 9-6, ξ_i is the number of lanes on the ramp, and $N(0,1)$ is the value of a random number drawn from a normal distribution with mean 0 and variance 1. In this, and all following test cases, the demand level changes each 20-seconds and the simulation step is set to approximately 5-seconds. The route-proportional matrices used in each 20-minute segment of this test case are listed in Appendix A.

The congestion levels at each interchange were set to 0.7, 0.8, 1, 0.4, and 0.75 for ramps 1, 2, 3, 4, and 5, respectively, for this and all other simulation test cases. The congestion level was not changed during the simulations, nor was the congestion level affected by the number of vehicles in the queue. Modifying the congestion level based on the number of vehicles in the queue, and to simulate the progression of traffic at each intersection is an important topic of future work.

When MILOS was run, β was set to 10 and β_2 was set to 1000β. The time-horizon for upper-layer optimization runs was set to 20 minutes and the time-horizon for PC-RT rate regulation optimization problems was 5 minutes (5 metering rates changes of 1-minute duration each). The capacity of each segment was computed based on the calibrated density-speed relationship.

Time period / input stream	External	Ramp 1	Ramp 2	Ramp 3	Ramp 4	Ramp 5
0 - 20 minutes	3644	1868	432	600	552	480
21 - 40 minutes	4832	2676	540	480	396	708
41 - 80 minutes	3836	2184	540	588	660	632
81 – 100 minutes	3036	1884	340	388	460	432
101 - 140 minutes	1000	500	300	300	400	400

Table 9- 6. Mean input rates, test case #1

Results for test case #1

Table 9-7 lists the average and standard deviation of the performance indices total travel time (TTT, vehicle-hours), total queue time (QT, vehicle-hours), corridor average speed (AS, km/hr), recovery time, (RT, hrs), maximum total queues (MQV), and maximum total vehicles in the system (MIS). Recovery time was computed as the time when all segments of the freeway return to a density below their respective capacities and all queues are reduced to less than 5 vehicles. Maximum total queues and maximum total vehicles in the system for each algorithm correspond to the simulation iterations displayed in Figures 9-19 through 9-22 of the total vehicles in the system and are not averages over all 5 iterations. Figures 9-4 through 9-7 depict the comparison of the performance distributions of each algorithm in total travel time, queue time, average speed, and recovery time, respectively. Each of the distributions is scaled to [0,1] for comparison purposes (so that each distribution has equal height).

Figures 9-8 through 9-11 depict the comparison of the freeway density of a single representative simulation (for the same initial conditions and random number streams) for each of the control methods. Figures 9-12 through 9-14 depict the comparison of the resulting queues that develop at each ramp during the same simulation. Note that in the "no control" case, no appreciable queues are built up over the simulation time, and thus no graph is included for this situation. Figures 9-15 through 9-18 depict the metering rates applied by each algorithm, where Figure 9-15 depicts the "no control" case, and thus represents the underlying demand process at each of the five on-ramps. Finally, Figures 9-19 through 9-22 depict the total vehicles in the system and total vehicles on the freeway for each of the algorithms. The upper portion of each figure displays the underlying total demand to the system, for comparison purposes. In Figures 9-19 through 9-22, the space between the two curves (if two curves are displayed) represents the total number of vehicles in ramp queues at any time instant. Note that, for comparison purposes, the scale of each graph is not identical.

Method	Avg. TTT	Std. Dev. TTT	Avg. QT	Std. Dev. QT	Avg. AS	Std. Dev AS	Avg. RT	Std. Dev. RT	MQV	MIS
No control	758.2	25.9	0.02	0.004	87.2	1.0	1.23	0.05	3	744
TR, w/QM	707.9	16.5	158.9	2.4	89.1	0.64	1.42	0.03	194	715
LP	698.5	7.7	220.6	15.1	89.6	0.30	1.31	0.16	330	802
MILOS	690.5	6.5	175.0	11.0	90.3	0.25	1.15	0.06	280	781

Table 9- 7. Performance results of test case #1

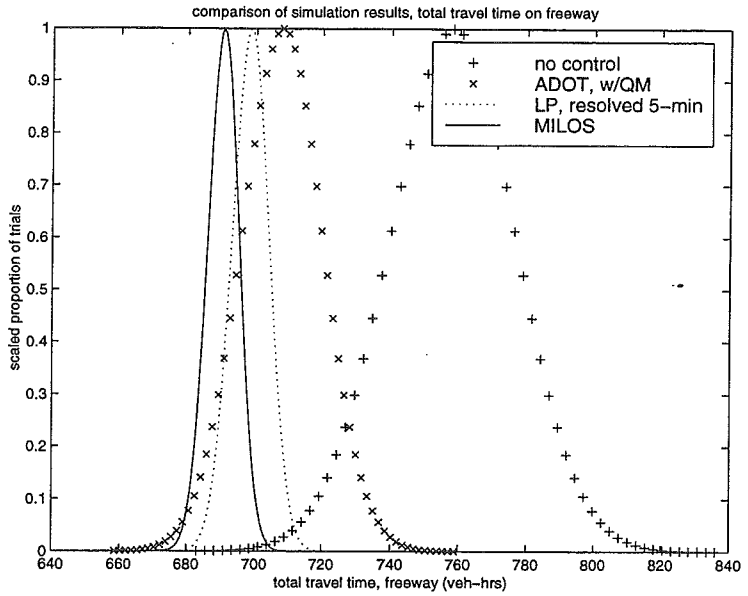


Figure 9- 4. Comparison of freeway travel time distributions

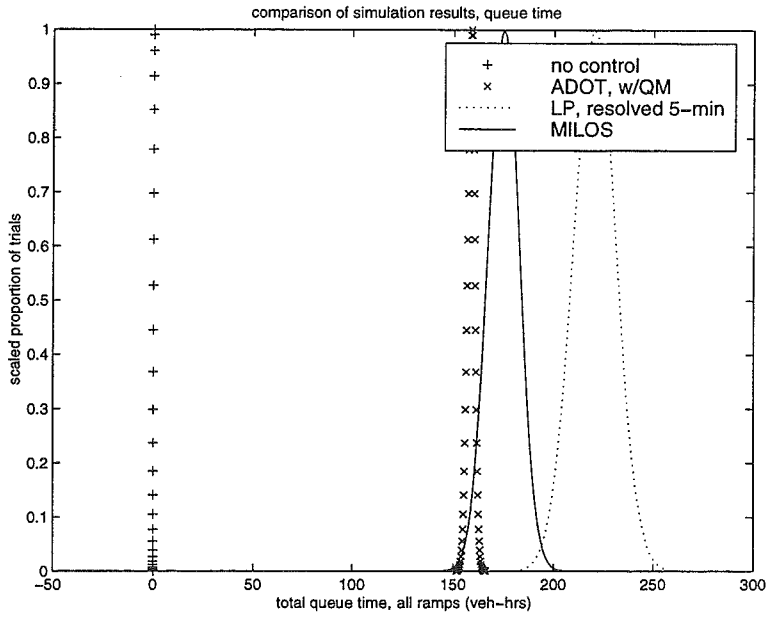


Figure 9- 5. Comparison of queue time distributions

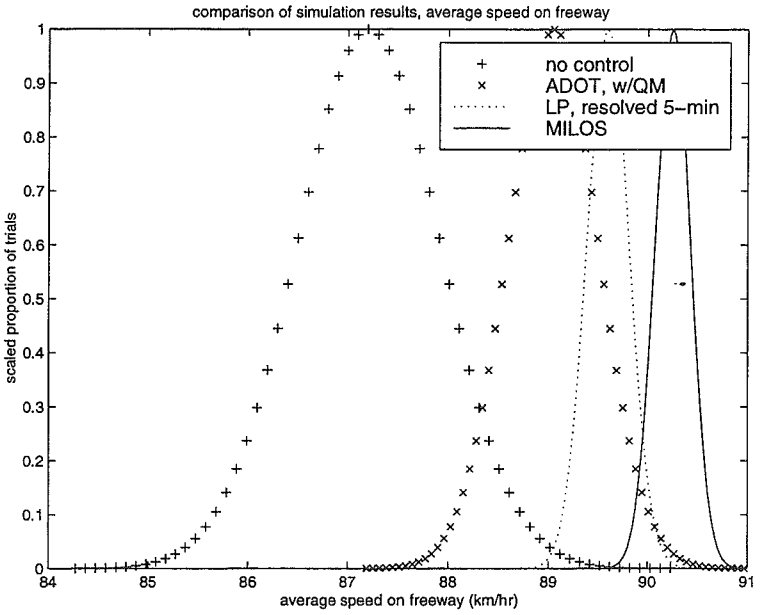


Figure 9- 6. Comparison of average speed distributions

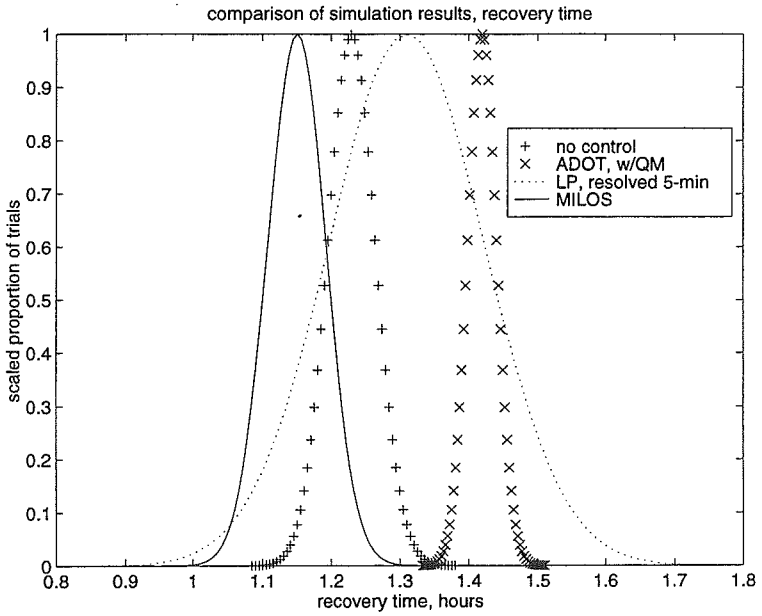


Figure 9- 7. Comparison of recovery time distributions

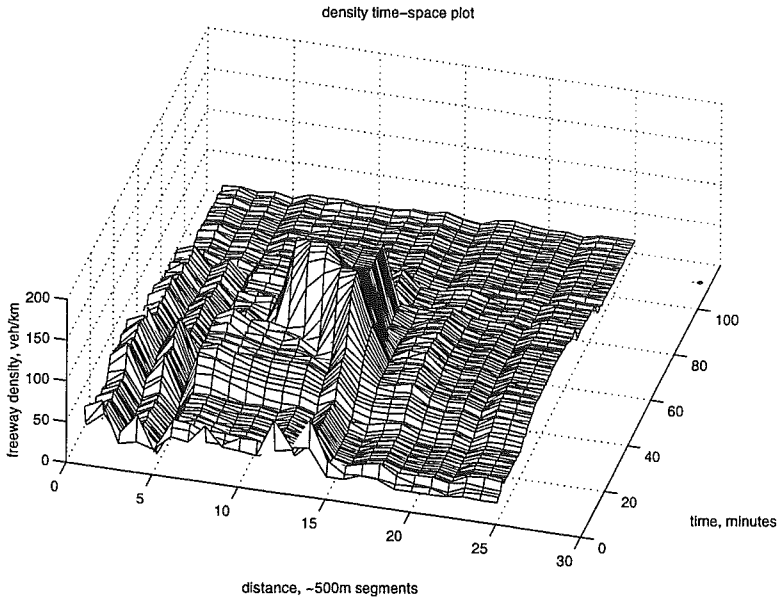


Figure 9- 8. Comparison of densities: no control

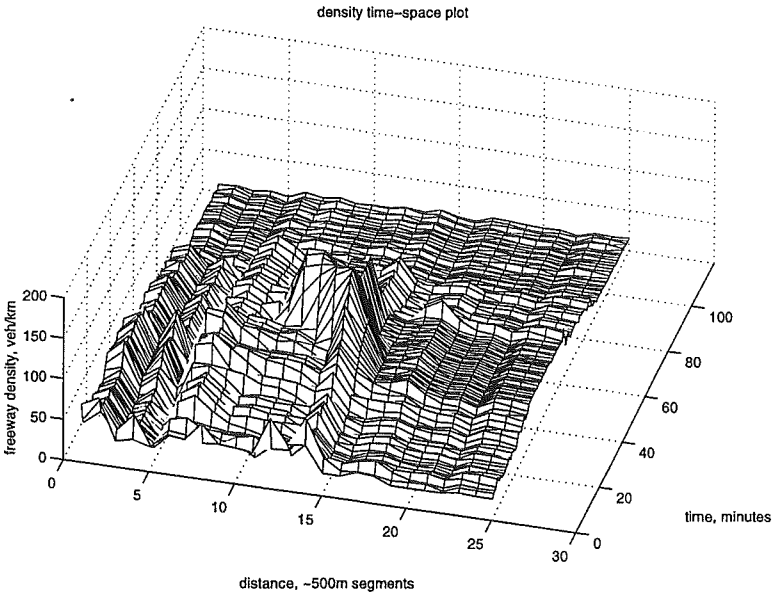


Figure 9- 9. Comparison of densities: TR, w/QM

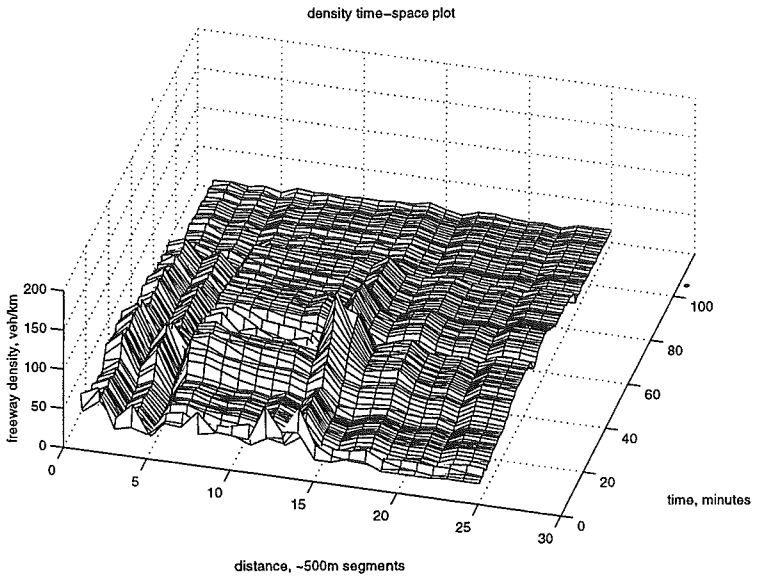


Figure 9- 10. Comparison of densities: LP, resolved each 5-minutes

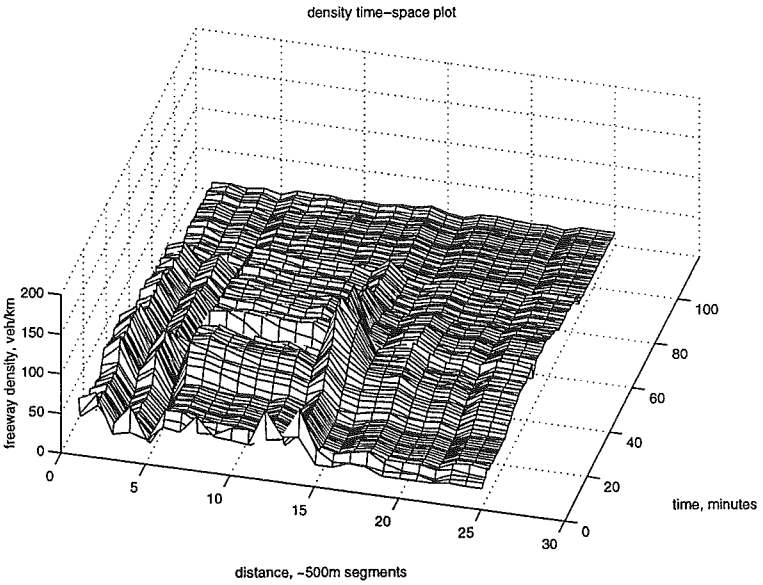


Figure 9- 11. Comparison of densities: MILOS

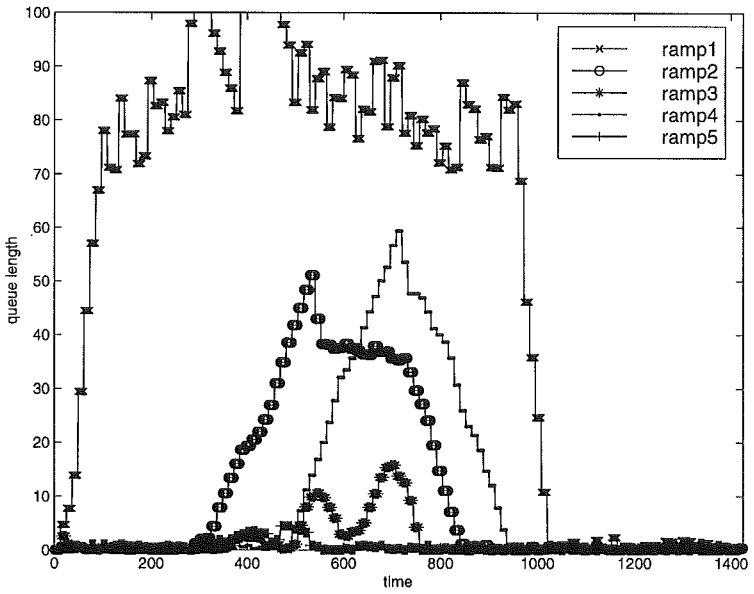


Figure 9- 12. Comparison of queue growth: TR w/QM

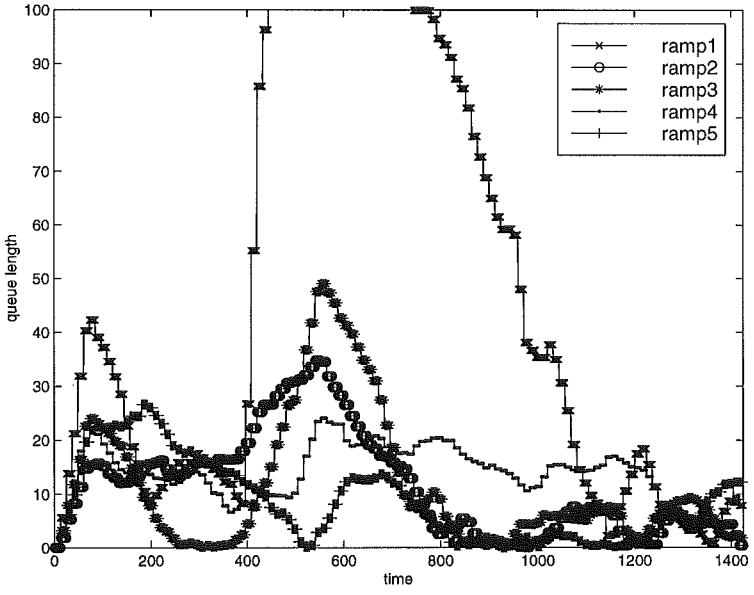


Figure 9- 13. Comparison of queue growth: LP, resolved each 5-minutes

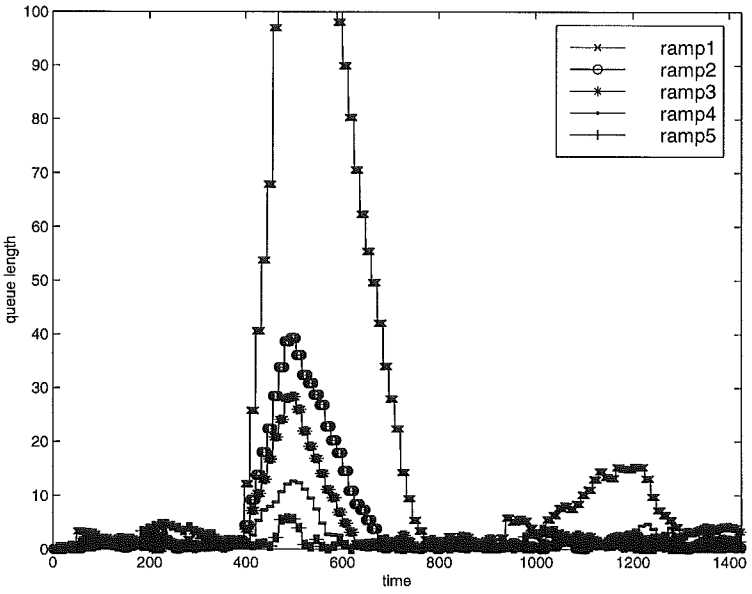


Figure 9- 14. Comparison of queue growth: MILOS

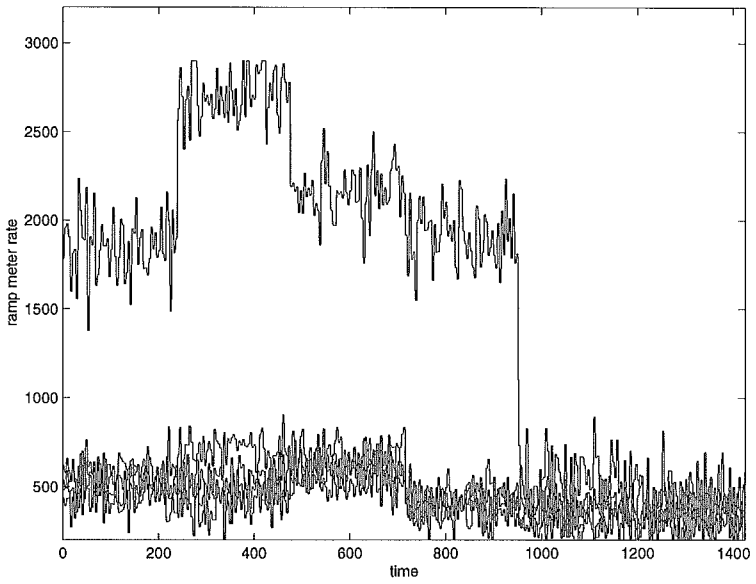


Figure 9- 15. Comparison of meter rates: no control (demands)

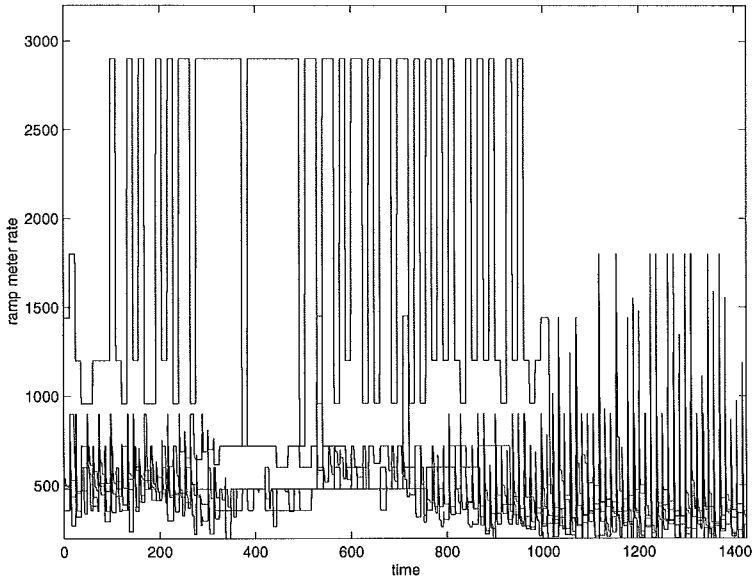


Figure 9- 16. Comparison of meter rates: TR w/QM

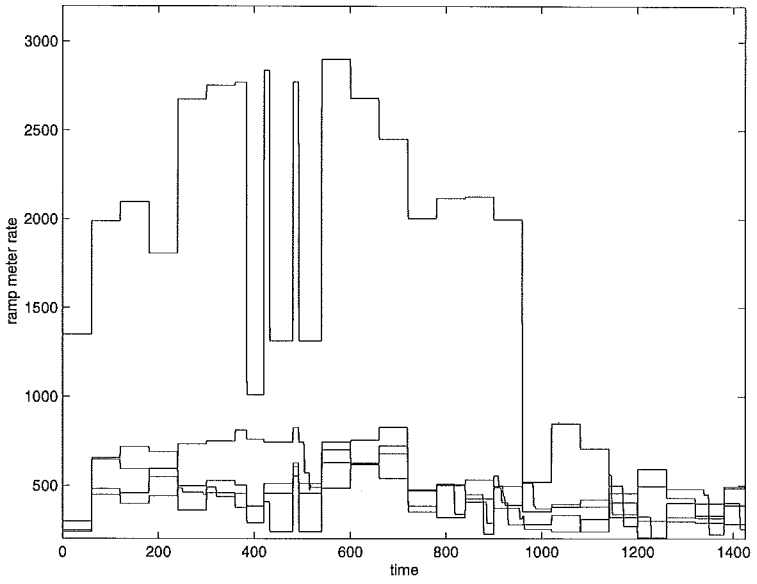


Figure 9- 17. Comparison of meter rates: LP, resolved each 5-minutes

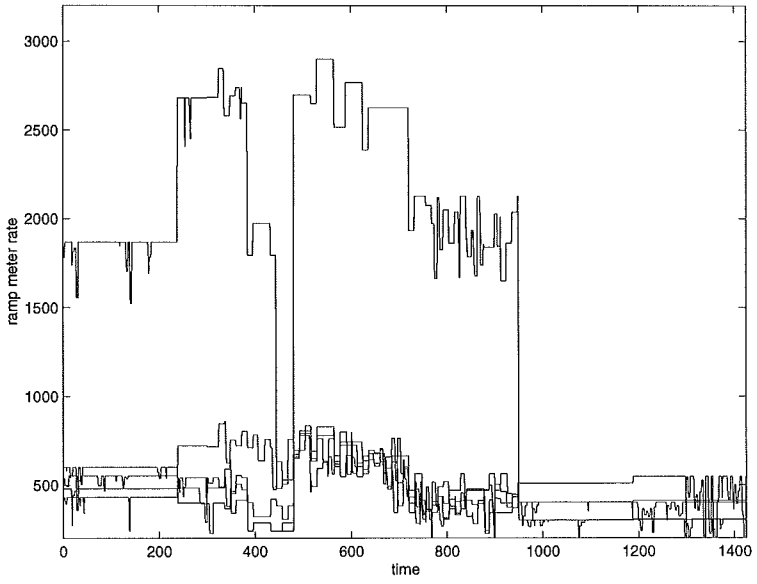


Figure 9- 18. Comparison of meter rates: MILOS

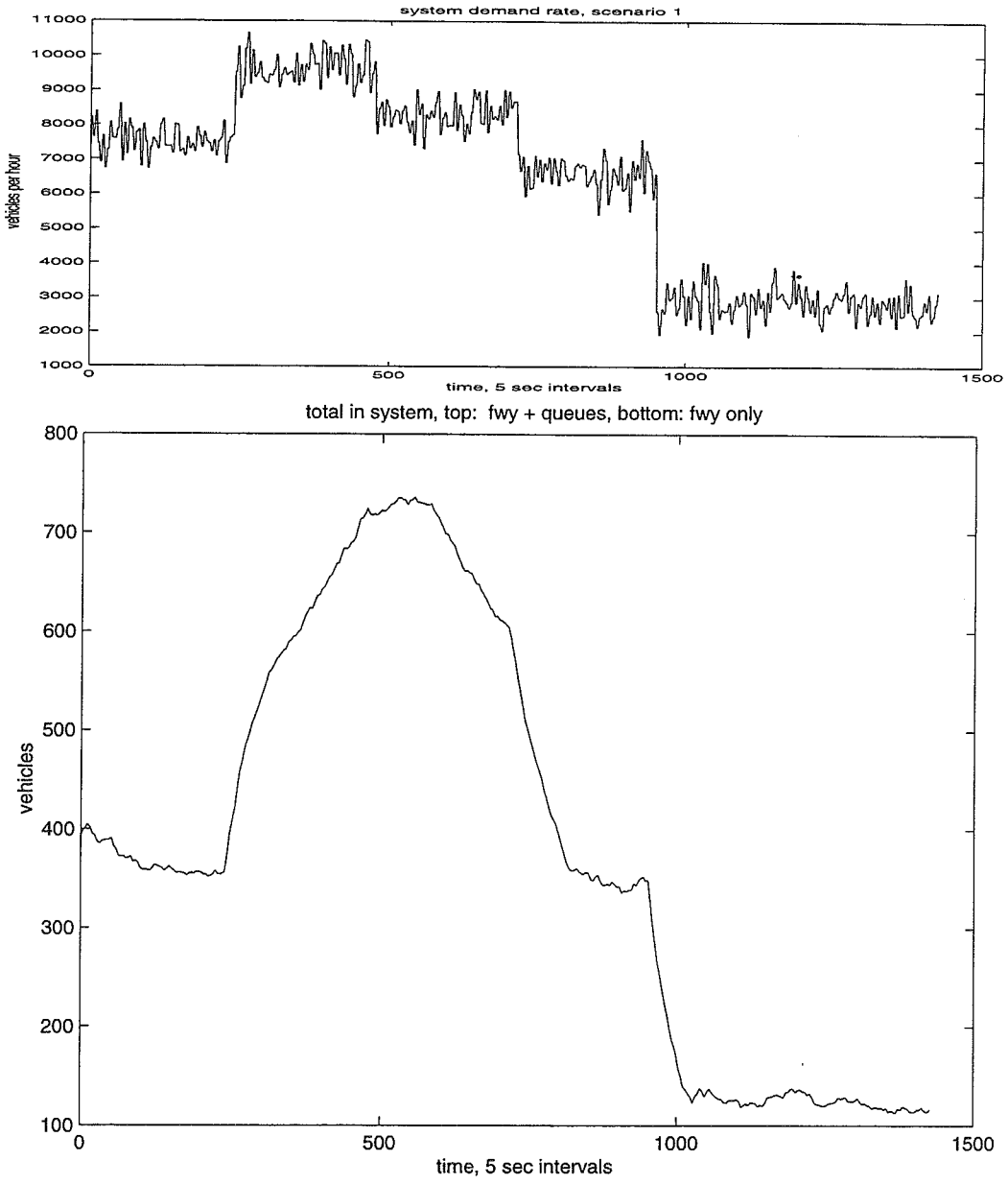


Figure 9- 19. Total vehicles in system: No control

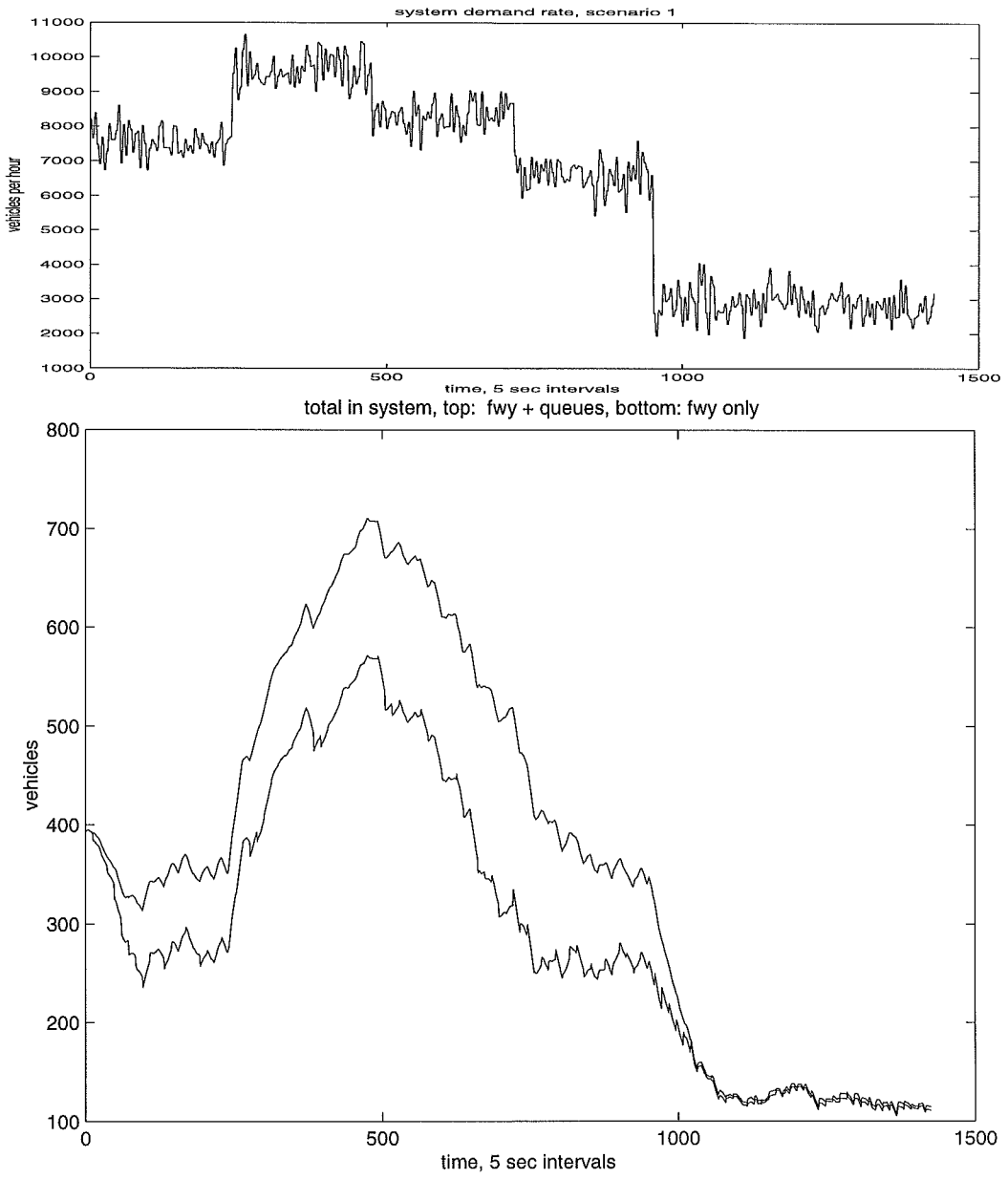


Figure 9- 20. Total vehicles in system: TR w/QM

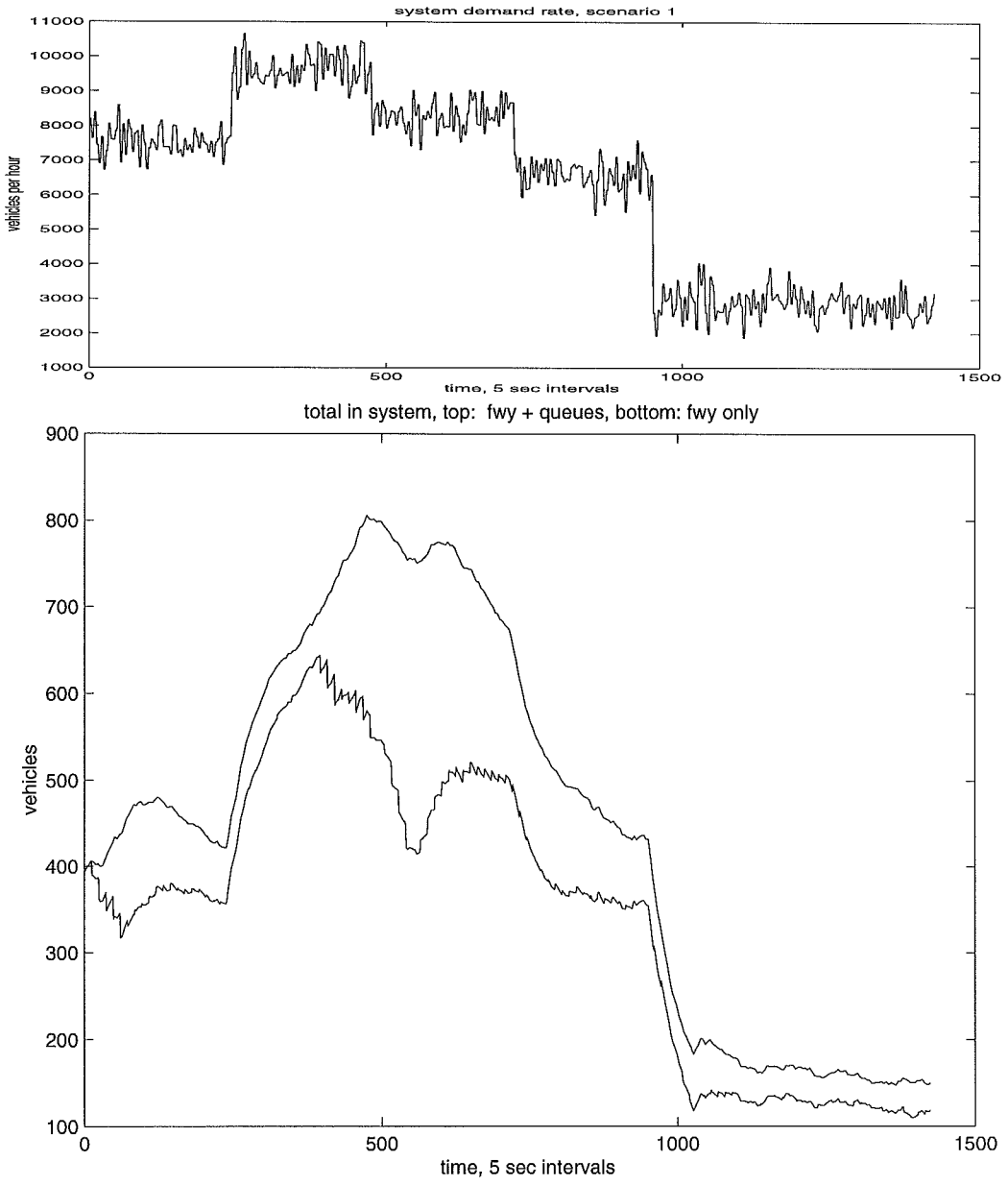


Figure 9- 21. Total vehicles in system: LP, resolved 5-min intervals

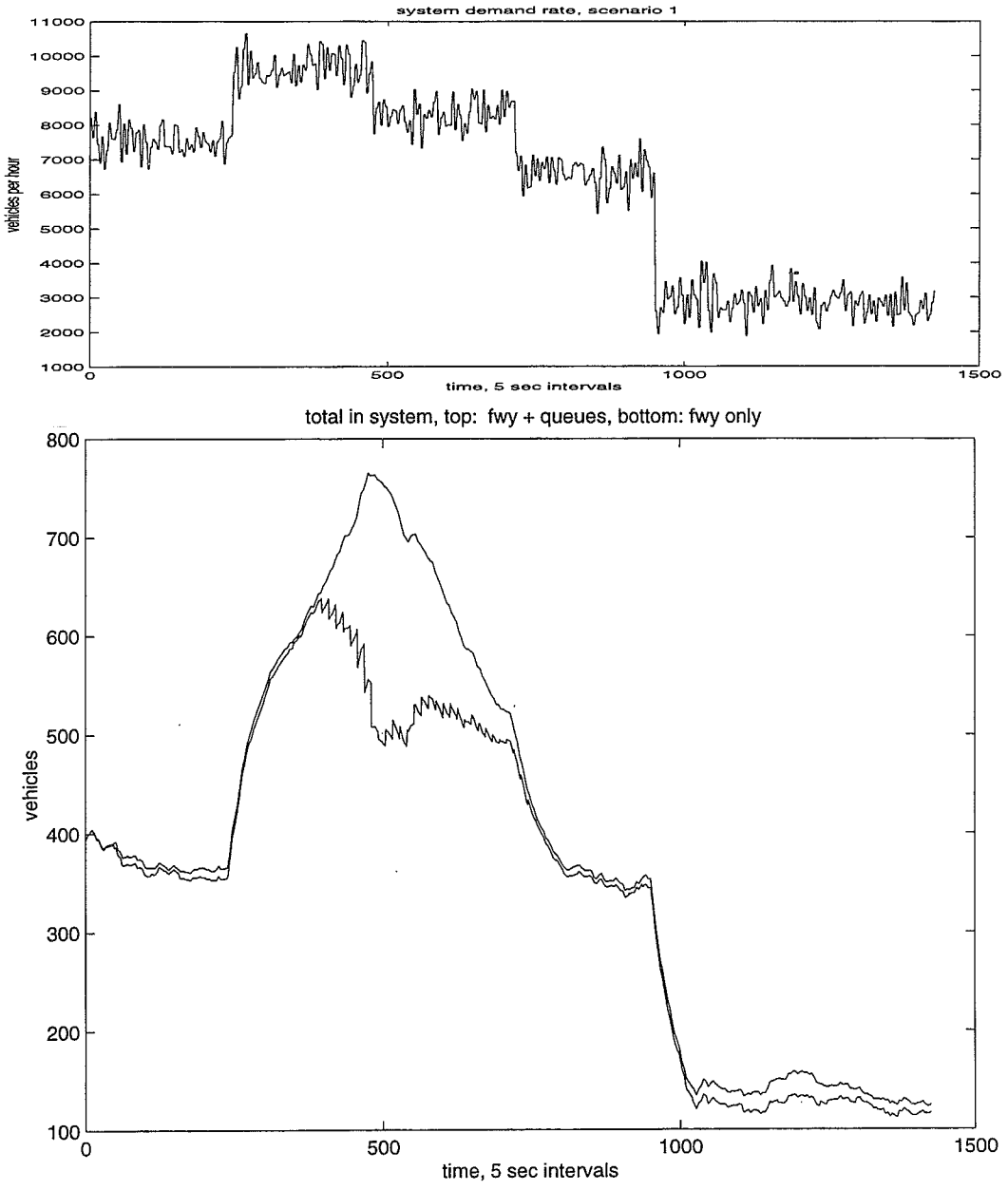


Figure 9- 22. Total vehicles in system: MILOS

Test case #2

In the second test case, a 220-minute simulation was run to represent a heavy-volume rush-hour period. During this simulation the volume and route-proportional rate tables were changed in 20-minute intervals with the first segments becoming progressively larger and the last 5 segments then becoming progressively smaller, as listed in Table 9-8. The route-proportional matrices used in each 20-minute segment of test case #2 are listed in Appendix A.

Time period / input stream	External	Ramp 1	Ramp 2	Ramp 3	Ramp 4	Ramp 5
0 - 20 minutes	4000	1350	200	300	250	220
21 - 40 minutes	4400	1450	350	400	350	280
41 – 60 minutes	4650	1600	530	500	450	380
61 – 80 minutes	4850	1500	450	600	550	480
81 – 100 minutes	5050	1600	550	500	600	540
101 – 120 minutes	4800	2050	600	480	500	750
121 – 140 minutes	4600	1600	550	480	400	700
141 – 160 minutes	4250	1400	580	580	660	650
161 – 180 minutes	4000	1200	520	500	500	600
181 – 200 minutes	3700	1100	400	500	300	400
201 – 220 minutes	3000	1050	300	300	400	500
221 – 260 minutes	1000	500	100	100	100	100

Table 9- 8. Average volume rates in each time segment, test case #2

Results for test case #2

Table 9-9 lists the average and standard deviation of the performance indices total travel time (TTT, vehicle-hours), total queue time (QT, vehicle-hours), corridor average speed (AS, km/hr), recovery time, (RT, hrs), maximum total queues (MQV), and maximum total vehicles in the system (MIS). Recovery time was computed as the time when all segments of the freeway return to a density below their respective capacities and all queues are reduced to less than 5 vehicles. Maximum total queues and maximum total vehicles in the system for each algorithm correspond to the simulation iterations displayed in Figures 9-39 through 9-42 of the total vehicles in the system and are not

averages over all 5 iterations. Figures 9-23 through 9-26 depict the comparison of the performance distributions of each algorithm in total travel time, queue time, average speed, and recovery time, respectively. Each of the distributions is scaled to [0,1] for comparison purposes (so that each distribution has equal height).

Figures 9-27 through 9-30 depict the comparison of the freeway density of a single representative simulation (for the same initial conditions and random number streams) for each of the control methods. Figures 9-31 through 9-34 depict the comparison of the resulting queues that develop at each ramp during the same simulation. Note that the queues built up for the "no control" case (as well as other algorithms where appropriate) can result from the inability of vehicles to enter the freeway due to congestion in the segment. Figures 9-35 through 9-38 depict the metering rates applied by each algorithm, where Figure 9-35 depicts the "no control" case, and thus represents the underlying demand process at each of the five on-ramps. Finally, Figures 9-39 through 9-42 depict the total vehicles in the system and total vehicles on the freeway for each of the algorithms. The upper portion of each figure displays the underlying total demand to the system, for comparison purposes. In Figures 9-39 through 9-42, the space between the two curves (if two curves are displayed) represents the total number of vehicles in ramp queues at any time instant. Note that, for comparison purposes, the scale of each graph is not identical.

Method	Avg. TTT	Std. Dev. TTT	Avg. QT	Std. Dev. QT	Avg. AS	Std. Dev AS	Avg. RT	Std. Dev. RT	MQV	MIS
No control	2588.8	42.5	144.7	13.2	79.5	0.54	3.54	0.25	956	2642
TR, w/QM	2293.6	62.1	454.3	23.5	81.8	0.69	3.39	0.35	407	1544
LP	1599.7	17.0	338.8	73.1	90.7	0.30	2.97	0.33	582	892
MILOS	1623.3	14.0	203.1	25.1	90.3	0.25	2.67	0.30	390	859

Table 9- 9. Performance comparisons, test case #2

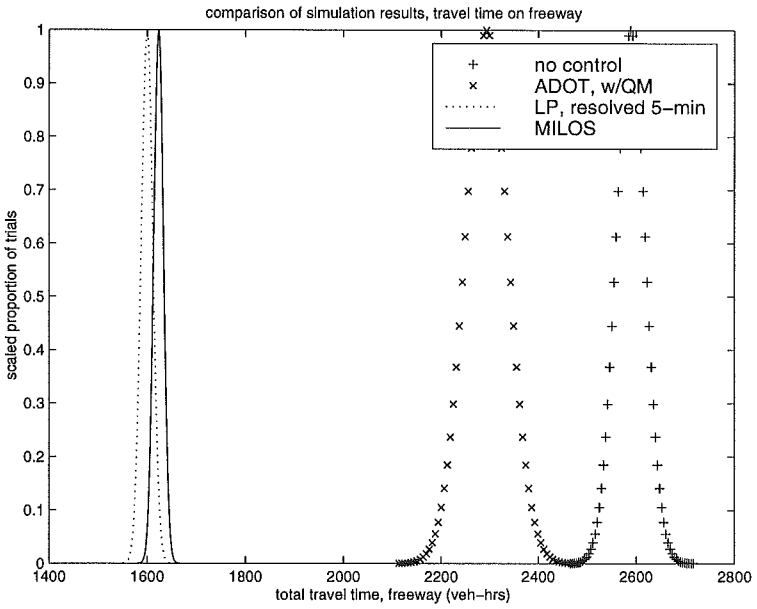


Figure 9- 23. Comparison of total travel time distributions

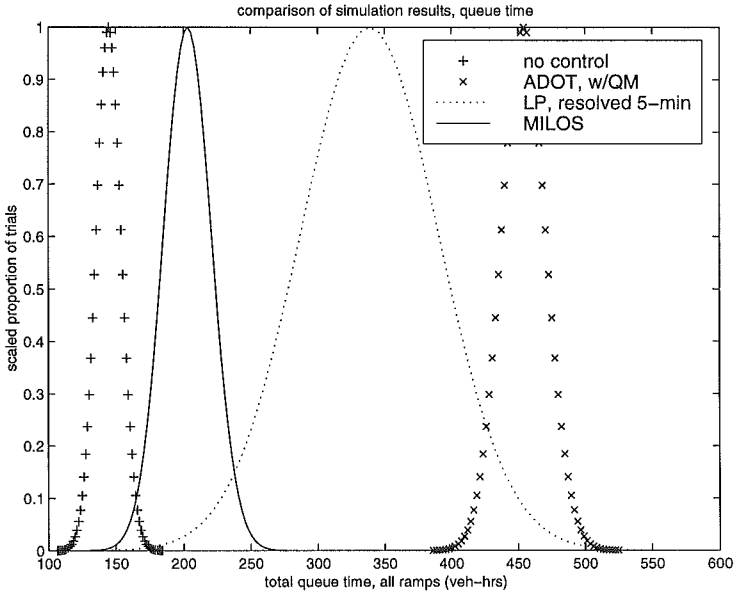


Figure 9- 24. Comparison of queue time distributions

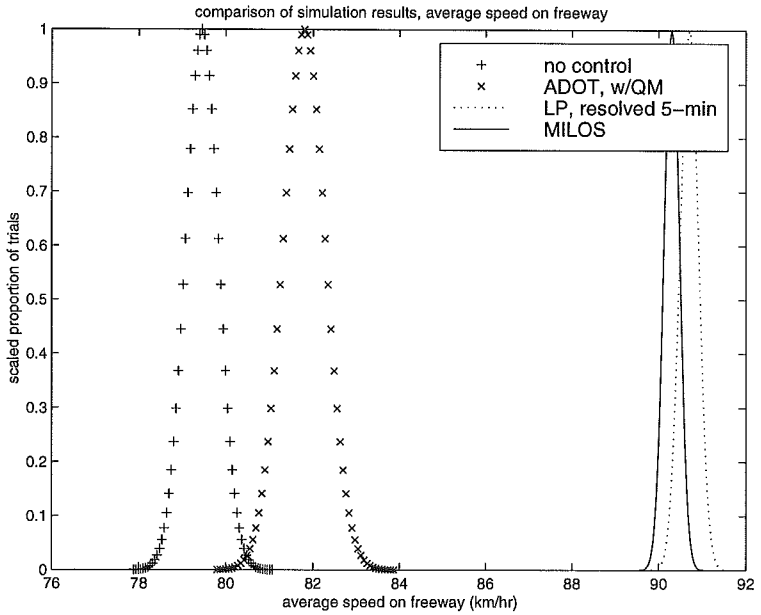


Figure 9- 25. Comparison of average speed distributions

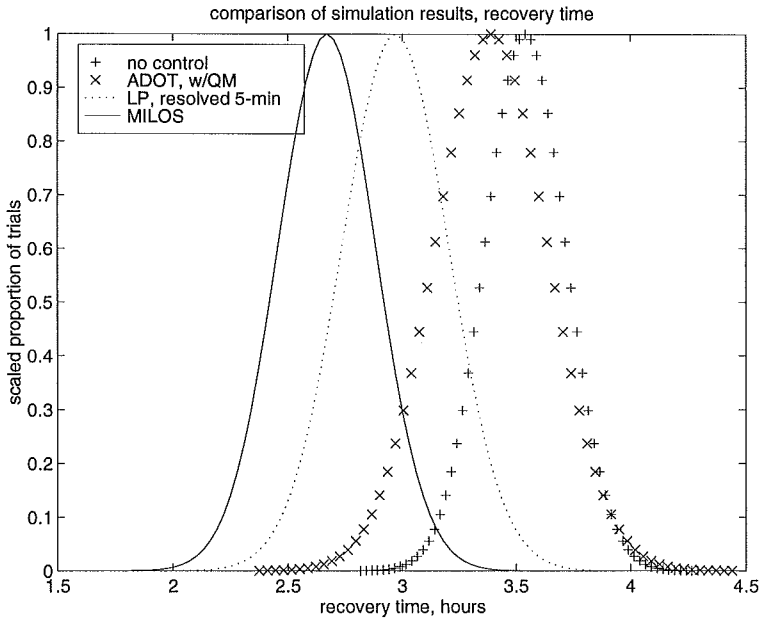


Figure 9- 26. Comparison of recovery time distributions

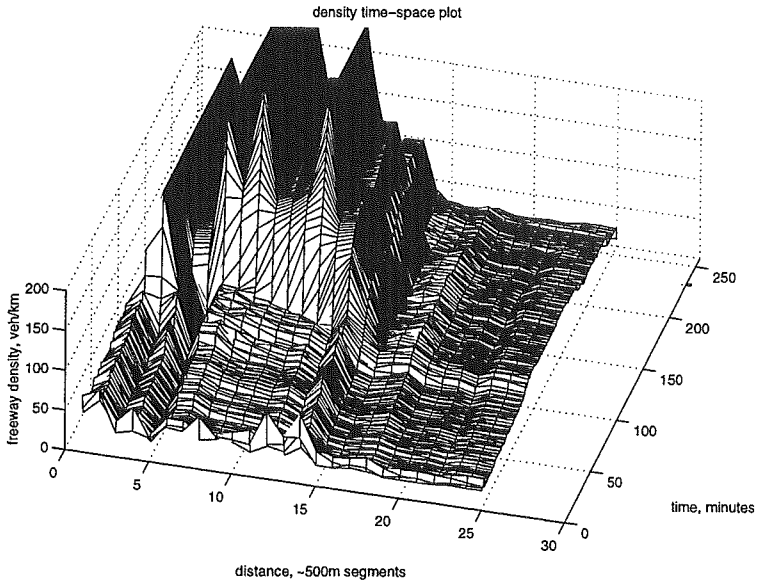


Figure 9- 27. Comparison of densities: no control

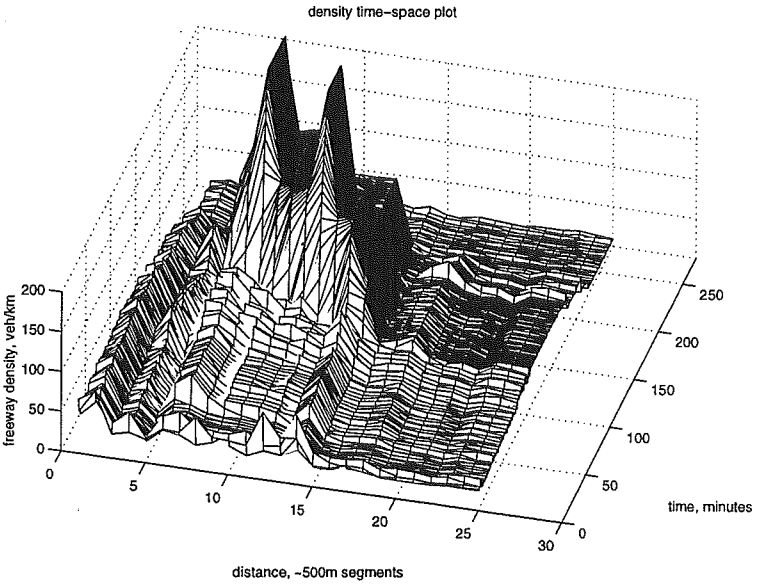


Figure 9- 28. Comparison of densities: TR, w/QM

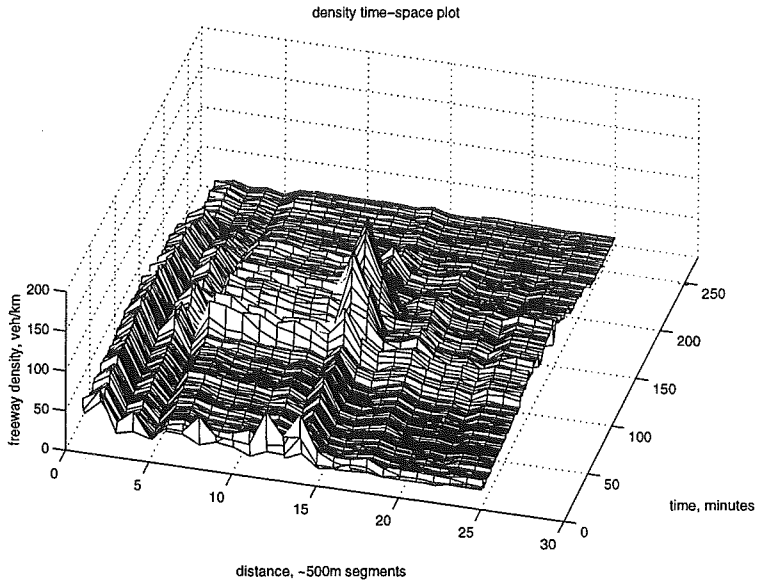


Figure 9- 29. Comparison of densities: LP, resolved each 5-minutes

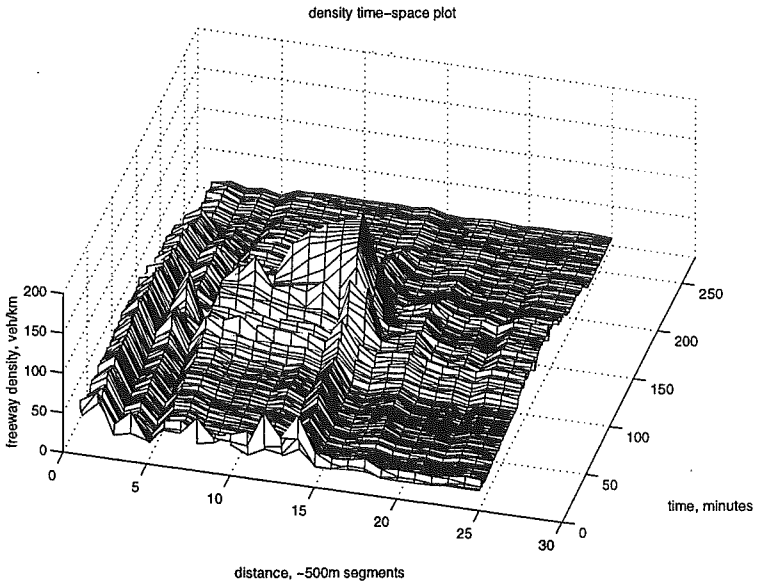


Figure 9- 30. Comparison of densities: MILOS

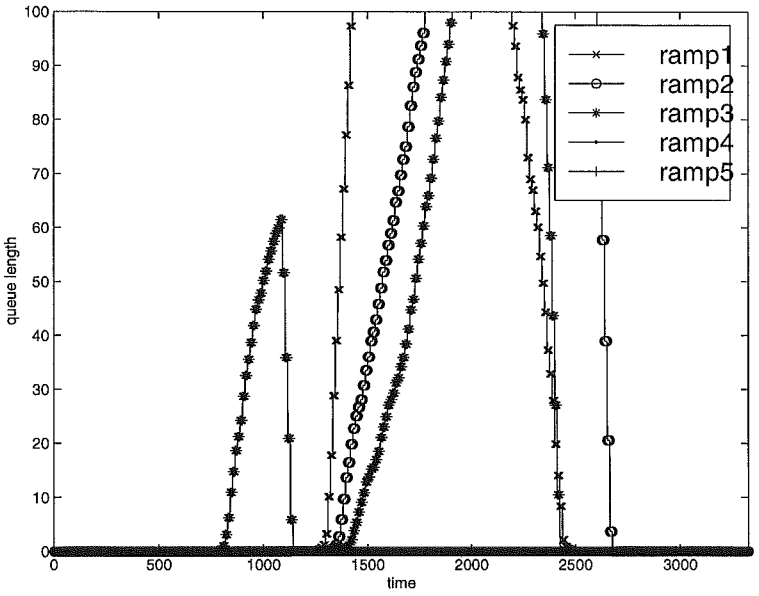


Figure 9- 31. Comparison of queue growth: no control

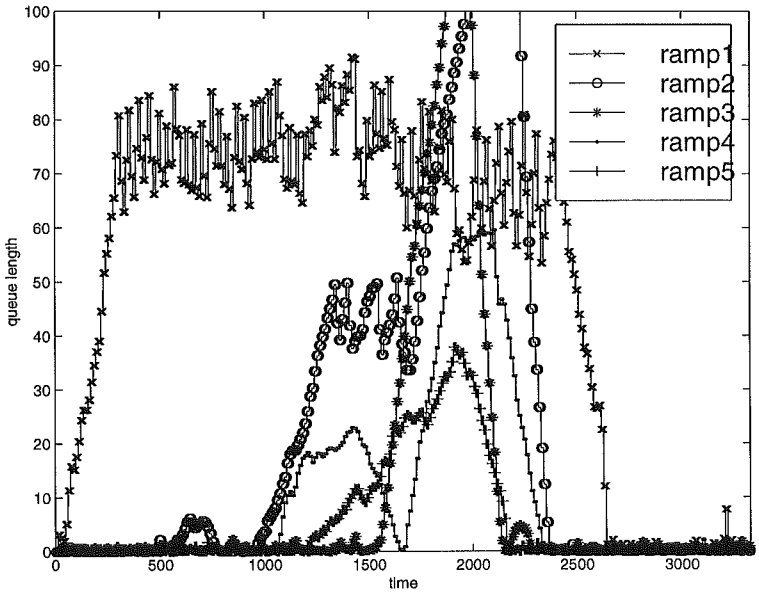


Figure 9- 32. Comparison of queue growth: TR w/QM

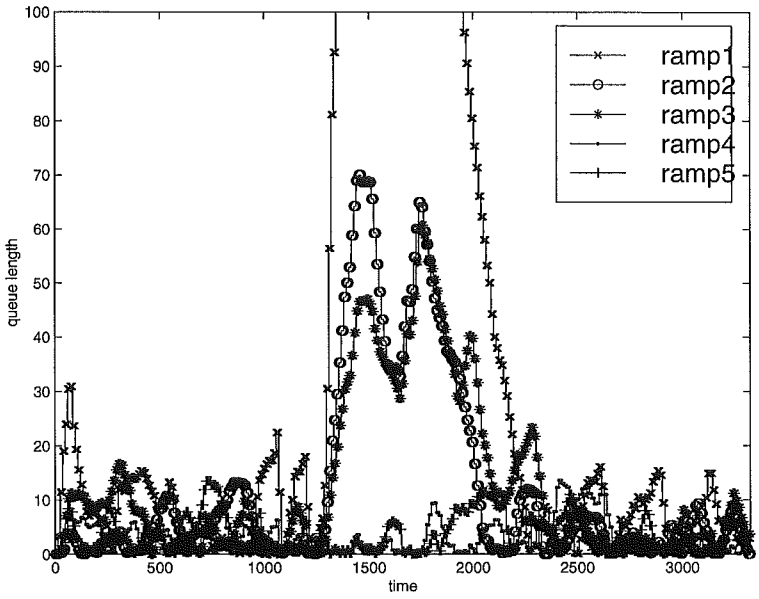


Figure 9- 33. Comparison of queue growth: LP, resolved each 5-minutes

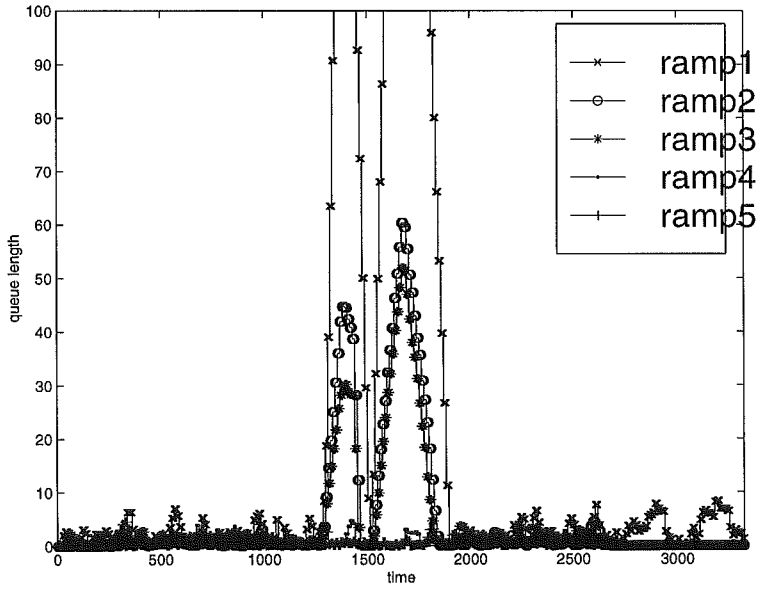


Figure 9- 34. Comparison of queue growth: MILOS

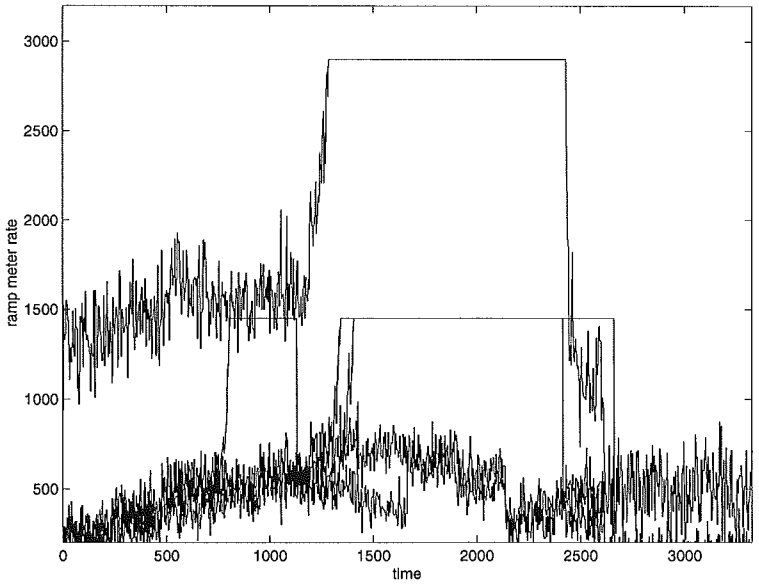


Figure 9- 35. Comparison of meter rates: no control (demands)

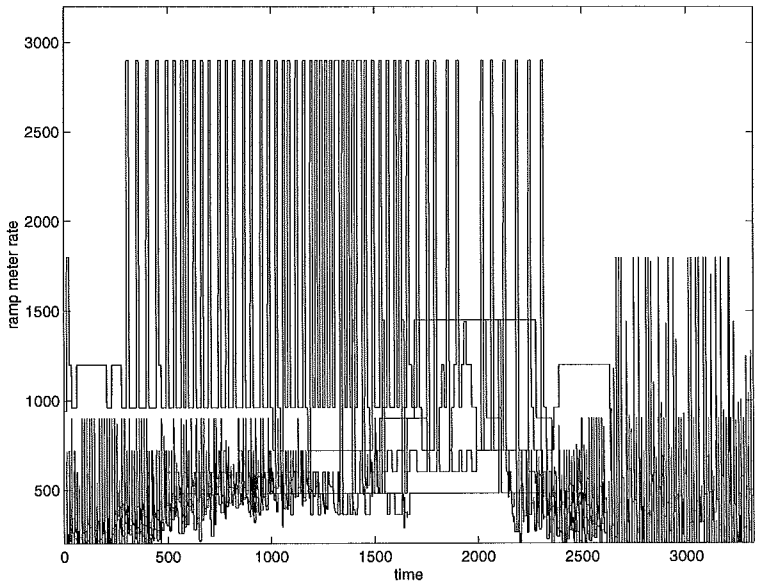


Figure 9- 36. Comparison of meter rates: TR w/QM

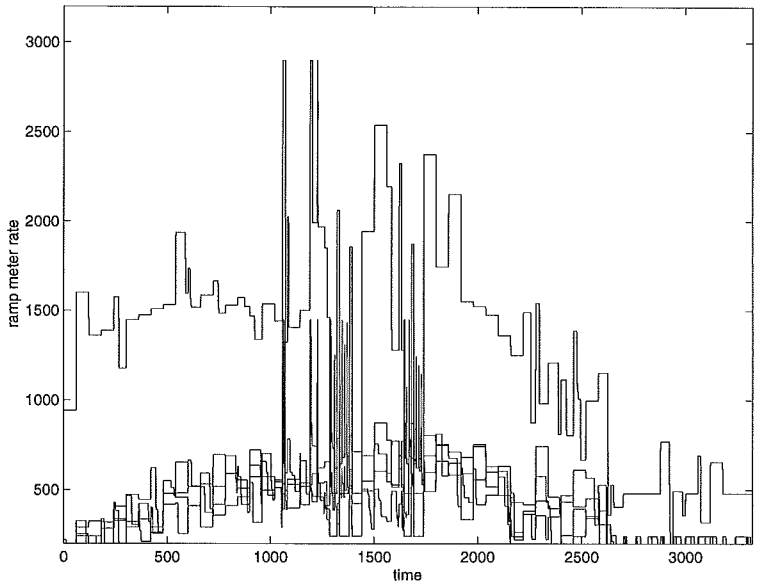


Figure 9- 37. Comparison of meter rates: LP, resolved each 5-minutes

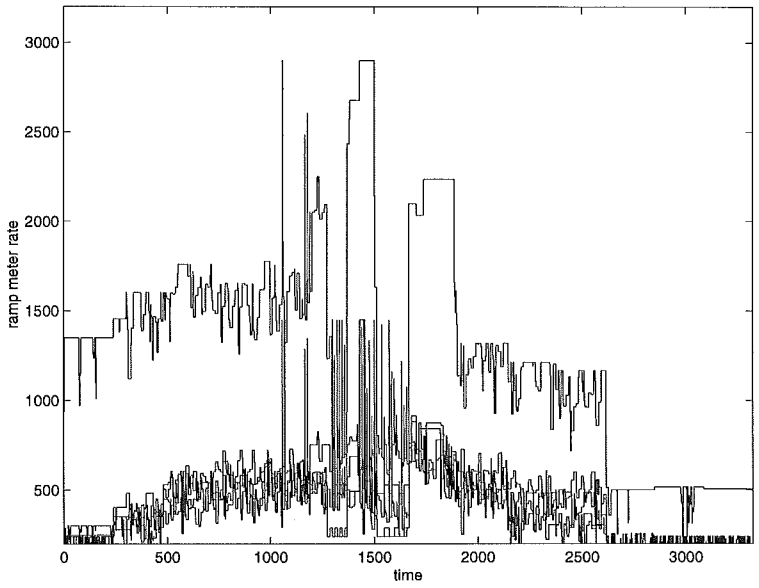


Figure 9- 38. Comparison of meter rates: MILOS

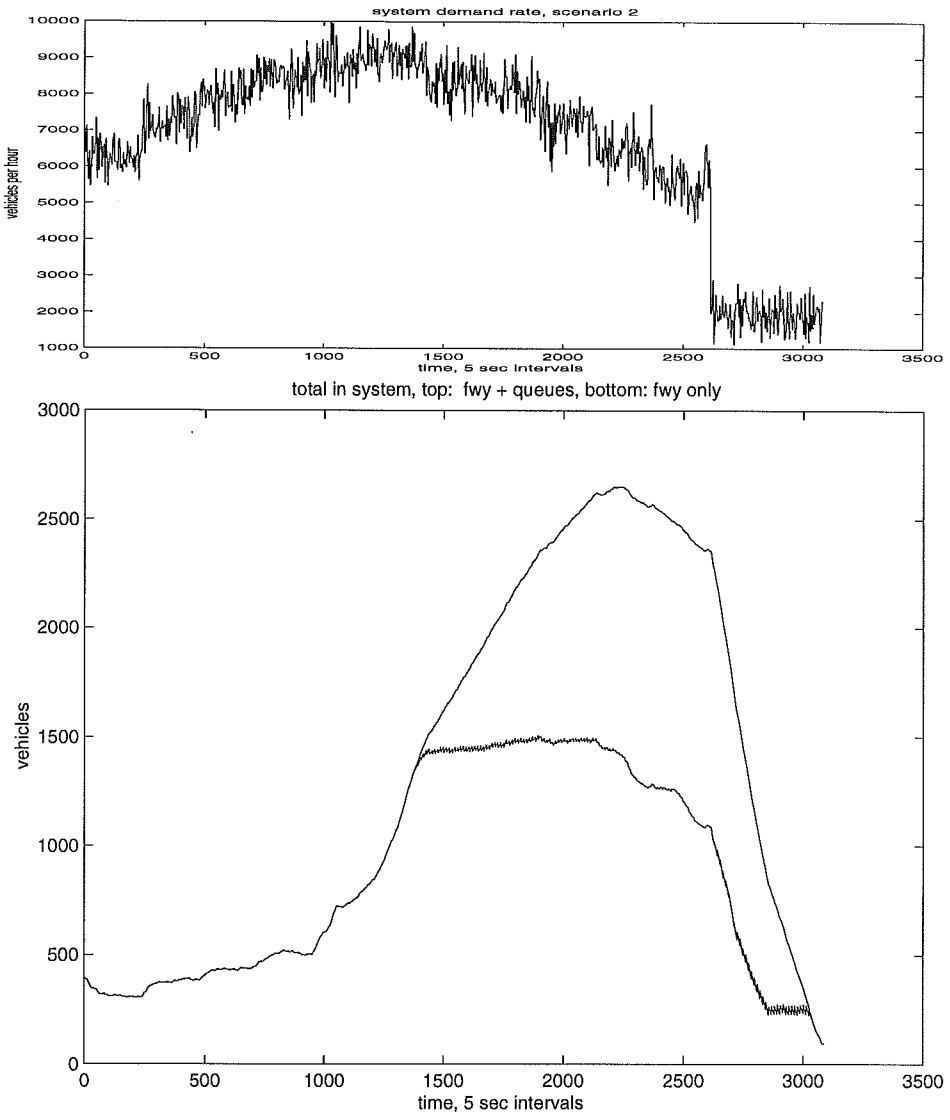


Figure 9- 39. Total vehicles in system: No control

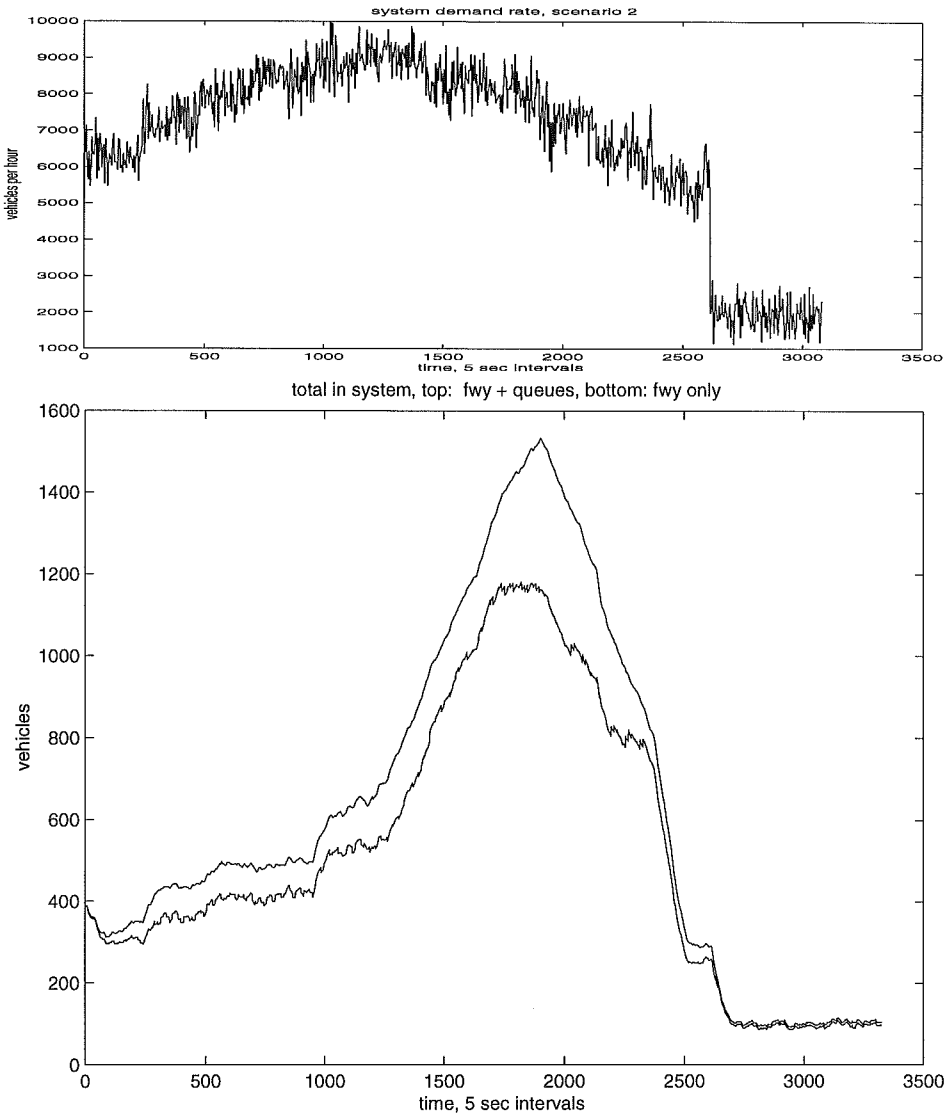


Figure 9- 40. Total vehicles in system: TR w/QM

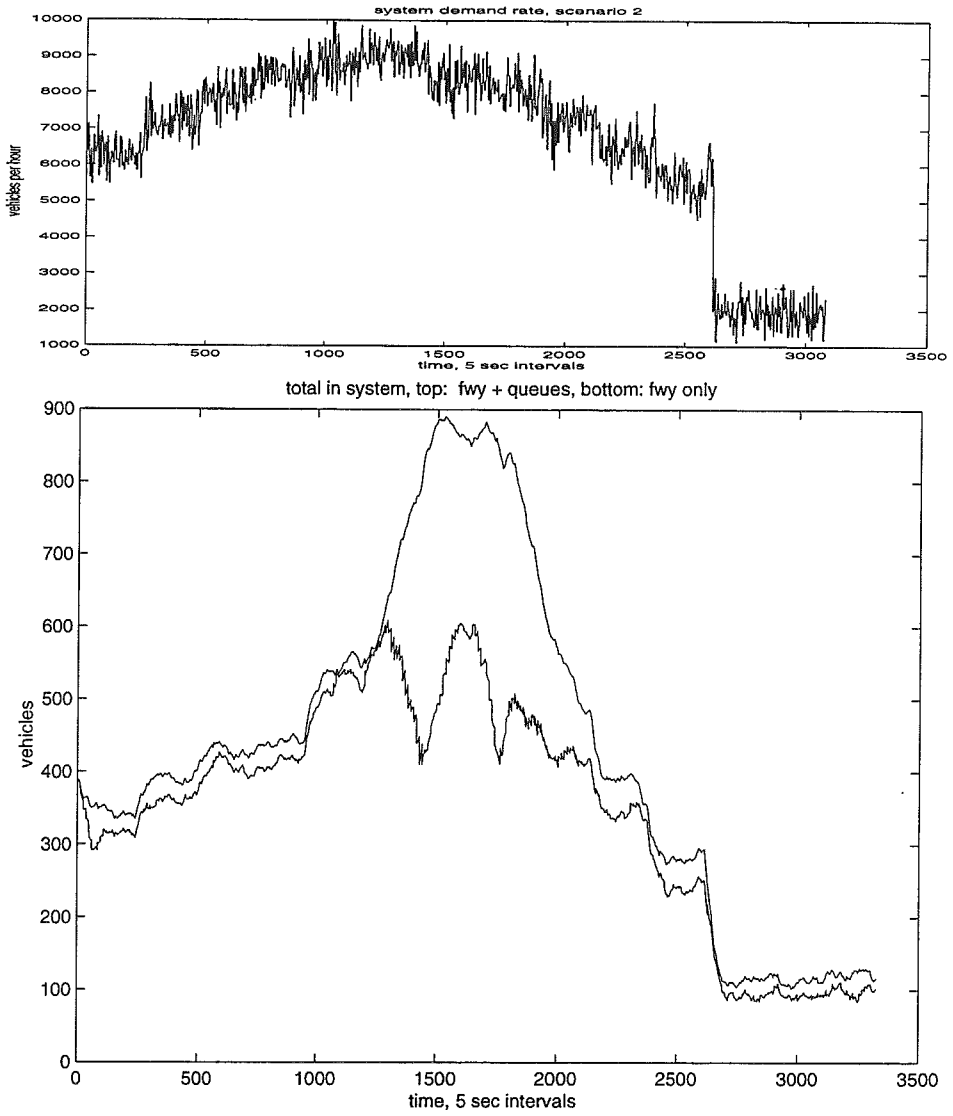


Figure 9- 41. Total vehicles in system: LP, resolved 5-min intervals

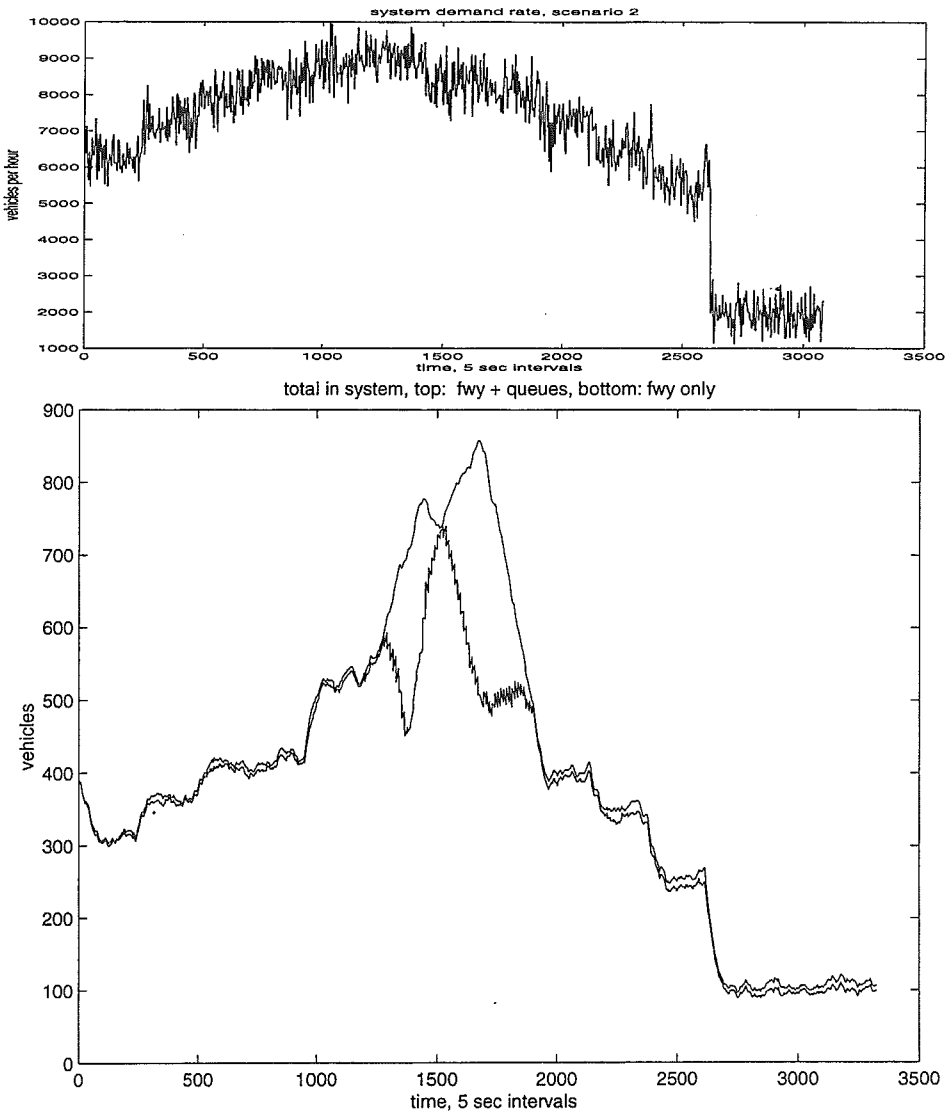


Figure 9- 42. Total vehicles in system: MILOS

The preceding figures and statistics in Table 9-9 indicate that MILOS results in the lowest maximum total queued vehicles and lowest vehicles in the system of all algorithms. Both MILOS and LP result in approximately 25% less freeway travel time and 66% fewer maximum vehicles in the system. MILOS also reduces average freeway travel time over no control and the locally traffic-responsive method and produces less queueing than LP or the locally traffic-responsive method. The improvement in average speed of MILOS is essentially equal to the improvement in speed produced by LP and significantly higher than the no control or traffic-responsive results. MILOS also displays significantly lower average recovery time over all methods.

Test case #3

In the third test case, a 220-minute simulation was run to represent a high-volume rush-hour period with an incident occurring 1-hour into the simulation and continuing for 30 minutes in section 6 of the SR202 model, as marked in Figure 9-43.

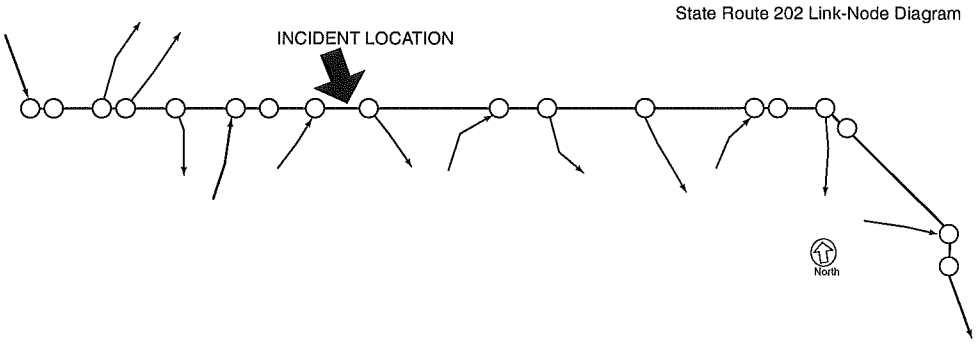


Figure 9- 43. SR202 model indicating incident location

During this simulation the volume and route-proportional rate tables were changed in 20-minute intervals. The average volumes are listed in Table 9-10. The variance of each demand distribution was given as in eqn. 9-1. The route-proportional matrices used in each 20-minute segment of test case #3 are listed in Appendix A.

Time period	External	Ramp 1	Ramp 2	Ramp 3	Ramp 4	Ramp 5
0 - 20 minutes	4000	1350	200	300	250	220
21 - 40 minutes	4400	1450	350	400	350	280
41 – 60 minutes	4650	1600	530	500	450	380
61 – 80 minutes*	4850	1500	450	600	550	480
81 – 100 minutes*	5050	1600	550	500	600	540
101 – 120 minutes	4800	2050	600	480	500	750
121 – 140 minutes	4600	1600	550	480	400	700
141 – 160 minutes	4250	1400	580	580	660	650
161 – 180 minutes	4000	1200	520	500	500	600
181 – 200 minutes	3700	1100	400	500	300	400
201 – 220 minutes	3000	1050	300	300	400	500
221 – 260 minutes	1000	500	100	100	100	100
*incident occurs						

Table 9- 10. Average volume rates in each time segment, test case #3

Results for test case #3

Table 9-11 lists the average and standard deviation of the performance indices total travel time (TTT, vehicle-hours), total queue time (QT, vehicle-hours), corridor average speed (AS, km/hr), recovery time, (RT, hrs), maximum total queues (MQV), and maximum total vehicles in the system (MIS). Recovery time was computed as the time when all segments of the freeway return to a density below their respective capacities and all queues are reduced to less than 5 vehicles. Maximum total queues and maximum total vehicles in the system for each algorithm correspond to the simulation iterations displayed in Figures 9-60 through 9-64 of the total vehicles in the system and are not averages over all 5 iterations. Figures 9-44 through 9-47 depict the comparison of the performance distributions of each algorithm in total travel time, queue time, average speed, and recovery time, respectively. Each of the distributions is scaled to [0,1] for comparison purposes (so that each distribution has equal height).

Figures 9-48 through 9-51 depict the comparison of the freeway density of a single representative simulation (for the same initial conditions and random number streams) for each of the control methods. Figures 9-52 through 9-55 depict the comparison of the resulting queues that develop at each ramp during the same simulation. Note that the queues built up for the "no control" case (as well as other algorithms where appropriate) can result from the inability of vehicles to enter the freeway due to congestion in the segment. Figures 9-56 through 9-59 depict the metering rates applied by each algorithm, where Figure 9-56 depicts the "no control" case, and thus represents the underlying demand process at each of the five on-ramps. Finally, Figures 9-60 through 9-64 depict the total vehicles in the system and total vehicles on the freeway for each of the algorithms. The upper portion of each figure displays the underlying total demand to the system and the incident duration is marked on each graph of total vehicles in the system. As in previous test cases, the space between the two curves represents the total number of

vehicles in ramp queues at any time instant. Note that, for comparison purposes, the scale of each graph is not necessarily identical.

Method	Avg. TTT	Std. Dev. TTT	Avg. QT	Std. Dev. QT	Avg. AS	Std. Dev AS	Avg. RT	Std. Dev. RT	MQV	MIS
No control	3256.4	25.9	743.6	44.2	69.6	0.26	3.75	0.08	620	1425
TR, w/QM	3375.0	31.7	1325.0	54.8	68.5	0.34	3.67	0.25	913	2220
LP	2025.0	20.6	1255.0	55.8	83.6	0.34	3.33	0.27	1048	1474
MILOS	2079.1	63.8	1088.0	199.3	82.8	1.01	3.04	0.18	852	1330

Table 9- 11. Performance comparison, test case #3

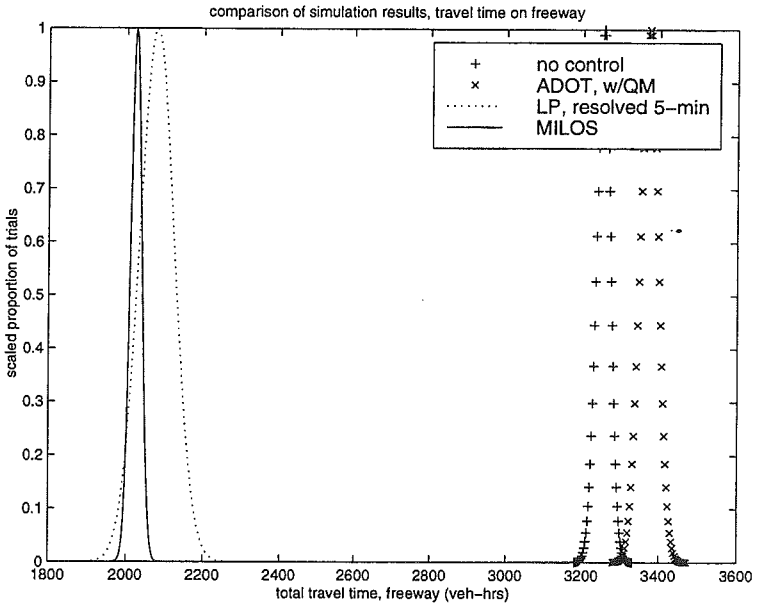


Figure 9- 44. Comparison of total travel time distributions

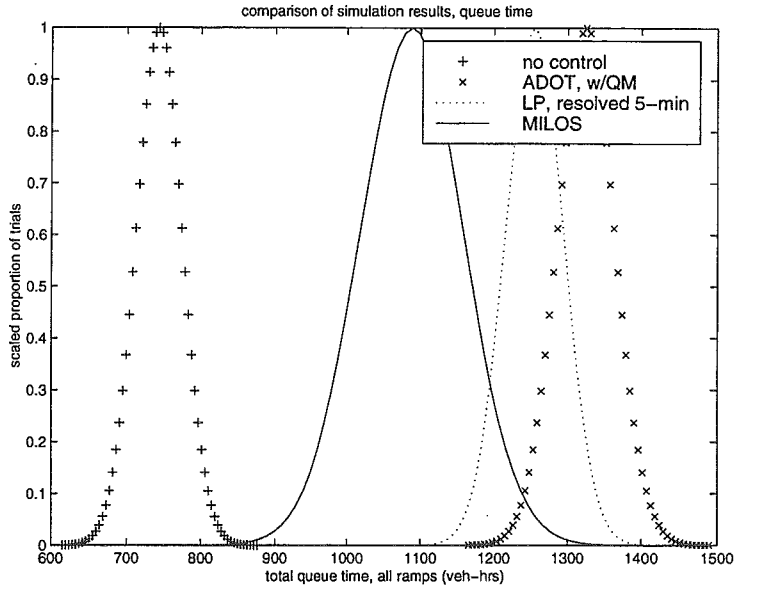


Figure 9- 45. Comparison of queue time distributions

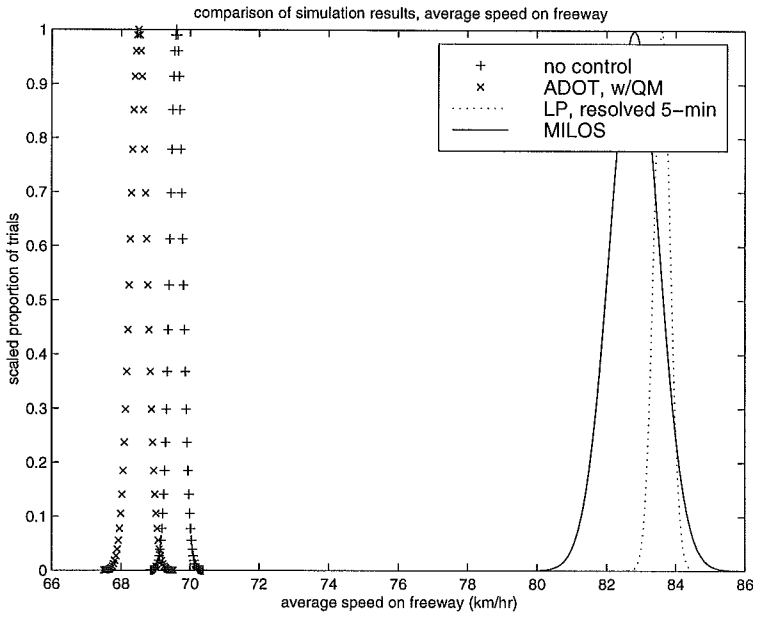


Figure 9- 46. Comparison of average speed distributions

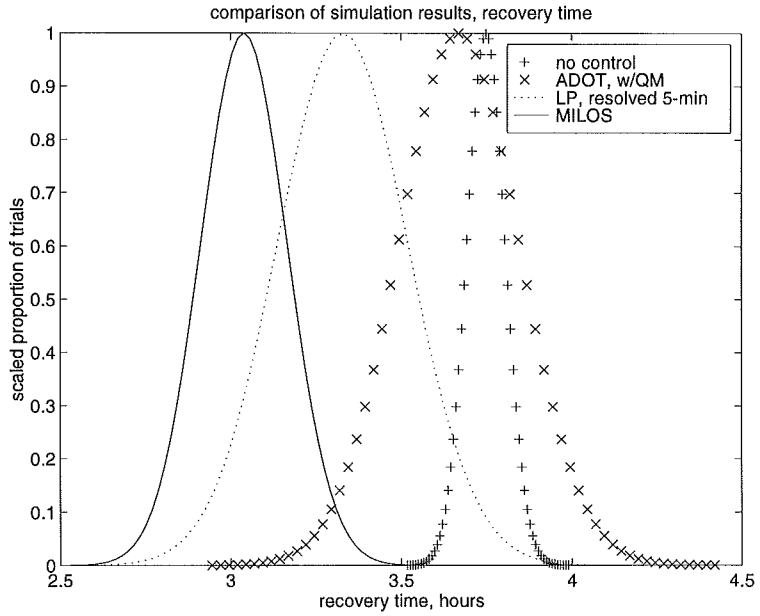


Figure 9- 47. Comparison of recovery time distributions

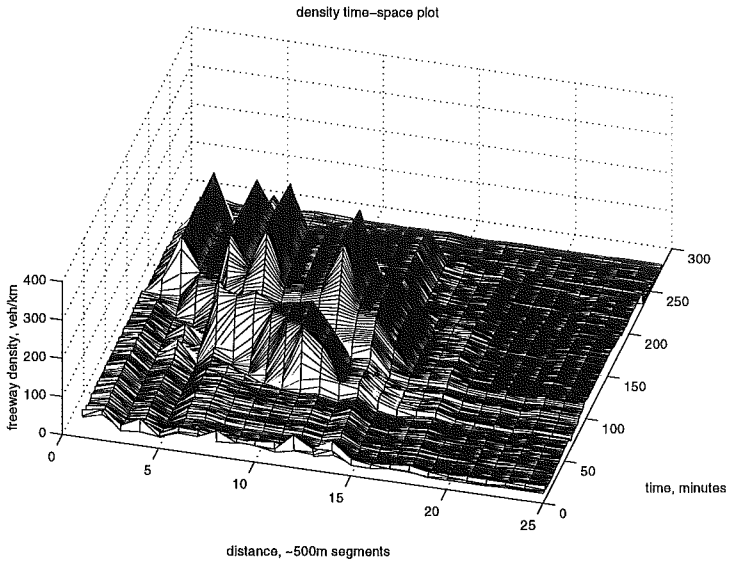


Figure 9- 48. Comparison of densities: no control

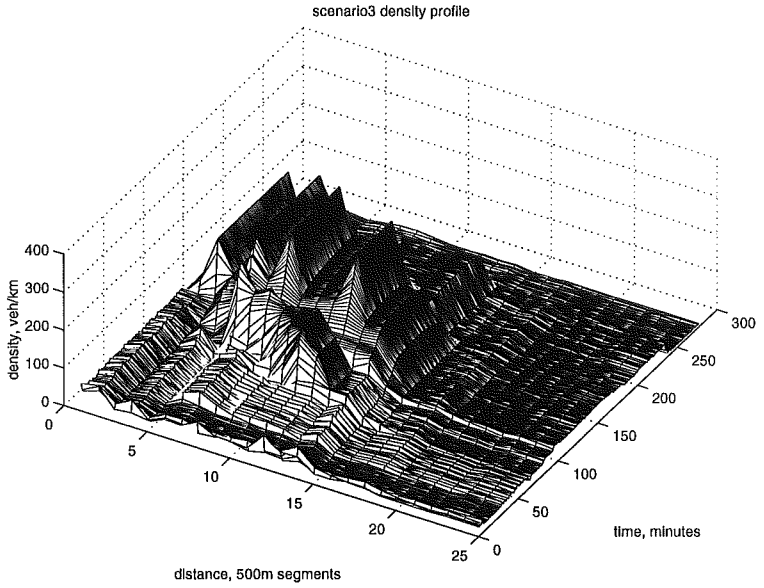


Figure 9- 49. Comparison of densities: TR, w/QM

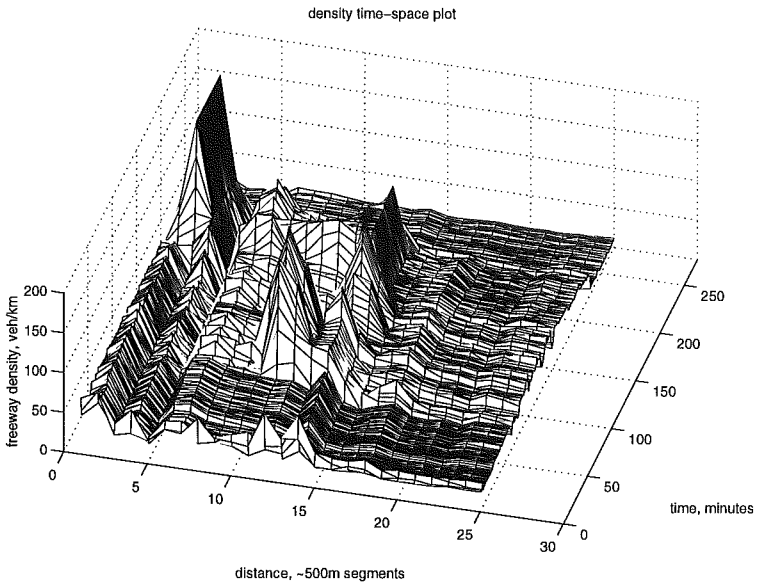


Figure 9- 50. Comparison of densities: LP, resolved each 5-minutes

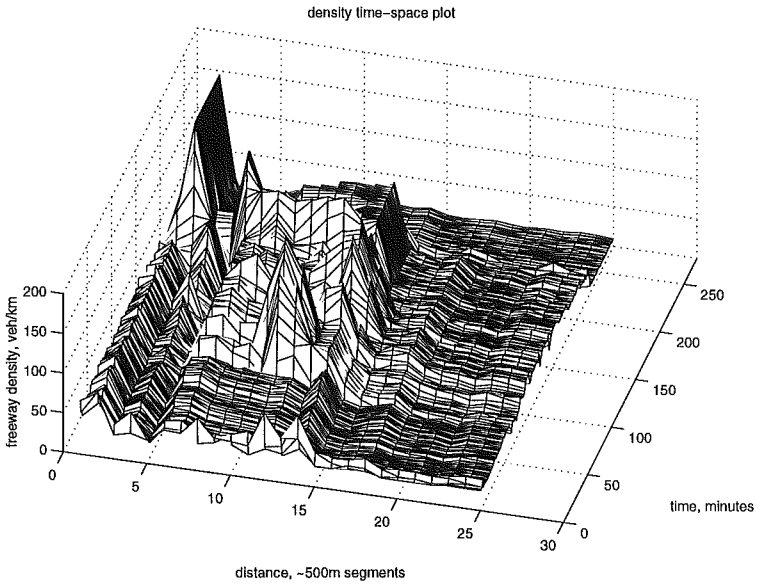


Figure 9-51. Comparison of densities: MILOS

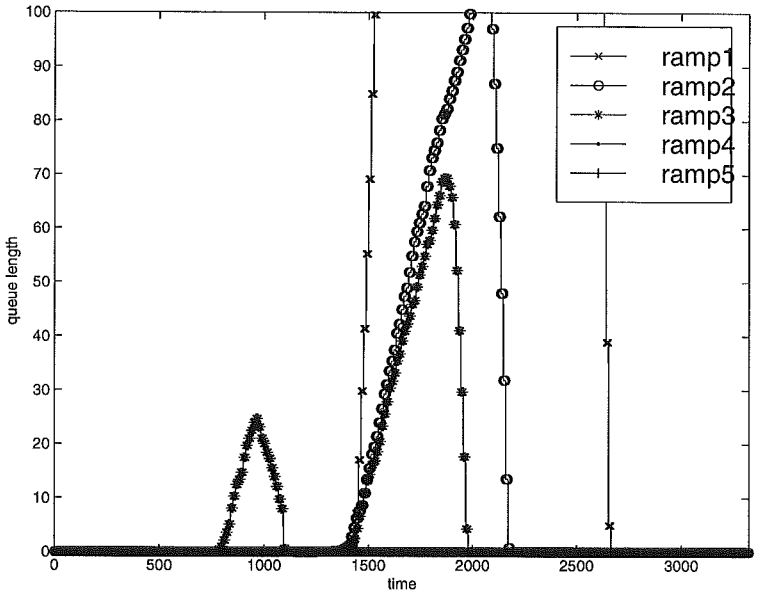


Figure 9- 52. Comparison of queue growth: no control

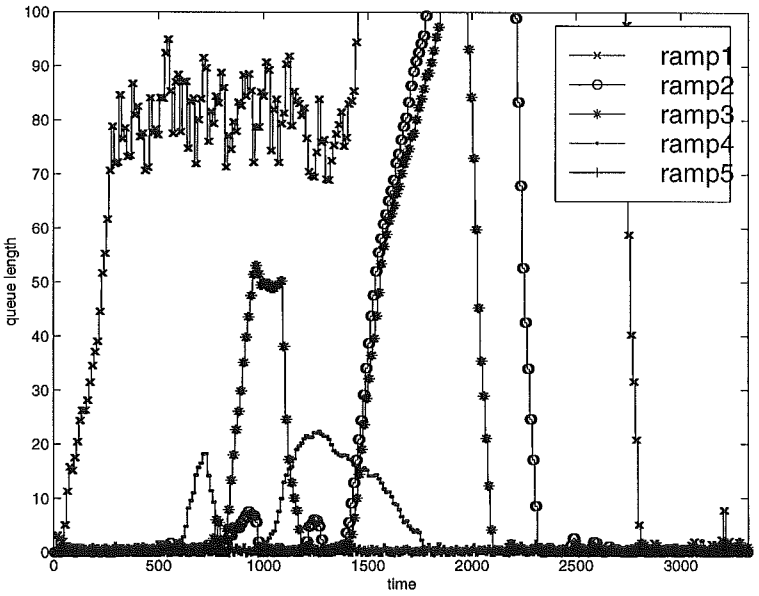


Figure 9- 53. Comparison of queue growth: TR w/QM

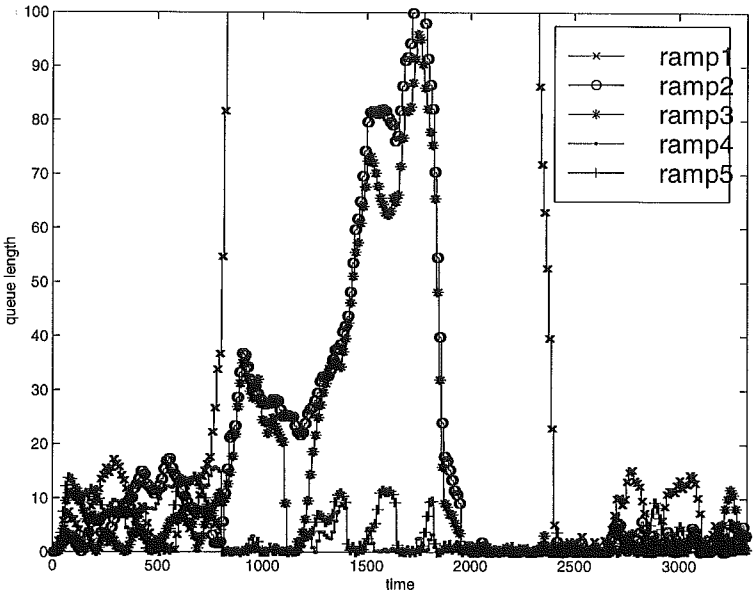


Figure 9- 54. Comparison of queue growth: LP, resolved each 5-minutes

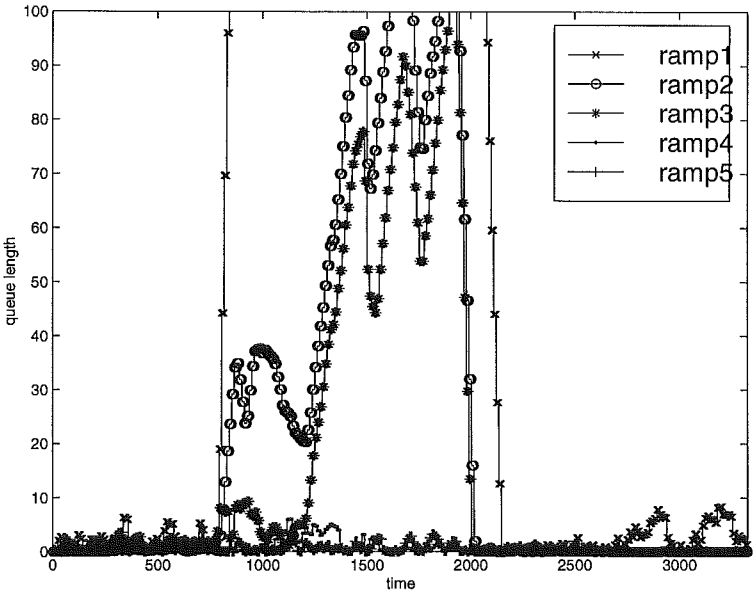


Figure 9- 55. Comparison of queue growth: MILOS

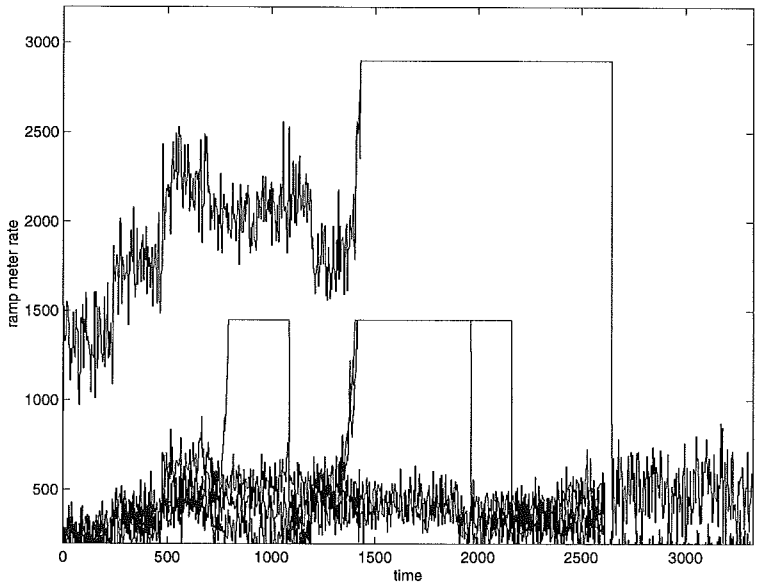


Figure 9- 56. Comparison of meter rates: no control (demands)

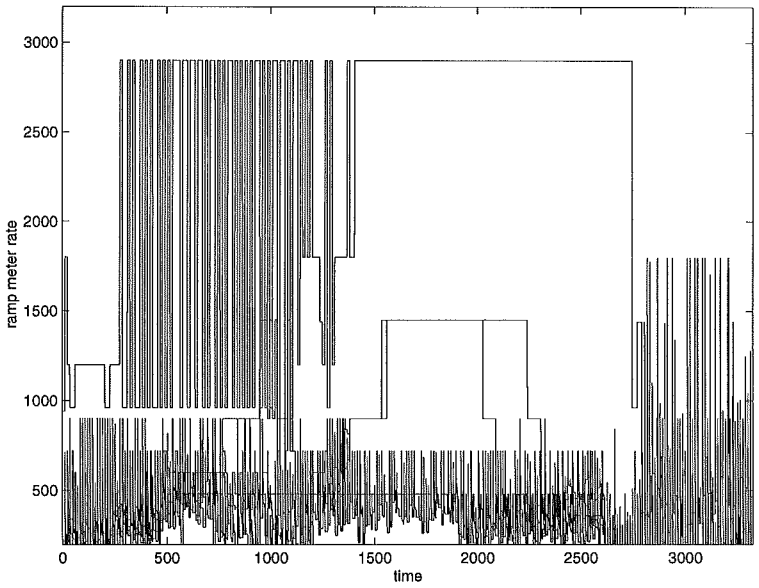


Figure 9- 57. Comparison of meter rates: TR w/QM

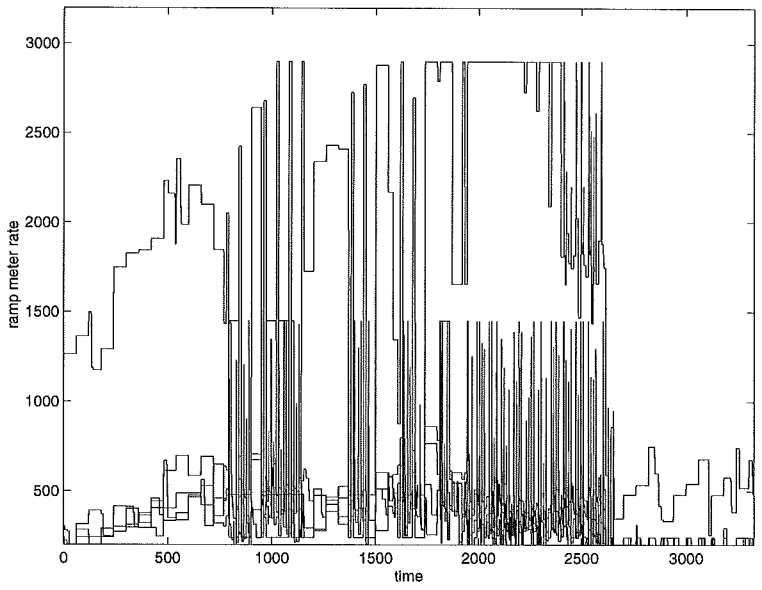


Figure 9- 58. Comparison of meter rates: LP, resolved each 5-minutes

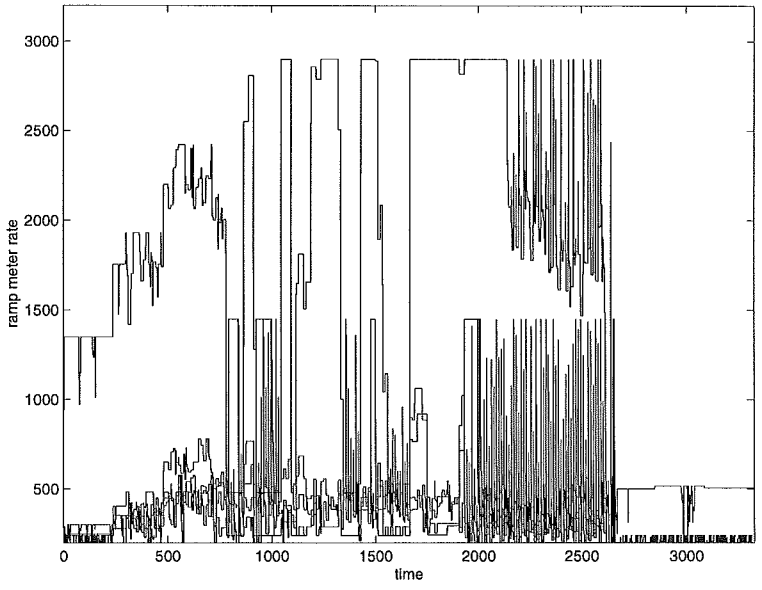


Figure 9- 59. Comparison of meter rates: MILOS

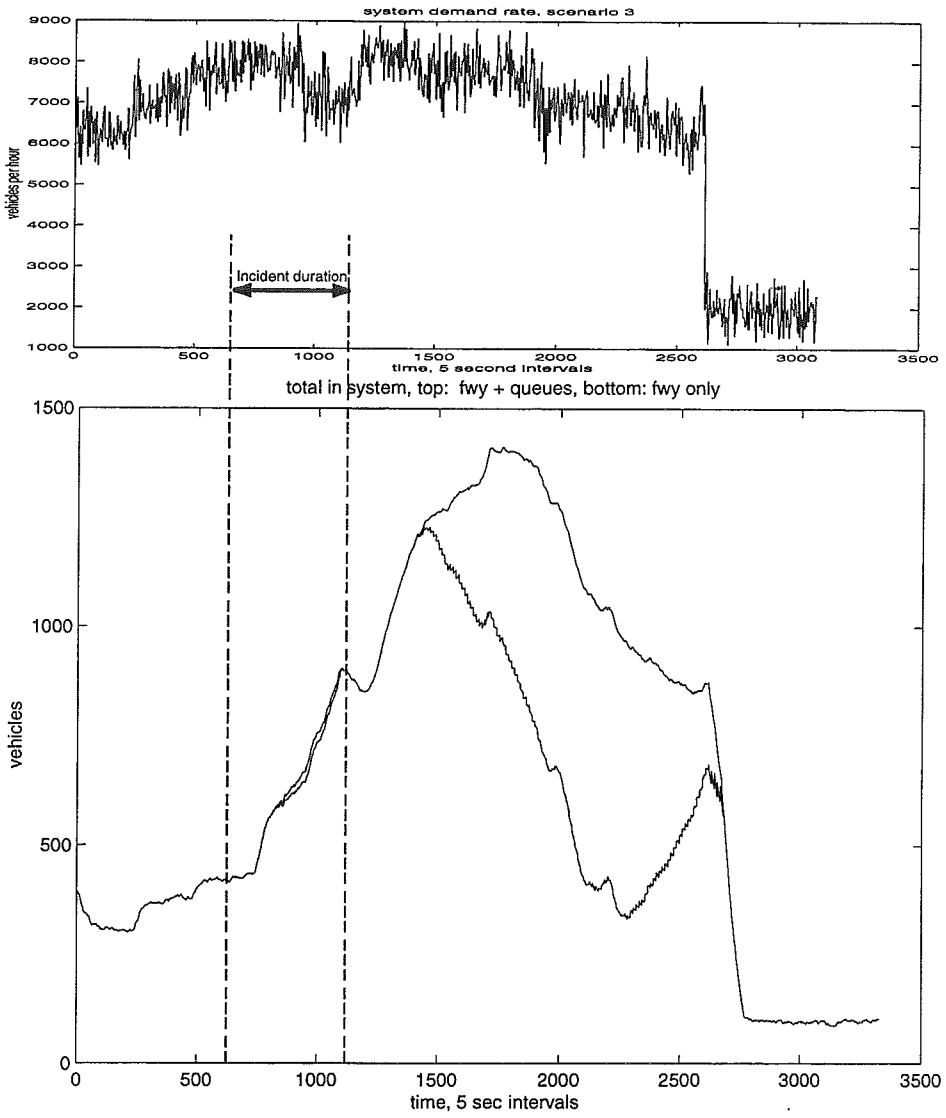


Figure 9- 60. Total vehicles in system: No control

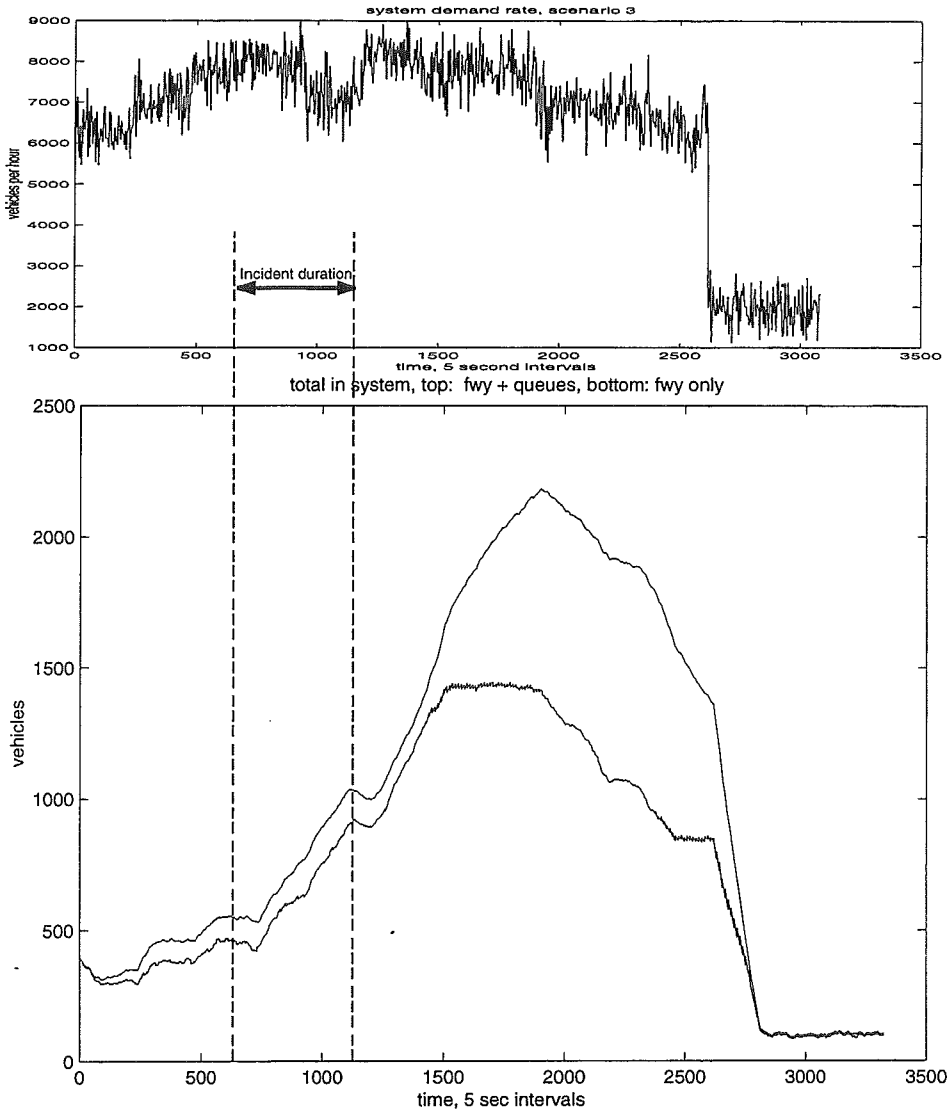


Figure 9- 61. Total vehicles in system: TR w/QM

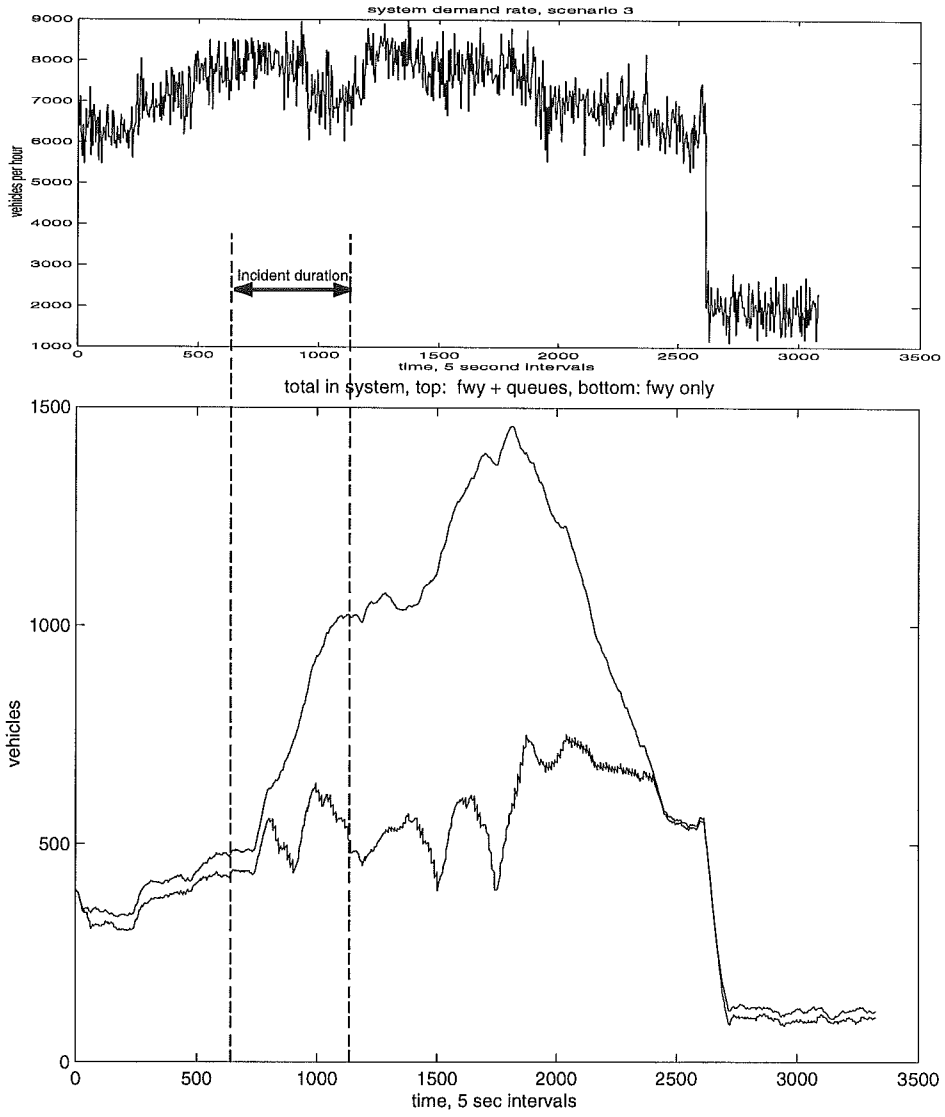


Figure 9- 62. Total vehicles in system: LP, resolved 5-min intervals

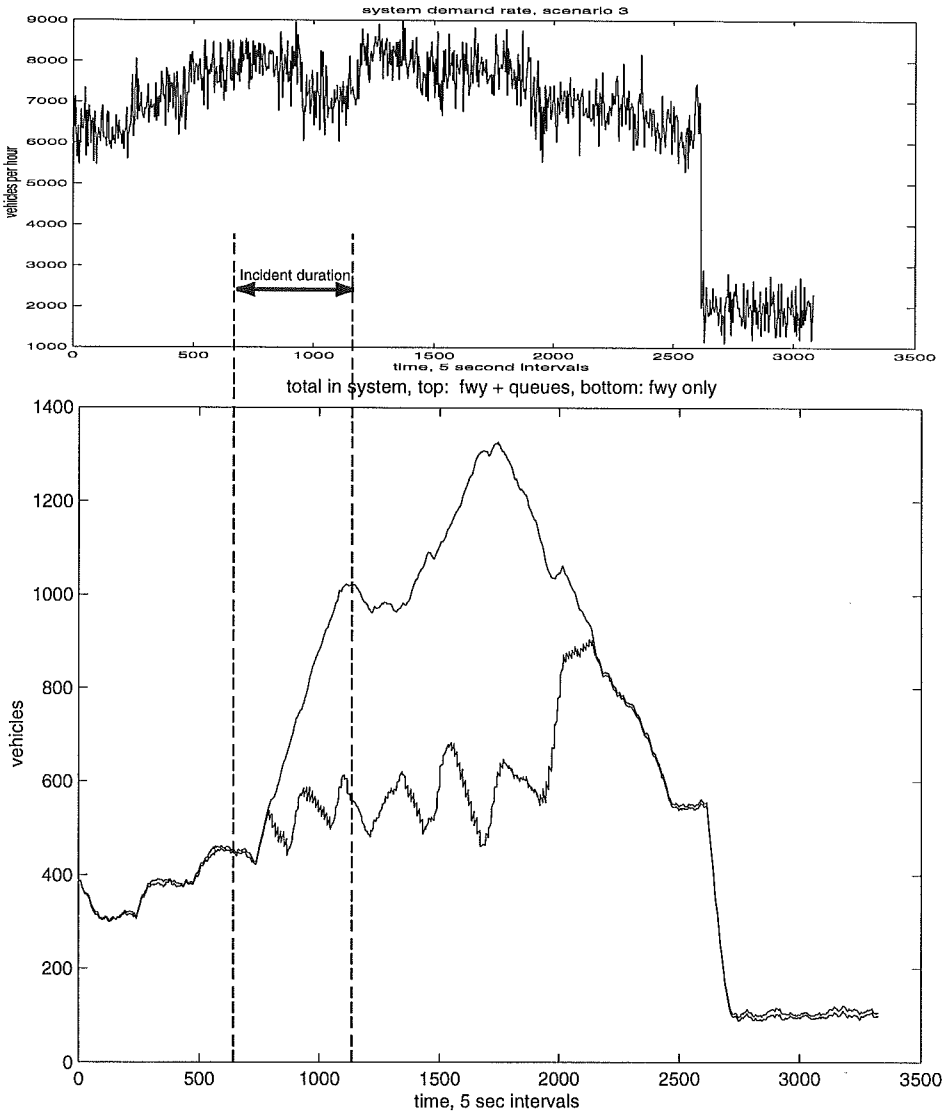


Figure 9-63. Total vehicles in system: MILOS

From the preceding figures and the statistics in Table 9-11 it is indicated that MILOS results in significantly lower average recovery time over all methods for test case #3. MILOS displays approximately 36% less freeway travel time and 20% less *total* travel time than the no control case. MILOS also produces less queuing than LP or the traffic-responsive method and has the lowest maximum total vehicles in the system of any algorithm. The improvement in average speed of MILOS is essentially equal to the improvement in speed produced by LP and 19% higher than the network average speed of the no control or locally traffic-responsive cases.

Summary

Table 9-12 lists the range of percentage differences of the average performance and standard deviation of performance of MILOS versus the other control methods used in this simulation experiment for the three test cases. For example, use of the MILOS algorithm produced from 8% to 36% less freeway travel time, 3% to 18% higher average speed and 6% to 24% earlier recovery time over the no-control case. These results could be considered conclusive since the range of results does not include zero for the three test cases. In contrast, the standard deviation of the freeway travel time under MILOS ranged from +146% to –74% versus the no control case, depending on the scenario. As such, it would not be concluded that MILOS produces either lower or higher average variability over the no control case.

MILOS vs. Method	Avg. FTT %	Std. Dev. TTT %	Avg. QT %	Std. Dev. QT %	Avg. AS %	Std. Dev AS %	Avg. RT %	Std. Dev. RT %	MQV %	MIS %
No control	-8 to –36	+146 to –74	+N/A to +46	+N/A to +90	+18 to +3	+288 to –75	-6 to –24	+125 to +20	+N/A to –59	+4 to –67
TR, w/QM	-29 to –38	+101 to –77	+40 to –55	+263 to +6	+20 to +10	+197 to –63	-17 to –24	+20 to –28	-4 to –59	-40 to –67
LP	+3 to –1	+209 to -17	-13 to -40	+257 to –65	+1 to –1	+197 to –17	-8 to –12	-9 to -62	-15 to -32	-2 to –9

Table 9-12. Comparison of MILOS results with alternatives

In a qualitative sense, the comparison of the MILOS method with the three other cases can be summarized in the Table 9-13. The numerical ranges are replaced with the qualitative judgments: Significantly lower, Marginally lower, Essentially same, Inconclusive, Marginally higher, and Significantly higher. If the range includes zero, but is extremely wide, the judgment is "inconclusive". If the range included zero but was much less variable than "inconclusive" results, the judgment was indicated as "same".

MILOS vs. Method	Avg. FTT %	Std. Dev. TTT %	Avg. QT %	Std. Dev. QT %	Avg. AS %	Std. Dev AS %	Avg. RT %	Std. Dev. RT %	MQV %	MIS %
No control	Sig. lower	Inc.	Sig. higher	Sig. higher	Sig. higher	Inc.	Sig. lower	Sig. Higher	Marg. lower	Marg. lower
TR, w/QM	Sig. lower	Inc.	Inc.	Sig. higher	Sig. higher	Inc.	Sig. lower	Same	Sig. lower	Sig. lower
LP	Same	Inc.	Sig. lower	Inc.	same	Inc.	Sig. lower	Sig. Lower	Sig. lower	Marg. lower

Table 9- 13. Qualitative comparisons of MILOS versus other algorithms

Chapter 10: Conclusions

General results

From the three simulation test cases, it is concluded that use of the MILOS ramp metering control method results in lower total freeway travel time, higher average freeway speed, earlier recovery time from congested conditions, lower maximum total queue lengths, and lower maximum total vehicles in the system than the other control methods evaluated. The results of the variability of most performance measures of MILOS versus the other methods tested were largely inconclusive. However, when the average performance of MILOS was significantly separated from other methods, the variability was not large enough to invalidate the conclusion of the performance measure being significantly larger or smaller.

MILOS versus "no control"

The general costs and benefits of ramp metering methods, such as minimizing the total travel time of freeway users, using freeway capacity efficiently, and decreasing freeway congestion and shock waves resulting from merging platoons were indicated to be valid when comparing MILOS versus doing no metering (or the "no control" case). The expected benefit of ramp metering to reduce the variance of corridor trip times over "no control" could not be shown, although it was indicated that the average corridor speed was increased significantly and the resulting variance of speed did not cause the speed distribution of the MILOS method to overlap the speed distribution of "no control" in all test cases.

It is a somewhat surprising result that MILOS results in typically lower maximum total queuing and significantly fewer total vehicles in the system than doing no ramp metering control. These results are explained by, under the "no control" policy, the freeway is allowed to become significantly congested in several sections with on-ramps, the flow into those sections from the ramps becomes impossible. Thus, the conclusion is that active management of queues may result in spillback during heavy-volume and/or

incident conditions, as indicated in Figures 9-14, 9-34, and 9-55 for the MILOS method, but these spillback conditions are less severe than the corresponding spillback conditions for the "no control" case (as reflected in the maximum total queuing measure of performance). Plus, these queues are only created when necessary to maintain capacity flow on the freeway during incident conditions or during heavy-flow periods.

MILOS versus the locally traffic-responsive method

The locally traffic-responsive metering method being evaluated by ADOT is responsive to freeway conditions for light to moderate flows, but once the queue reaches the maximum storage capacity, the resulting rate exhibits strong oscillations between the maximum rate and the rate derived from the local freeway conditions, such as indicated in Figure 9-28. In test case #2, this oscillatory behavior becomes detrimental to total system performance since the maximum total vehicles in the system is actually increased over the no control case and the average speed on the freeway is actually reduced slightly.

MILOS is not susceptible to such oscillations, except when the freeway conditions become severely congested, such as during the incident of test case #3. When this occurs, MILOS must oscillate between restrictive rates to help dissipate the freeway congestion and very high flow rates to disperse the spilled-back ramp queue. Comparing such resulting metering rate profiles such as Figures 9-57 and 9-59 indicates that the oscillatory condition (low rate, high rate, low rate, high rate, etc. in consecutive minutes) occurs far less frequently for the MILOS method than the traffic-responsive method. This oscillatory condition is eliminated since in MILOS the area-wide optimization problem produces the high or low rates uniformly in the corridor based on the freeway conditions (i.e. if the freeway is severely congested, restrictive ramp metering must be enacted to reduce, and eventually remove, the congested condition(s)) but the local PC-RT rate regulation optimization problems are only afforded a limited range in which the rate can be modified to. Thus, at the local level, if the queue is reaching its capacity, or has already exceeded it, the local processor does not have the "authority" to switch the

rate to the saturation flow rate which would invalidate the coordinated decisions of what the rate should be made by the area-wide layer.

MILOS versus LP method

The performance of MILOS and LP methods are essentially equal in average freeway travel time and average freeway speed, but MILOS significantly outperforms LP in terms of total queuing time (13-40% less queuing time, 15-32% lower maximum total vehicles in queues), and recovery time (8-12% earlier). These results are due to two factors. First, the MILOS algorithm distributes queuing at the area-wide level among several ramps according to the interchange cost parameters. The LP approach may grow excessive queues at one ramp and allow heavy on-ramp volume at others, somewhat arbitrarily, thus contributing to larger maximum vehicles, higher queue times, and slower recovery. In addition, the PC-RT rate regulation layer of the MILOS hierarchy, not included in the LP metering method, makes more "intelligent" metering decisions at the local, real-time layer, than LP as illustrated in Figures 10-1 and 10-2.

Figure 10-1 indicates the metering rates applied by LP and MILOS during an iteration of test case #1. Figure 10-2 then compares sections A and B, as noted in Figure 10-1, side-by-side for the two metering algorithms. Notice in Figure 10-1 that MILOS is in "incident management mode" for a shorter amount of time than LP remains in this mode, according to the demand estimates and queues built up by the two systems. MILOS is able to apply more restrictive metering rates during the incident management because the active metering of the PC-RT rate regulation method earlier in the simulation run was able to dissipate the unnecessary queues at the ramps. The LP method, without such a pro-active real-time control mechanism, cannot do this. As a result, MILOS can return to the "Normal" metering condition sooner than LP, as shown in Figure 10-1.

As shown in these figures, LP has the tendency to build unnecessary queues at ramps when it estimates the upstream ramp demand lower than its true value if the current demand has fluctuated lower than its average value, since its new estimate of the demand

to the ramp for the next time-horizon (20-minutes) is based on this flow estimate. In contrast, MILOS uses the SPC-based anomaly detection scheme to identify each minute whether or not the current fluctuation is anomalous and plan metering rates that react to such a fluctuation, but only local to this ramp. The area-wide estimate of the average demand does not change in MILOS unless the outer control limit is breached at least for two consecutive observation periods (2 minutes). Thus, although MILOS changes the metering rates more frequently than the LP metering method, as shown in these figures, the changes on an area-wide basis, relative to each other, are fewer than the LP method, since the area-wide coordination problem is re-solved less frequently in MILOS than in the LP method. Hence, less total queuing and faster response time results with use of the MILOS method over LP since the demand estimation procedure of MILOS (two-level SPC-based anomaly detection) does not over-react to the stochastic fluctuations of the underlying demand processes (i.e. LP's "no level" re-solve of the area-wide coordination problem each 5-minutes).

Summary

This research has presented a hierarchical structure for solution of the large-scale freeway management problem that is constructed to address the key features of the large-scale freeway management problem (dynamic state changes, stochasticity, multi-dimensionality, unpredictability, partial-observability, and existence of multiple objectives). This hierarchical structure decomposes the freeway control problem into subproblems along temporal/spatial boundaries as appropriate. Although hierarchical treatment of freeway management is not new, the specific hierarchy proposed in this project is novel, in particular the identification of subnetworks from a large-scale freeway system and the basis for interaction between the area-wide layer and the locally traffic-responsive layer is entirely new.

A popular and useful macroscopic model of freeway traffic flow was modified to more accurately represent the ramp-freeway interface under the presence of congestion indicating that queues can develop at ramp entrances when the freeway density is too

high for those vehicles to merge into the system. The area-wide coordination component of the hierarchical control system considers the impact of queue growth on the adjacent interchanges in the optimization model. This optimization model is based on models available in the literature but incorporates several novel additions: (1) a new multi-criterion objective function and trade-off structure based on the congestion level of each adjacent interchange, (2) an alternative treatment of queue growth constraints, including both physical queue storage size and the time horizon of the optimization problem, and (3) modeling of demands from surface-street interchange flows. In addition, this formulation of the area-wide optimization problem is guaranteed to be have a feasible solution, unlike previous formulations.

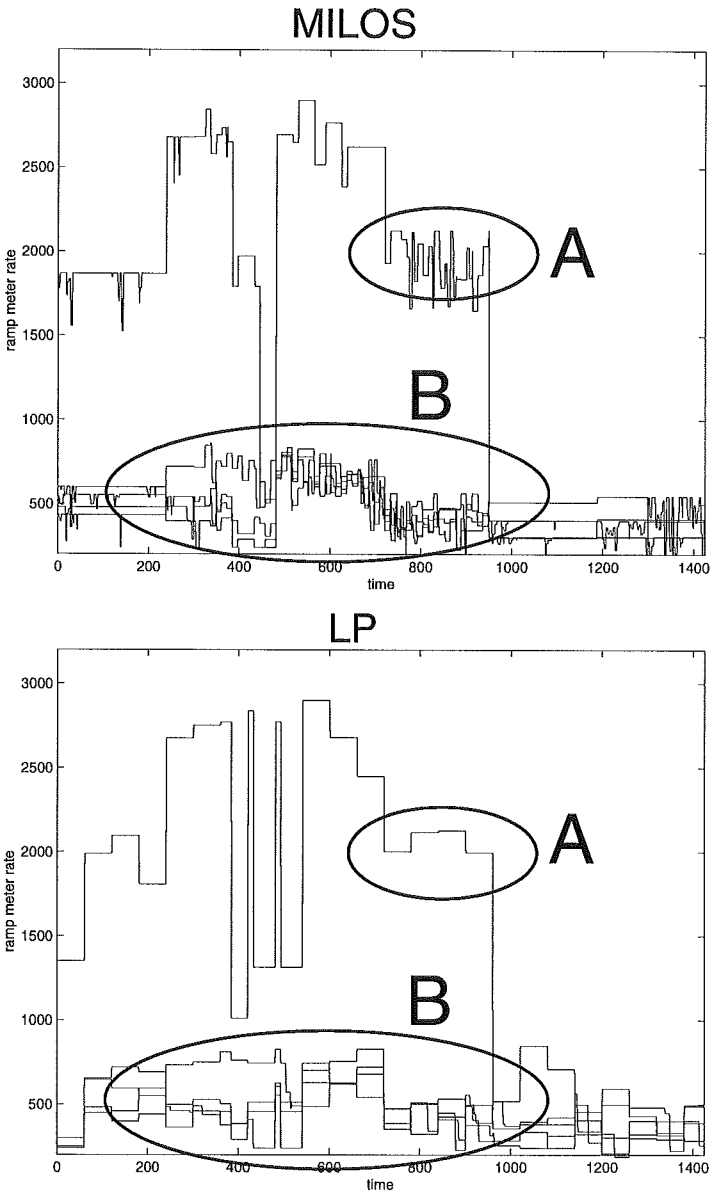


Figure 10- 1. Comparison of typical metering rates, MILOS and LP

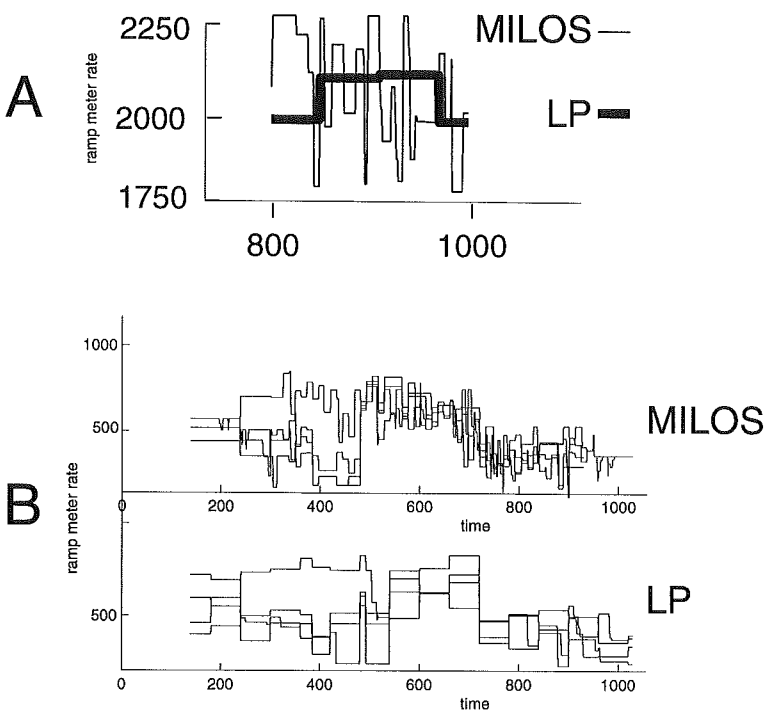


Figure 10- 2. Side-by-side comparison of metering rates for LP and MILOS

The solution of the area-wide coordination problem is then modified in real-time by the locally traffic-reactive, PC-RT rate regulation algorithm. The PC-RT optimization formulation increases the capacity of the freeway/surface-street interface by pro-actively planning opportunities to disperse queues or hold back additional vehicles when appropriate to do so. Although the basis for this optimization model is not new (i.e. linearization of the nonlinear macroscopic flow model of Chapter 4), the formulation of the scenario-based linear-programming problem is entirely new and its interaction with the SPC-based anomaly detection method is completely novel. The link to the solution of the area-wide coordination problem of Chapter 5 using the dual information is entirely new for dissimilar optimization problem structures (i.e. the area-wide optimization problem is an input-output QP, the PC-RT rate regulation optimization problem is a dynamic difference equation LP).

The SPC-based demand/flow monitoring system is a new method for management of stochastic variation in dynamic control systems. This flow monitoring system re-schedules the optimization of the area-wide coordination problem and the PC-RT rate

regulation optimization problems according to the underlying statistics of the demand processes. The SPC-based anomaly detection scheme can also be used to detect changes to the route-proportional matrices of the freeway network, vital for solution of the area-wide coordination problem.

A simulation experiment was presented that evaluated the MILOS hierarchical system against "no control", a locally traffic-responsive ramp metering policy currently being evaluated by ADOT, and a policy to resolve an area-wide LP coordination problem in 5-minute intervals (with a 20-minute time horizon) on a relatively small, but realistic, freeway management problem in the metropolitan Phoenix, Arizona. Three test cases were presented for a "burst" of heavy-volume, a 3-hour commuting peak, and a 3-hour commuting peak with a 30-minute incident in the middle of the network. The performance results indicated that MILOS is able to reduce freeway travel time, increase freeway average speed, and improve recovery performance of the system when flow conditions become congested.

Directions for future research

There are several promising areas for further development of the MILOS hierarchy;

(1) The *subnetwork identification* module needs development to construct a method for defining the subnetwork decomposition structure. Such a method would allow automatic reaction to changes in the prevailing network conditions over long time scales.

(2) The area-wide coordination optimization problem relies on the provision of a route-proportional matrix. Integration and testing of MILOS with a route-proportional matrix estimation routine is required to transition MILOS to a deployable system for real-world ramp metering.

(3) Currently, MILOS does not consider diversion of vehicles from lengthy ramp queues. A diversion model should be integrated and tested with MILOS to provide accurate flow predictions.

(4) MILOS assumes that the ramp demand can be obtained exactly. An estimation method for deriving the ramp demand from the interchange flows and interchange turning probabilities should be included in MILOS.

(5) In the PC-RT subproblem optimization formulation, the linearized macroscopic freeway flow model is used to predict freeway flow dynamics over five to seven minute time scales for pro-active metering rate determination. The range of descriptive accuracy of the linear approximation to the nonlinear macroscopic model should be studied in more detail to analytically derive the upper and lower bounds of the density constraints.

(6) Analysis of real-world detector data time-series is necessary to analytically derive the parameter $\theta$and further develop the form of the function $f(\overline{R},T)$ in the SPC anomaly detection module. In addition, updates to θ would require further analysis of the structure of detector data time-series. Some measure of "acceptable performance" in simulation experiments should be chosen as the initial goal of choosing of the inner control limit specification $A'_2 = \theta A_2$ in the SPC anomaly detection module.

(7) Another direction of future work on the MILOS ramp metering control algorithm is to test the strategy with real-world data and/or implement the scheme in a real-world freeway subnetwork.

APPENDIX A: Route-proportional matrices

Route proportional matrices for test case #1

Time period 1
OD = [1,0.69,0.614,0.614,0.528,0.475,0.304,0.304,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.448,0.448,0.392,0.392,0.392;...
0,0,0,0,0.756,0.688,0.405,0.405,0.354,0.354,0.354;...
0,0,0,0,0,0.819,0.446,0.446,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.802,0.802,0.802;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 2
OD = [1,0.71,0.66,0.66,0.594,0.535,0.417,0.417,0.375,0.375,0.375;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.865,0.767,0.586,0.586,0.525,0.525,0.525;...
0,0,0,0,0.817,0.72,0.543,0.543,0.486,0.486,0.486;...
0,0,0,0,0,0.809,0.569,0.569,0.505,0.505,0.505;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 3
OD = [1,0.72,0.677,0.677,0.616,0.548,0.4,0.4,0.36,0.36,0.36;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.878,0.771,0.548,0.548,0.491,0.491,0.491;...
0,0,0,0,0.835,0.727,0.51,0.51,0.456,0.456,0.456;...
0,0,0,0,0,0.796,0.513,0.513,0.455,0.455,0.455;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 4
OD = [1,0.72,0.677,0.677,0.616,0.548,0.4,0.4,0.36,0.36,0.36;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.878,0.771,0.548,0.548,0.491,0.491,0.491;...
0,0,0,0,0.835,0.727,0.51,0.51,0.456,0.456,0.456;...
0,0,0,0,0,0.796,0.513,0.513,0.455,0.455,0.455;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

otherwise
OD = [1,0.72,0.677,0.677,0.616,0.548,0.4,0.4,0.36,0.36,0.36;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.878,0.771,0.548,0.548,0.491,0.491,0.491;...
0,0,0,0,0.835,0.727,0.51,0.51,0.456,0.456,0.456;...
0,0,0,0,0,0.796,0.513,0.513,0.455,0.455,0.455;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Route proportional matrices for test cases #2 and #3

Time period 1
OD = [1,0.69,0.614,0.614,0.528,0.475,0.304,0.304,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.448,0.448,0.392,0.392,0.392;...
0,0,0,0,0.756,0.688,0.405,0.405,0.354,0.354,0.354;...
0,0,0,0,0,0.819,0.446,0.446,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.802,0.802,0.802;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 2
OD = [1,0.69,0.624,0.624,0.528,0.475,0.284,0.284,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.478,0.478,0.392,0.392,0.392;...
0,0,0,0,0.756,0.688,0.405,0.405,0.334,0.334,0.334;...
0,0,0,0,0,0.819,0.446,0.446,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.762,0.762,0.762;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 3
OD = [1,0.69,0.614,0.614,0.528,0.495,0.304,0.304,0.217,0.217,0.217;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.428,0.428,0.392,0.392,0.392;...
0,0,0,0,0.756,0.688,0.405,0.405,0.334,0.334,0.334;...
0,0,0,0,0,0.819,0.446,0.446,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.892,0.892,0.892;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];

Time period 4

```
OD = [1,0.69,0.614,0.614,0.528,0.475,0.304,0.304,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.715,0.468,0.468,0.322,0.322,0.322;...
0,0,0,0,0.756,0.688,0.405,0.405,0.354,0.354,0.354;...
0,0,0,0,0,0.819,0.446,0.446,0.355,0.355,0.355;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.782,0.782,0.782;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 5

```
OD = [1,0.71,0.62,0.62,0.594,0.535,0.417,0.417,0.375,0.375,0.375;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.865,0.767,0.556,0.556,0.525,0.525,0.525;...
0,0,0,0,0.817,0.72,0.543,0.543,0.516,0.516,0.516;...
0,0,0,0,0,0.809,0.569,0.569,0.505,0.505,0.505;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 6

```
OD = [1,0.71,0.66,0.66,0.594,0.535,0.467,0.467,0.375,0.375,0.375;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.865,0.767,0.586,0.586,0.525,0.525,0.525;...
0,0,0,0,0.817,0.72,0.543,0.543,0.426,0.426,0.426;...
0,0,0,0,0,0.809,0.569,0.569,0.505,0.505,0.505;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.833,0.833,0.833;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 7

```
OD = [1,0.71,0.66,0.66,0.594,0.535,0.437,0.437,0.375,0.375,0.375;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.865,0.787,0.586,0.586,0.525,0.525,0.525;...
0,0,0,0,0.817,0.72,0.563,0.563,0.486,0.486,0.486;...
0,0,0,0,0,0.809,0.569,0.569,0.525,0.525,0.525;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.803,0.803,0.803;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 8

```
OD = [1,0.72,0.677,0.677,0.616,0.518,0.4,0.4,0.36,0.36,0.36;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.878,0.771,0.518,0.518,0.491,0.491,0.491;...
0,0,0,0,0.835,0.727,0.51,0.51,0.456,0.456,0.456;...
0,0,0,0,0,0.796,0.483,0.483,0.455,0.455,0.455;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.823,0.823,0.823;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 9

```
OD = [1,0.82,0.677,0.677,0.616,0.528,0.38,0.38,0.36,0.36,0.36;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.858,0.771,0.548,0.548,0.471,0.471,0.471;...
0,0,0,0,0.855,0.727,0.51,0.51,0.456,0.456,0.456;...
0,0,0,0,0,0.856,0.513,0.513,0.425,0.425,0.425;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.783,0.783,0.783;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 10

```
OD = [1,0.79,0.714,0.714,0.628,0.575,0.404,0.404,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.448,0.448,0.492,0.492,0.492;...
0,0,0,0,0.756,0.688,0.415,0.415,0.354,0.354,0.354;...
0,0,0,0,0,0.819,0.476,0.476,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.832,0.832,0.832;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

Time period 11

```
OD = [1,0.69,0.614,0.614,0.508,0.475,0.304,0.304,0.267,0.267,0.267;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,1,0.815,0.725,0.458,0.458,0.392,0.392,0.392;...
0,0,0,0,0.756,0.688,0.405,0.405,0.354,0.354,0.354;...
0,0,0,0,0,0.819,0.546,0.546,0.385,0.385,0.385;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,1,0.782,0.782,0.782;...
0,0,0,0,0,0,0,0,0,0,0;...
0,0,0,0,0,0,0,0,0,1,1;...
0,0,0,0,0,0,0,0,0,0,0];
```

References

B. Allen and G. Newell, (1976) "Some issues relating to metering or closing of freeway ramps, Part I and II", Transportation Science, vol. 10, no. 3, 243-268.

E. Arnold, (1987) "Changes in travel in the Shirley Highway Corridor 1983-1986", Virginia Transportation Research Council, 43.

Y. Asakura, (1995) "LP type dynamic on-ramp traffic control model for urban expressway", Proceedings of 4th International Conference on Applications of Advanced Technologies in Transportation Engineering, June 27-30, 434-438.

K. Ashok and M. Ben-Akiva , (1993) "Dynamic origin-destination matrix estimation for real-time traffic management systems", *Transportation and Traffic Theory*. Ed. C. Daganzo. Elsevier Science, 465-484.

J. Banks, (1990) "Freeway speed-flow-concentration relationships: More evidence and interpretations", Transportation Research Record, 1225, 53-60.

J. Banks, (1992) "Effect of response limitations on traffic-responsive ramp metering", Transportation Research Record, 1394, 17-25.

J. Banks, (1996) "Another look at apriori relationships among traffic flow characteristics", Transportation Research Record, 1510, 1-10.

M. Bazaraa, H. Sherali, and C. Shetty, (1993) *Nonlinear programming theory and algorithms*, 2nd edition, Wiley & Sons, New York, 385-392.

J. Bender, (1991) "An overview of systems studies of automated highway systems", IEEE Transactions on Vehicular Technology, Vol. 40, No. 1, 82-99.

L. Benmohomed and S. Meerkov, (1994) "Feedback control of highway congestion by a fair on-ramp metering", Proceedings of 33rd conference on decision and control, Lake Beuna Vista, FL, 1994. 2437-2442.

D. Berg, (1990) "Predictive Algorithm Improvements for a Real-time Ramp Control System", ed. N. L. Nihan, ITE 1990 Compendium of Technical Papers.

J. Bernassou, and A. Titli, (1982) *Interconnected Dynamical Systems: Stability, Decomposition, and Decentralization*. North-Holland, Amsterdam.

C. Blumentritt, et al., (1981) "Guidelines for selection of ramp control systems", NCHRP report 232, Transportation Research Board.

K. Brewer, J. Buhr, D. Drew, and C. Messer, (1969) "Ramp Capacity and Service Volume as Related to Freeway Control", Highway Research Record, 279, 70-86.

D. Brubaker, (1992) "Fuzzy-Logic System Solves Control Problem", EDN, 121-126.

J. Buhr, et al., (1969) "A Moving Vehicle Merging Control System", Highway Research Record, 279, 121-136.

C. Carlson and C. Glen, (1978) "Ramp control on I-35E: Review of operational experience 1970-1977", Minnesota Department of Transportation, Traffic Engr. Section, 35.

G-L. Chang, P. Ho, and C. Wei, (1992) "A Dynamic System-optimum Control Model for Commuting Traffic Corridors", 71st Annual Meeting Transportation Research Board, Washington, DC

G-L. Chang, J. Wu, and S. Cohen, (1994) "An Integrated Real-Time Ramp Metering Model for Non-Recurrent Congestion: Framework and Preliminary Results", 73rd Annual Meeting of the Transportation Research Board, Washington, DC

V. Chankong and Y. Haimes. (1983) *Multiobjective Decision Making: Theory and Methodology*. North-Holland Series on System Science and Engineering, Volume 8., A. Sage (editor). North-Holland.

C. Chen, et al., (1974) "Entrance ramp control for travel rate maximization on expressway", Transportation Research, Vol. 8, 503-508.

B. Coifman, (1996) "New methodology for smoothing freeway loop detector data: Introduction to digital filtering", Transportation Research Record, 1554, 142-152.

K. Courage, (1968) "A Freeway Corridor Surveillance, Information, and Control System", Research Report, 438-8, Texas Transportation Institute.

CPLEX, (1996) *CPLEX 3.0 users manual.*

M. Cremer and H. Keller, (1987) "A New Class of Dynamic Methods for the Identification of Origin-Designation Flows", Transportation Research, 21B, 117-132.

M. Cremer, S. Schoof, and J. Perrin, (1990) "On control strategies for urban traffic corridors", Control, Computers, and Communications in Transportation: selected papers from IFAC/IFORS/IFIP Symposium. Paris, France. pp. 213-219.

G. Davis, (1993) "Estimating Freeway Demand Patterns and Impact of Uncertainty on Ramp Controls", Journal of Transportation Engineering, 119, 4, 489-503.

G. Dantzig and P. Wolfe, (1961) "The decomposition algorithm for linear programming", Econometrica, Vol. 9, No. 4.

J. Decker, (1998) private communication. ADOT Technical Advisory Committee meeting, April 15, 1998.

Y. Ding, P. Mirchandani, and S. Nobe (1997), "Prediction of network loads based on origin-destination synthesis from observed link volumes", Transportation Research Record, 1607, 95-104.

D. Drew, K. Brewer, J. Buhr, and R. Whitson, (1969) "Multilevel Approach to the Design of a Freeway Control System", Highway Research Record, 157, 40-55.

D. Drew, et al., (1966) "The Development of an Automatic Freeway Merging Control System", Research Report, Texas Transportation Institute, 24-19.

G. Euler, (1990) "Intelligent vehicle/highway systems: Definitions and applications", ITE Journal, November, 17-22.

W. Fan and E. Asmussen, (1990) "A Hierarchical Computerized Adaptive Control Strategy for the Freeway System", Scientific Publication, VK7703.301, Delft University of Technology, Department of Transportation Planning and Highway Engineering, Traffic Safety Division.

FHWA, (1985). *Highway Capacity Manual.* US Government Printing Office. Washington, DC.

N. Gartner, (1983) "OPAC: A demand-responsive strategy for real-time traffic signal control", Transportation Research Record, 906, 75-81.

D. Gettman, (1998) "A multiobjective integrated large-scale optimized ramp metering control system for freeway/surface-street traffic management", Ph.D. Thesis. University of Arizona.

N. Goldstein and K. Kumar, (1982) "A Decentralized Control Strategy for Freeway Regulation", Transportation Research, Part B: Methodological, Vol. 16B, No. 4, 279-290.

R. Gordon, (1996) "Algorithm for controlling spillback from ramp meters", Transportation Research Record, 1554, 162-171.

J. Gray, P. Lavallee, and A. May, (1990) "Segment-Wide On-Line Control of Freeways to Relieve Congestion and Improve Public Safety", Final Report, FHWA/CA/UCB-ITS-RR-90-11, Institute for Transportation Studies, University of California, Berkeley.

K. Haboian, (1996) "A case for freeway mainline metering", Transportation Research Record, 1494, 11-20.

H. Haj-Salem, M. Papageorgiou, M. Blosseville, (1990) "ALINEA: A local Feedback Control Law for On-Ramp Metering - A Real Life Study", 3rd IEE International Conference on Road Traffic Control, London, U.K., 194-198.

Y. Haimes, (1990) *Hierarchical multiobjective analysis of large-scale systems.* Hemisphere Publishing.

M. Hallenbeck and J. Nisbet, (1993) "Freeway and Arterial Integrated Control System", Final technical report GC8719, Task 8, Washington State Transportation Center.

B. Han and R. Reiss, (1994) "Coordinating Ramp Meter Operation with an Upstream Intersection Traffic Signal", 73rd Annual Meeting of the Transportation Research Board, Washington, DC.

K. Head, P. Mirchandani, and D. Sheppard, (1992) "Hierarchical framework for real-time traffic control", Transportation Research Record, 1360, 82-88.

B. Hellinga and M. Van Aerde (1997) "Examining the potential of using ramp metering as a component of an ATMS", Transportation Research Record, 1494, 75-83.

J. Hibbard, et al., (1990) "An Overview of the Ramp Metering Subsystem for the Phoenix Freeway Management System", ITE 1990 Compendium of Technical Papers, 51-54.

W. Hines and D. Montgomery, (1990) *Probability and Statistics in Engineering and Management Science.* 3rd Edition, John Wiley & Sons, 580-602.

A. Hobeika, R. Sivanandan, S. Subramaniam, K. Ozbay, and Y. Zhang, (1993) "Real-Time Traffic Diversion Model: Conceptual Approach", Journal of Transportation Engineering, 119, 4, 515-535.

V. Hurdle and P. Datta, (1982) "Speeds and flows on an urban freeway: Some measurements and a hypothesis", Transportation Research Board, 905, 127-137.

L. Isaksen and H. Payne, (1973) "Suboptimal control of linear systems by augmentation with application to freeway traffic regulation", IEEE Transactions on Automatic Control, Vol. AC-18, No. 3, 210-219.

L. Jacobson, (1989) "Integrated Freeway/Arterial Management in Washington State", Engineering Foundation Conference, Santa Barbara, California, 199-214.

L. Jacobson, H. Kim, and O. Meyhar, (1989) "A Real-Time Metering Algorithm for Centralized Control", Transportation Research Record, 1232, 17-26.

H. Ji, (1996) "Freeway traffic systems: Prediction and control", Proceedings of 36th conference on decision and control, 1996. 1815-1819.

S. Kahng, C. Jeng, J. Campbell, and A. May, (1984) "Segment-wide Traffic Responsive Freeway Entry Control: Freeway Corridor Modeling, Control Strategy, and Implementation Plan", Final Report FHWA/CA/TO-84-5, U.S. Department of Transportation, Washington, DC.

U. Karaaslan, P. Varaiya, and J. Walrand, (1990) "Two proposals to improve freeway traffic flow", Proceedings of 1990 American Control Conference, Boston, MA. Vol. 3, 2539-2544.

D. Kirk, (1970) *Optimal Control Theory: An Introduction.* Prentice-Hall.

E. Kwon, (1991) "A New Approach for Real-Time Prediction of Traffic Demand-Diversion in Freeway Corridors", Applications of Advanced Technologies to Transportation Engineering, Ed. Stephanedes and Sinha, Minneapolis, MN, ASCE Press.

L. Lapidus and R. Luus, (1967) *Optimal Control of Engineering Processes.* Blaisdell Publishing, Waltham, MA.

M. Lighthill and G. Whitham, (1955) "On Kinematic Waves II: A Theory of Traffic Flow on Long Crowded Roads", Proceedings of the Royal Society, London A229, 317 - 345.

J. Lindley, (1987) "Urban freeway congestion: quantification of the problem and effectiveness of potential solutions", ITE Journal, January, 27-32.

J. Lindley, (1987) "A methodology for quantifying urban freeway congestion", Transportation Research Record, 1132, 1-7.

L. Lipp, L. Corcoran, G. Hickman, (1991) "Benefits of Central Computer Control for Denver Ramp-Metering System", Transportation Research Record, 1320, 3-6.

D. Looze, P. Houpt, N. Sandell, and M. Athans, (1978) "On Decentralized Estimation and Control with Application to Freeway Ramp Metering", IEEE Transactions on Automatic Control, AC-23, 2, 268-275.

S. Madanat, S. Hu, and J. Krogmeier, (1996) "Dynamic estimation and prediction of freeway O-D matrices with route switching considerations and time-dependent model parameters", Transportation Research Record, 1537, 98-104.

M. Mahmoud, (1977) "Multilevel systems control and applications: a survey", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-7, No. 3, 125-142.

B. Marsden, (1981) "Ramp meters and travel quality in Austin, Texas", Texas Department of Highways and Public Transportation, 27.

A. May, (1979) "A Dynamic Freeway Control System Hierarchy", Engineering Foundation Conference on Research Directions in Computer Control of Urban Traffic Systems, Pacific Grove, CA, 287-310.

A. May, (1987) "Freeway Simulation Models Revisited", Transportation Research Record, 1132, 94-99.

W. McShane and R. Roess, (1990) *Traffic Engineering.* Prentice-Hall, Inc. Englewood Cliffs, NJ.

C. Messer, (1969) "A design and synthesis of a multi-level freeway-control system and a study of its operational control plan". Ph.D. Dissertation, Texas A&M University, UMI #70-11561.

P. Michalopoulos, (1986) "Integrated Modeling of Freeway Flow and Application to Microcomputers", Traffic Engineering and Control, Vol. 27, No. 4, 198-208.

P. Michalopoulos, P. Yi, and A. Lyrintzis, (1993) "Continuum Modeling of Traffic Dynamics for congested freeways", Transportation Research, Part B. Vol 27B, No. 4, 315-332.

F. Middelham and S. Smulders, (1991) "Isolated Ramp Metering: Real-Life Study in the Netherlands", Technical Report. DRIVE-Office, Brussels, Belgium.

Mobility 2000, (1989) *Intelligent highway and vehicle systems: a report of Mobility 2000.* San Antonio, TX, ed. W. Harris and G. Bridges, TTI, Feb. 15-17.

J. Moore and P. Jovanis, (1985) "Statistical designation of traffic control subareas", Transportation Engineering, Vol. 11, No. 3, 208-223.

K. Moskowitz, (1965) "Analysis and projection of research on traffic surveillance, communication, and control", NCHRP report no. 87, Transportation Research Board.

L. Newman, et al., (1970) "An Evaluation of Ramp Control on the Harbor Freeway in Los Angeles", Highway Research Record, 303, 44-55.

N. Nihan, (1991) "Predictive Algorithm Improvements for a Real-Time Ramp Control System", Technical Report WA-RD 213.1, Washington State Department of Transportation.

N. Nihan and D. Berg, (1992) "A Predictive Algorithm for a Real-Time Ramp Control System", ITE Journal, 29-32.

I. Okutani, (1987) "The kalman filtering approach in some transportation and traffic problems", Proceedings of 10th International Symposium on transportation and traffic theory, Cambridge, MA, 139-158.

M. Papageorgiou and G. Schmidt, (1980) "On the hierarchical solution of nonlinear optimal control problems", Large Scale Systems, Vol. 1, 265-271.

M. Papageorgiou, (1980) "A New Approach to Time-of-Day Control Based on a Dynamic Freeway Traffic Model", Transportation Research, Vol. 14B, 349-360.

M. Papageorgiou and R. Mayr, (1982) "Optimal decomposition methods applied to motorway traffic control", International Journal of Control, Vol. 35, No. 2, 269-280.

M. Papageorgiou, (1983) *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control*, Springer-Verlag, Berlin.

M. Papageorgiou, (1983) "A Hierarchical Control System for Freeway Traffic", Transportation Research, Vol. 17B, 251-261.

M. Papageorgiou, (1984) "Multilayer Control System Design Applied to Freeway Traffic", IEEE Transactions on Automatic Control, Vol. AC-29, No. 6, 482-490.

M. Papageorgiou, H. Haj-Salem, J. Blosseville, N. Bhouri, and J. Perrin, (1989) "Macroscopic Modeling of Traffic Flow on the Boulevard Peipherique in Paris", Transportation Research, Vol. 23B, 29-47.

M. Papageorgiou, H. Haj-Salem, J. Blosseville, N. Bhouri, and J. Perrin, (1990) "Modeling and Real-Time control of Traffic Flow on the Boulevard Peripherique in Paris", IFAC Control, Computers, Communication in Transportation, Vol. 12, 205-211.

M. Papageorgiou, H. Hadj-Salem, and J. Blosseville, (1991) "ALINEA: A Local Feedback Control Law for On- Ramp Metering", Transportation Research Record, 1320, 58-64.

M. Papageorgiou, A. Messmer, J. Azema, and D. Drewanz, (1995) "A neural network approach to freeway network traffic control", Control Engineering Practice, Vol. 3, No. 12, 1719-1726.

M. Papageorgiou, (1995) "An integrated control approach for urban traffic corridors", Transportation Research, Part C, Vol. 3C, No. 1, 19-30.

H. Payne, (1971) "Models of Freeway Traffic and Control", Mathematical Models of Public Systems, Ed. G. Bekey, Vol. 1., No. 1, 51-61.

H. Payne, (1979) "FREFLO: A macroscopic Simulation Model of Freeway Traffic.", Transportation Research Record, 722, 68-77.

H. Payne, D. Brown, J. Todd, (1985) "Demand-Responsive Strategies for Interconnected Freeway Ramp Control Systems, Volume I: Metering Strategies", FHWA/RD-85/109, U.S. Department of Transportation, Washington, DC.

H. Payne and W. Thompson, (1974) "Allocation of Freeway Ramp Metering Volumes to Optimize Corridor Performance", IEEE Transactions on Automatic Control, AC-19, 3, 177-186.

F. Pooran and R. Sumner, (1992) "Coordinated Operation of Ramp Metering and Adjacent Traffic Signal Control Systems. Volume II: Executive summary", FHWA-RD-92-088, Farradyne Systems, Inc., U.S. Department of Transportation, Washington, DC.

F. Pooran, J. Farhad, and H. Lieu, (1994) "Evaluation of System Operating Strategies for Ramp Metering and Traffic Signal Coordination", Proceedings of the 4th Annual meeting of IVHS America, Atlanta, GA., Vol. 2, 528-534.

D. Powell, (1997) Private Communication. ADOT technical advisory committee meeting, September 4, 1997.

H. Rahka and M. Van Aerde, (1997) "Statistical analysis of day-to-day variations in real-time traffic flow data", Transportation Research Record, 1510, 26-34.

S. Rajan, J. Blackburn, J. Lien, and V. Subbarao, (1986) "On-Ramp Traffic Control on the Black Canyon Freeway", Final Report, FHWA/AZ (86-211) Center for Advanced Research in Transportation, Arizona State University.

A. Rathi, E. Lieberman, and M. Yedlin, (1985) "Enhanced FREFLO: Modeling of congested environments", Transportation Research Record, 1112, pp 61-70.

D. Robertson, I. Dennis, and R. Bretherton, (1991) "Optimizing networks of traffic signals in real-time: the SCOOT method", IEEE Transactions on Vehicular Technology, VT-40, No. 1, 11-15.

P. Richards, (1956) "Shock waves on the highway", Operations Research, 42-51.

N. Sandell, et. al, (1978) "Survey of decentralized control methods for large scale systems", IEEE transactions on automatic control, AC-23, no. 2, 108-128

T. Sasaki and T. Akiyama, (1987) "Fuzzy On-Ramp Control Model on Urban Expressway and Its Extension", Transportation and Traffic Theory, Ed. N. Gartner and H. Wilson, Elsevier, New York, NY.

S. Sen and K. Head, (1997) "Controlled optimization of phases at an intersection", Transportation Science, 31, 5-17.

D. Siljak, (1991) *Decentralized Control of Complex Systems.* Mathematics in Science and Engineering Series, Vol. 184, W. Ames (editor), Academic Press.

B. Smith and M. Demetsky, (1996) "Multiple-interval freeway traffic flow forecasting", Transportation Research Record, 1554, 136-141.

Y. Stephanedes, E. Kwon, and P. Michalopoulos, (1989) "Demand Diversion for Vehicle Guidance, Simulation, and Control in Freeway Corridors", Transportation Research Record, 1220, 12-20.

Y. Stephanedes, E. Kwon, and K. Chang, (1992) "Control emulation method for evaluating and improving traffic-responsive ramp metering strategies" Transportation Research Record, 1360, 42-45.

Y. Stephanedes and A. Chassiakos, (1992) "Comparative Performance Evaluation of Incident Detection Algorithms" Transportation Research Record, 1360, 50-57.

Y. Stephanedes and K-K. Chang, (1991) "Optimal Ramp Metering Control for Freeway Corridors", Applications of Advanced Technologies in Transportation Engineering, Ed. Y. Stephanedes and C. Sinha, Minneapolis, MN, ASCE Press.

Y. Stephanedes and K.-K. Chang, (1993) "Optimal Control of Freeway Corridors", Journal of Transportation Engineering, Vol. 119, No. 4, 504-513.

Technical Advisory Committee, (1998) Personal communication. RHODES-ITMS corridor control progress meeting. Arizona Department of Transportation, Phoenix, AZ. April 16, 1998.

USDOT, (1976) *Traffic Control Systems Handbook*, U.S. Government Printing Office, Washington, DC.

USDOT, (1992) *IVHS strategic plan: report to congress.* U.S. Government Printing Office, Washington, DC.

M. Van Aerde and S. Yagar, (1988) "Dynamic Integrated Freeway/Traffic Signal Networks: Problems and Proposed Solutions", Transportation Research Part A, Vol. 22A, No. 6, 435-443.

M. Van Aerde, et al., (1987) "A Review of Candidate Freeway-Arterial Corridor Traffic Models", Transportation Research Record, 1132, 53-65.

C. Wang, (1972) "On a Ramp-Flow Assignment Problem", Transportation Science, Vol. 6, 114-130.

J. Wang and A. May, (1973) "Computer Model for Optimal Freeway On-Ramp Control", Highway Research Record, 469.

J. Wattleworth and D. Berry, (1967) "Peak Period Analysis and Control of a Freeway System", Highway Research Record, 157, 1-21.

J. Wattleworth and K. Courage, (1968) "An Evaluation of Two Types of Freeway Control Systems", Technical Report 4888-6, Texas Transportation Institute.

E. Weits, (1988) "Stochastic model of traffic flows on freeways", Report PB90-175753. Center for Mathematics and Computer Science, Amsterdam, Netherlands.

I. Wilson, (1979) "Foundations of hierarchical control", International Journal of Control, Vol. 29, 899-933.

J. Wright, (1993) "Twin cities integrated traffic management system", Proceedings of 1993 annual meeting of IVHS America, Session 18.

S. Yagar, (1989) "Metering Freeway Access", Transportation Quarterly, Vol. 43, No. 2, 215-224.

H. Yang and S. Yagar, (1994) "Some developments in traffic control of freeway-arterial corridor systems", IFAC transportation systems conference, Tianjin, PRC. 293-298.

T. Yoshino, T. Sasaki, and T. Hasegawa, (1991) "Traffic Control System on the Hanshin Expressway", ASCE 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, MN, 288-292.

L. Yuan and J. Kreer, (1971) "Adjustment of Freeway Ramp Metering Rates to Balance Entrance Ramp Queues", Transportation Research, Vol. 5, No. 2, 127-133.

H. Zhang and S. Ritchie, (1995) "An integrated traffic responsive ramp control strategy via nonlinear state feedback", 74th Annual Meeting of Transportation Research Board, Washington, DC.

H. Zhang, S. Ritchie, and Z. Lo, (1995) "On the optimal control problem: when does ramp metering work?", 74th Annual Meeting of the Transportation Research Board, Washington, DC.

Y. Zhang and A. Hobeika, (1997) "Diversion and signal retiming for a corridor under incident conditions", Proceedings of 77th Annual meeting of Transportation Research Board, Wasington, DC.