Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/TX-94/1232-24		

4. Title and Subtitle	5. Report Date
AN ADVANCED REAL-TIME RAMP METERING SYSTEM (ARMS): THE SYSTEM CONCEPT	November 1994
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Jyh-Charn S. Liu, Junguk L. Kim, Yilong Chen, Ying Hao, Soojung Lee, Taeyoung Kim, and Michael Thomadakis	Research Report 1232-24

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Texas Transportation Institute The Texas A&M University System College Station, Texas 77843-3135	11. Contract or Grant No. Study No. 0-1232

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Texas Department of Transportation Research and Technology Transfer Office P. O. Box 5080 Austin, Texas 78763-5080	January 1993 - April 1994
	14. Sponsoring Agency Code

15. Supplementary Notes

Research performed in cooperation with the Texas Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration.
Research Study Title: Urban Highway Operations Research and Implementation Program

16. Abstract

This research report presents a three-level, highway ramp-metering control scheme. In the first level, ramp controllers distributively compute ramp metering rates based on system-wide information. The system adapts quickly to changing traffic conditions; it is modular and allows scalable and robust implementation. The O-D prediction algorithm is adaptive and very accurate. The second level consists of an optimal, self-learning, congestion predictor algorithm that may predict all short-term traffic flow-breakdowns. The algorithm, using self-learning, utilizes sequences of traffic patterns to improve its prediction accuracy. The third level is a congestion resolution scheme which overcomes many of the drawbacks of existing techniques. It balances the congestion resolution time with the service quality of the surface streets.

Simulation results show that application of the three-level control algorithm will improve the freeway and surface street service.

17. Key Words	18. Distribution Statement
Ramp Metering, Congestion Management, Traffic Flow-Breakdown Prediction, Optimization, Simulation	No restrictions. This document is available to the public through NTIS: National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	92	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

AN ADVANCED REAL-TIME RAMP METERING SYSTEM (ARMS): THE SYSTEM CONCEPT

by

Jyh-Charn S. Liu
Assistant Research Specialist
Texas Transportation Institute

Junguk L. Kim
Assistant Research Specialist
Texas Transportation Institute

and the
Following Research Assistants

Yilong Chen
Ying Hao
Soojung Lee
Taeyoung Kim
Michael Thomadakis

Research Report 1232-24
Research Study Number 0-1232
Research Study Title: Urban Highway Operations Research and Implementation Program

Sponsored by the
Texas Department of Transportation
In Cooperation with
U.S. Department of Transportation
Federal Highway Administration

November 1994

Texas Transportation Institute
The Texas A&M University System
College Station, Texas 77843-3135

IMPLEMENTATION STATEMENT

This report presents an Advanced Real-time Ramp Metering System (ARMS), which consists of three levels of control algorithms integrated for control of freeway ramp metering systems.

The first level control algorithm treats traffic as free-flow; ramp metering controllers make ramp metering decisions distributively, based on system-wide traffic flow information. Our scheme improves existing ramp metering schemes by its rapid adaptation to current traffic conditions. It is a modular control algorithm that allows for scalable and robust implementation. This approach effectively reduces the possibility of recurrent congestion and achieves a higher level of freeway service. The algorithm for prediction of O-D traffic distribution is quite accurate, and it adapts to traffic pattern changes quickly.

The second level control algorithm is an optimal, self-learning congestion predictor. This algorithm is able to predict short-term traffic flow-breakdown and prevent congestion. It is a stochastic, model based pattern-recognition technique which predicts imminent traffic congestion based on sequences of traffic patterns. Our approach improves the prediction accuracy over existing snapshot based bottleneck prediction techniques, since it can capture interdependencies of consecutive traffic states. The self-learning capability allows the congestion prediction algorithm to start from an arbitrary set of traffic patterns and improves the accuracy of prediction with results obtained from on-line operation.

The third level control algorithm is a congestion resolution scheme. This algorithm overcomes drawbacks of existing approaches by flexible balance of the congestion resolution time of the freeway, and the service quality of the surface streets. Simulation results show that application of the three-level control method will improve the freeway and surface street service. The proposed techniques enhance the capability of solving transportation problems. To support full implementation of our scheme, further research is required in the following areas: (1) Simulation testing of the free flow control scheme. The simulation work is needed to optimize the throughput-gain and congestion-risk functions. (2) Optimization of our congestion detection algorithms using actual traffic data. Based on the test results, we would be able to optimize the model parameters so that prominent traffic parameters can be reflected in the model. The proposed system provides a comprehensive solution that cohesively integrates different system elements including traffic control algorithms, resilient system architecture, and scalable deployment strategies for implementation of the next generation of freeway ramp metering systems.

DISCLAIMER

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 view or policies of the Texas Department of Transportation (TxDOT) or the Federal Highway Administration (FHWA). This report does not constitute a standard, specification, or regulation nor is it intended for construction, bidding, or permit purposes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

This report presents an *Advanced Real-time ramp Metering System*, ARMS, which consists of three levels of control algorithms integrated for flexible control of freeway ramp metering systems. The objective of ARMS is to provide real-time control of a system of ramp meters both to reduce the onset of congestion and to explicitly address congestion resolution.

The first level of control algorithm treats traffic as free-flow. We collect system-wide traffic flow information and then distributively compute metering decisions for a target area to optimize a system-wide objective function. Our free-flow control scheme differs from existing algorithms by taking into account the risk of congestion when the ramp metering rate (or freeway throughput, equivalently) increases. As a result, we reduce effectively the possibility of recurrent congestion caused by peak demand to achieve a better level of freeway service. The algorithm is modular, and thus scales easily for incremental implementation.

The objective of the second level of control algorithm is mainly to predict and prevent traffic flow breakdowns caused by traffic fluctuation. We use a stochastic, model based pattern recognition technique to capture sequences of traffic patterns for flow breakdown prediction. The prediction algorithm has self-learning capability: it begins with an arbitrary set of traffic patterns for congestion prediction, and the accuracy of congestion prediction can improve with the actual results obtained from on-line operation.

The third level control scheme is a congestion resolution scheme which can balance the congestion resolution time and ramp input rates. The dynamic congestion resolution algorithm overcomes the drawbacks of existing approaches by taking into account the traffic conditions on both the freeway and surface streets. It resolves congestion in a flexible manner based on the management criteria.

1 INTRODUCTION

We can broadly divide the ramp metering techniques into three categories: *pretimed*, *traffic-responsive*, and *predictive*. In pretimed schemes, the time of day, weather condition, and historical traffic data determine the ramp metering rates [9]. The main disadvantage of pretimed systems is that they cannot adapt to the actual freeway traffic situation. To cope with this problem, traffic-responsive systems have been developed [3]. In the traffic-responsive approach, we utilize detectors and computers to determine mainline flow and ramp demand in the immediate vicinity of the ramp and to set an appropriate metering rate. There are two basic types of traffic-responsive ramp metering schemes: flow-based and gap-based. In the flow-based approach, detectors periodically provide traffic volume and occupancy data from both upstream and downstream to the ramp metering controller. Algorithms based on traffic-flow theories determine ramp metering rates based on the current flow rate and occupancy values. The gap-based approach attempts to maximize the volume entering the freeway by looking for "gaps" in the main traffic stream. Upstream detectors provide gap information [8]. Predictive schemes have been developed to predict traffic conditions in the near future and prevent conditions such as flow breakdowns.

The detectors and ramp meters of a ramp metering system can be *isolated* or *coordinated*. In an isolated system, ramp meters make decisions based only on the local traffic data [3, 16, 25]. In a system-wide control approach, ramp meters coordinate with each other to make system-wide control decisions [10, 17].

In this report we develop an Advanced Real-time Ramp Metering System (ARMS) freeway systems. ARMS consists of three integrated control schemes to handle different situations: *free flow control*, *congestion prediction*, and *congestion resolution*. In the free flow control scheme, we take the congestion risk into account in setting ramp metering rates to maintain free traffic flow and to reduce the probability of congestion. By incorporating *congestion risk* and *throughput gain* functions into the *utility function*, we can achieve the system-wide control optimum by a hill-climbing method. To handle traffic fluctuation, the congestion prediction scheme predicts imminent occurrence of congestion. When necessary, we may further reduce the ramp metering rates the free flow control algorithm calculates based on the prediction results to avoid flow breakdown due to traffic fluctuation. If congestion occurs under inevitable situations, the congestion resolution algorithm helps dissipate congestion by controlling the upstream ramps feeding the congested area.

The remainder of this report is organized as follows. In section 2 we discuss some representative ramp metering algorithms. In section 3 we present an overview of the ARMS. Section 4 describes the free flow control algorithm. Section 5 presents the prediction algorithm. Section 6 presents the congestion resolution algorithm. Section 7 concludes this report.

2 BACKGROUND

Many different ramp metering systems have been designed, tested, and operated in the past. In this section, we classify the current ramp metering schemes and briefly discuss merits and demerits of those approaches.

Figure 1 illustrates the classification of these traffic systems.

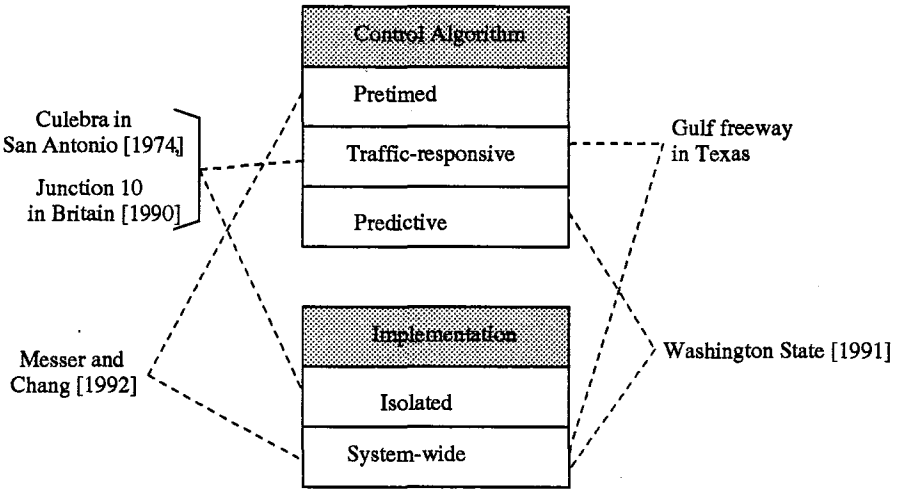


Figure 1: Classification of the surveyed traffic systems

2.1 SYSTEM-WIDE PRETIMED APPROACH

Messer and Chang [17] studied a system-wide ramp metering control strategy based on static optimization methods. They used a linear programming model, which was proposed initially by Wattleworth [30], as the basic technique to search for optimal ramp metering rates. Messer and Chang used total traffic input (to the freeway system) as the system performance index, and in their optimization algorithm they also considered traffic demands, freeway layouts, freeway capacities and other factors such as weather and human behavior in their model. Their approach can also incorporate other objectives such as balancing ramp queues, and full utilization of the storage capacity of entrance ramps.

The optimization routine executes in fixed time intervals. They assumed that traffic data in one control time slice does not change.

The main constraints of the optimization algorithm include:

1. the demand should not exceed the capacity of a section,
2. the demand for a merging ramp should not exceed the capacity of the ramp,
3. the queue length of an entrance ramp should not exceed its queue storage capacity,
4. queues on different ramps should be balanced, and
5. the ramp metering rate should be between the minimum allowable value and the ramp demand.

Any ramp metering systems must observe the first two constraints. The third and fourth constraints take the queuing effect of entrance ramps into consideration. Moreover, the queuing analysis considers possible detouring of vehicles on a ramp queue. Messer and Chang used the *Origin-Destination matrix* to estimate the corresponding coefficients. We can define the Origin-Destination matrix using the geometric structure of freeway and historical information on the drivers' routes [1, 18, 32]. Simulations on real traffic data show that Messer and Chang's algorithm has a better performance than other similar optimization algorithms [17]. By using their approach, we can optimize ramp metering rates under some conditions; the implementation is comparatively simple. Messer and Chang's approach greatly improved previous linear programming models because they solve the control variables for several consecutive time intervals together to reduce the possibility of short time fluctuation.

2.2 ISOLATED TRAFFIC-RESPONSIVE APPROACH

Many isolated traffic-responsive ramp-metering systems have been developed in the last two decades. We discuss two classical examples here: the Culebra entrance ramp-metering system in San Antonio, Texas [16] and the Junction 10 ramp-metering system on freeway M6, Birmingham, Britain [20].

The Texas Transportation Institute developed the Culebra entrance ramp metering system [16] on the southbound IH 10 freeway in San Antonio, Texas. This ramp metering system was the first system of its kind in the State of Texas. The main implementation objective was efficient and reliable remote control. The system engages ramp metering when the measured

traffic density exceeds a certain level. The signal controller selects one of four metering rates based on the mainline density levels. Since the system controls only one entrance ramp, its performance effect has been minimal. On-line test results show that traffic conditions on the freeway have not significantly changed by the ramp control scheme.

A similar system was installed on Junction 10 of M6 freeway near Birmingham, Britain, 1986 [20]. This system controlled the ramp signal by a stage selection algorithm in which the stage is a duration of a red/green signal. If the traffic demand, which is measured on-line, exceeds the capacity, or if the average speed of vehicles in a merging lane becomes lower than a pre-set threshold value, a red signal is set, provided that no overflow occurs to the ramp queue. This ramp metering system reportedly improves freeway operations. It reduced significantly the average travel time on upstream and downstream sections of the metered ramp. It also significantly increased the traffic flow at the downstream sections. Despite the improved average traffic conditions, flow breakdowns occurred continuously at the installation site.

Papageorgiou *et al.* [21] proposed a feedback-control based approach called *ALINEA*. The ALINEA approach views traffic flow as the process being controlled and the on-ramp rate is the control variable. By deriving the feedback control law, we set the ramp metering rate so traffic flow will not exceed system capacity. This method is simple and should be effective if a steady state on the free flow exists.

All systems discussed so far try to set ramp metering rates such that the traffic demand will not exceed capacity. Since the traffic condition is dependent on many factors, the relations among demand, capacity, and control decisions are not always deterministic. Therefore, some probability-based algorithms have been developed to deal with the nondeterministic correlation between system parameters. A typical example reflecting this approach is the *fuzzy-set* based control scheme proposed by Chen and May [3] for entrance ramp control at the San Francisco-Oakland Bay Bridge. Their main observation about conventional schemes is that traffic conditions are usually defined in a *crisp* (deterministic) manner. For example, we often define congestion as the occupancy being greater than a fixed threshold. Hence, in conventional approaches we control freeway traffic in an identical manner, either when the occupancy is very close to the threshold, or when it is very low. The fuzzy-logic based mechanism tried to improve the flexibility of control decisions based on different control rules, which are in the form of:

IF *<traffic condition>* **THEN**
<metering action A>

ELSE
<metering action B>

Based on these control rules, ramp metering rates are manually or automatically reduced when occupancy increases; a metering rates increase, however, when vehicles on the entrance ramps have waited for a long time, or when occupancy decreases. We can apply several control rules to a particular traffic condition, and one or more control actions associated with the best suited control rules can execute. This allows the operator to handle ambiguous situations during the recovery period of an incident. For example, as congestion dissipates gradually, incident control rules become less significant than non-incident control rules. Therefore, control decisions made by the two different control rules are weighted to generate the final control decision. Otherwise, under the conventional *crisp* logic, or with a *crisp* definition of "incidents," an abrupt transition of control may occur.

Chen and May used the FREECON2 dynamic freeway model to simulate the fuzzy controller approach. Reportedly, the controller reduces the impact of incidents more effectively, compared to the existing controllers. Its ability to reduce congestion is 40 to 100 percent better than an ideal controller. However, the fuzzy controller dissipates ramp queues approximately zero to five minutes slower than the existing control. Also, the operator needs to take over the control to achieve objectives such as maximizing overall throughput, maintaining safety near the incident, and minimizing adverse impact on local streets.

2.3 SYSTEM-WIDE TRAFFIC-RESPONSIVE APPROACH

System-wide ramp metering systems are designed to improve the drawbacks of isolated ramp-metering systems. Currently, a central computer controls several ramps, as in the Gulf Freeway Ramp Control System in Texas [8, 25]. This system integrated surveillance and control system functions in the same system. Traffic-responsive ramp metering controllers have been installed on eight entrance ramps and control decisions are made by a central computer based on demand-capacity and gap-acceptance control strategies. A closed circuit television system has been installed along the freeway for traffic surveillance. The final implementation of this ramp control system combines the demand-capacity and the gap-

acceptance approaches to provide fast response to changes in freeway operation. A linear programming model establishes the demand-capacity relationship in order to assign metering rates. Evaluation was done through a comprehensive on-line test. Results showed that total travel volume increased by more than 10% during the morning peak hours, the average speed decreased by 27%, and violation of ramp signals ranged from 1% to 13%.

2.4 TRAFFIC-RESPONSIVE WITH CONGESTION PREDICTION APPROACH

Although it has been shown that most current ramp metering systems improve freeway throughput and vehicle speed, they cannot reduce flow-breakdowns. As pointed out by Nihan and Berg [19], most responsive ramp metering systems determine ramp metering rates based on one-minute old historical data; that is, existing ramp-metering systems react to past traffic conditions and, thus, cannot effectively prevent flow-breakdowns. The Seattle control system is developed to overcome this problem [19].

The Seattle system defines a bottleneck as a congested freeway section in which the occupancy is higher than 18% and the I/O flow difference is positive. A predictive algorithm runs in each ramp controller to predict bottlenecks before they actually form. Upstream entrance ramps resolves bottlenecks cooperatively. Each control interval is one-minute long, and traffic data from the previous control interval are used to decide ramp metering rates for the current control interval. In each control interval, data from detectors on the freeway are sampled frequently to calculate flow rates and occupancy levels. If the occupancy level at the downstream station is higher than 18%, and if the I/O difference is positive, then we assign to each upstream entrance ramp affecting the section a bottleneck metering rate reduction (BMRR) value which takes into account their distance. The weighting factors that determine the BMRR values are manually adjustable. The total BMRR value is equal to the number of vehicles being stored in the congested sections. They calculate the final bottleneck metering rate for an entrance ramp by subtracting its BMRR value from its previous one minute volume, and by choosing the most restrictive result if multiple bottlenecks are detected.

The bottleneck algorithm is a reactive control algorithm because it begins to reduce the ramp meters in a target area after it detects a bottleneck. To avoid further congestion, Nihan and Berg proposed a predictive algorithm which can predict the bottleneck of a target area

a few minutes before it actually occurs. This reduces congestion by properly controlling the related ramps. A statistical pattern recognition technique identifies current traffic conditions for flow breakdown prediction. Reportedly, the downstream loop occupancy averaged over the past minute of the target freeway section, and the I/O difference of the downstream section lagged two minutes composes the best pattern [19]. Historical traffic patterns that actually led to flow breakdowns are collected and grouped in the pattern space. The patterns are then labeled as having a high probability of flow breakdown two minutes later. During normal operation, the system continuously compares current traffic patterns against stored patterns to identify possible breakdown conditions. The system activates the algorithm to predict breakdowns only in lightly congested traffic situations. In other cases it utilizes the bottleneck algorithm.

The WSDOT used the Seattle Traffic Systems Management Center (TSMC) ramp metering control computer to test the predictive algorithm on-line along with existing control strategies. They collected 15 minutes worth of volume and occupancy data for analysis at mainline stations between 236th St. SW and S. Spokane St, during 6:00 a.m. and 8:00 a.m. Several factors clouded the on-line test results of the predictive algorithm. These factors included the time frame of the on-line test, the size of the before and after data sets, and the fact that the test was on-line and was subject to uncontrollable external factors (e.g., driver behavior and equipment breakdowns). The predictive algorithm appeared to increase volume and decrease occupancy during some time periods when the freeway was lightly congested. Test results for heavily congested time periods were statistically insignificant.

In summary, none of the existing approaches can completely solve the freeway traffic control problem. The pretimed schemes have more stable control behavior; but they cannot adapt to traffic changes effectively. The traffic responsive approach can adapt to rapid traffic changes but is susceptible to unstable traffic flows due to frequent and possibly radical changes in control decisions. A good compromise between these two extremes is an incremental approach which has been successfully used in SCOOT and SCAT for surface street control. This incremental control approach forms the basic structure of the free-flow control algorithm of ARMS.

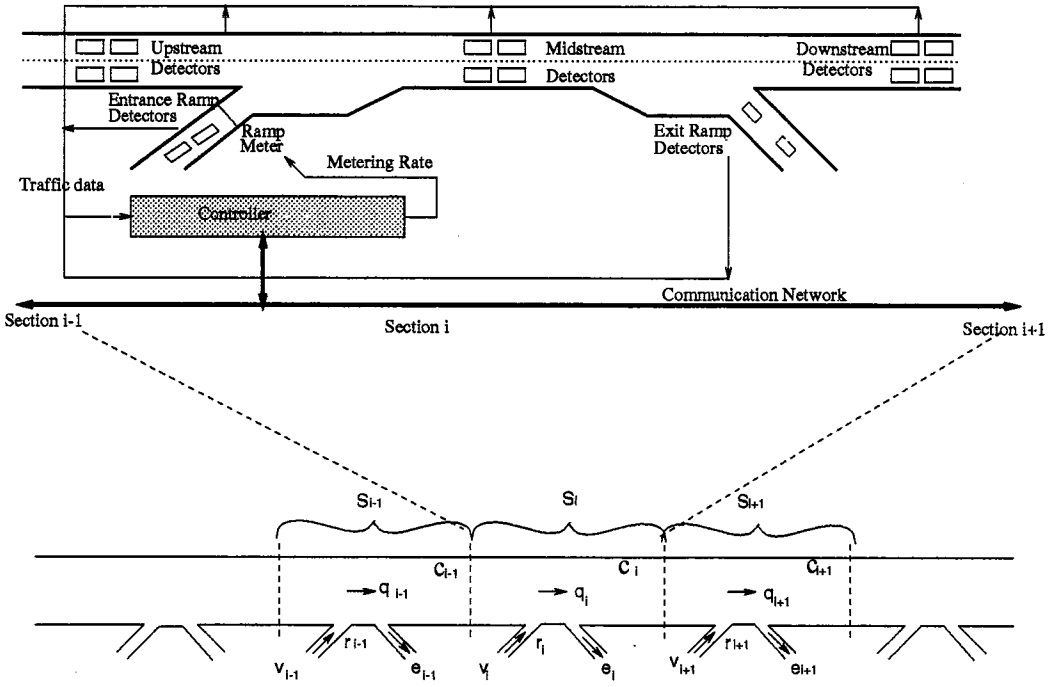
When traffic is close to freeway capacity, the chance of flow breakdown becomes significant. Therefore, we expect that in heavily loaded freeway systems, accurate prediction of congestion has a major impact on the system performance. A major difficulty in the design of congestion prediction algorithms is that different freeway stretches may have quite different characteristics before flow breakdown. Therefore, it is highly desirable for the congestion

prediction algorithm to have *self-learning* capability so accuracy of the prediction algorithm can improve over time in field operation. The self-learning capability forms the core of the congestion prediction algorithm of ARMS.

Finally, very little attention has been paid to the problem of congestion resolution in the past. Most existing approaches require the assistance of operators to resolve traffic congestions; thus, we cannot guarantee the optimality of the solution. For ARMS we design a novel approach in congestion resolution by taking into account both the freeway and surface streets. By adjusting the relative weighting factors of the freeway and surface street traffic, we can balance the congestion resolution time and its impact on the surface streets systematically.

3 THE ARMS SYSTEM MODEL

In our model, the freeway system consists of several consecutive freeway sections. Each freeway section consists of one entrance ramp and one exit ramp, and the entrance ramp is upstream to the exit ramp. We can apply the control scheme to different freeway layouts with proper modifications. Figure 2 illustrates our freeway model in which each entrance ramp has a ramp metering controller and neighboring controllers connect to each other through a communication network. We denote pair-wise inductive loop detectors located in three locations of a section as *upstream, midstream*, and *downstream* detectors [9]. The three locations connect to their local ramp metering controller to provide traffic information such as traffic flow and occupancy. Inductive loop detectors installed at the entrance and exit ramps also connect to their local controller to provide input and output traffic flow information.



n: the total number of sections in the concerned freeway area

S_i: the i^{th} freeway section

c_i: freeway capacity of S_i at the midstream detector

T: control time interval

q_i: freeway mainline flow rate at S_i

v_i: flow rate of the entrance ramp at S_i

r_i': current entrance ramp metering rate at S_i

r_i: new entrance ramp metering rate at S_i calculated by the algorithm

e_i: flow rate of the exit ramp at S_i

Q_i': current queue length of entrance ramp at S_i

Q_i: expected queue length of entrance ramp at S_i after the new metering rate is applied

Q_{max_i}: ramp queue capacity for entrance ramp at S_i

p_{ij}: percentage of the traffic entering the entrance ramp of S_i that departs at the exit ramp of S_j

c_{m_i}: merging lane capacity for entrance ramp i

Figure 2: A freeway system consisting of n consecutive sections

4 THE ARMS ALGORITHMS

ARMS, in its essence, consists of three levels of control algorithms: *free-flow control*, *congestion prediction*, and *congestion resolution*. The objective of free-flow control is to smooth peak demand so the possibility of recurrent congestion is reduced. The ARMS free-flow model treats traffic flow as a semi-static process in which traffic flow varies slowly with time. This variation is dependent mainly on the slow changes in freeway access demand, which we consider to be the major cause of recurrent freeway congestion. The *free-flow control algorithm* makes system-wide control decisions based on a free-flow model (assuming no congestion exists). We adjust ramp metering rates to adapt to prevailing traffic conditions in order to maximize freeway throughput and reduce the probability of recurrent congestion. The breakdown prediction scheme detects abnormal traffic fluctuation to avoid non-recurrent congestion. In this scheme we view the traffic flow as a dynamic process which varies rapidly with time. The flow breakdown prediction algorithm is essentially a self-learning pattern recognition technique. By proper integration of the free-flow and congestion prediction algorithms, throughput of the freeway system would improve at a reduced congestion probability. Finally, since congestion may occur, even when the entrance ramps are metered, we need an efficient congestion resolution scheme to cope with such situations. We design the congestion resolution scheme of the ARMS to balance congestion resolution time and freeway throughput.

Figure 3 shows the correlation among the three control algorithms. At the beginning of each control time interval, congestion detection and flow breakdown algorithms are executed. If the system detects congestion in the current control interval, it initiates the congestion resolution algorithm which first calculates an affected area of the congestion. The algorithm then overrules control decisions made by control algorithms of other levels. Once a freeway section is out of the affected area, normal control algorithms resume. If the algorithm predicts a possible flow breakdown may occur in a section, it reduces metering rates of the upstream entrance ramps to offset the decision made in the free-flow control algorithm. If no congestion or flow breakdown is detected in the current control interval, the controller in each ramp executes the decisions made by the free-flow control algorithm.

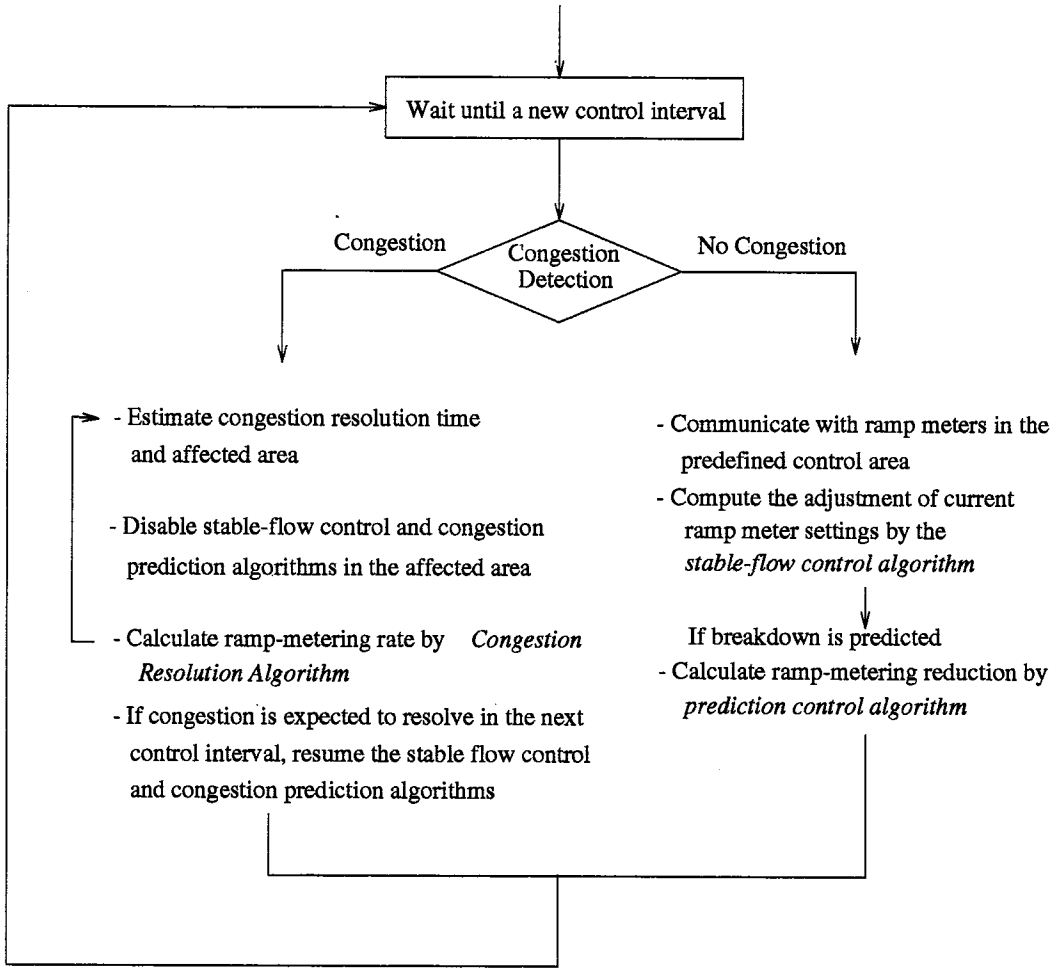


Figure 3: Integration of the three-level control algorithms

4.1 FREE-FLOW CONTROL

The objective of ramp-metering control schemes is to maximize freeway throughput subject to demand-capacity constraints. When input flow is less than freeway capacity, freeway throughput increases with the input flow. However, traffic flow may become unstable when input flow is close to the freeway capacity. Under such conditions, we consider the capacity of a freeway section as a random variable, not a constant. If the input flow increases beyond certain points, traffic flow will break down and the system throughput will drop sharply. Thus, optimizing the metering rates solely based on traffic volume and freeway capacity may lead to flow breakdown due to the unstable behavior of traffic flow. Figure 4 shows a typical relationship between freeway traffic throughput and its input flows from the entrance ramp and mainline.

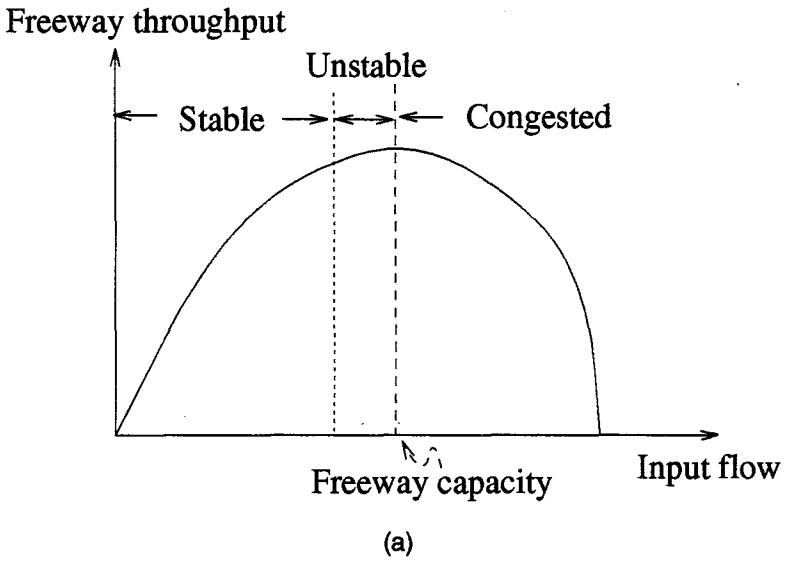


Figure 4: The relation between freeway input flow and throughput

When we set the ramp metering rate for a section, we need to consider its impact not only on the section at the current time, but also the impact on downstream sections later on. Although we can avoid congestion by setting the traffic level lower than the highway capacity by a safety margin of d (see Figure 5), the highway capacity can be permanently deprived even in normal operation.

Therefore, long-term throughput loss can be quite significant in this approach. To avoid congestion without permanent loss of capacity, we note that a high flow rate at an entrance ramp may cause congestion in a downstream section only under certain conditions. There-

fore, if the control strategy can adapt to current traffic conditions, we can serve the high traffic demand at entrance ramps without causing congestion. Otherwise, we should meter the high demand ramps at a lower rate to avoid congestion.

To maximize highway throughput without causing congestion, both the current throughput gain and the potential congestion risk are taken into consideration in our free-flow control scheme. We define a *utility function* U as the overall system performance index. We further divide U into a set of utility sub-functions $\{U_i,\ i = 1, 2, \cdots n\}$ for each section. $U = \sum_i U_i(G_i, R_i)$, $i = 1, 2, \cdots, n$, and U_i is associated with S_i to reflect its performance. U_i consists of the *throughput gain function*, G_i, and the *congestion risk function*, R_i. We discuss details on U_i, G_i and R_i later.

The main flow entering a control area, which is uncontrollable by the controllers in this area, and the input rates of entrance ramps within the control area contribute to its throughput. The congestion risk function for a section depends on the current total input flow rate to the section and the traffic contributed by upstream entrance ramps. Thus, R_i is a function of the upstream ramp metering rates: $\vec{r} = [r_1, r_2, \cdots, r_n]$.

Our control objective is to find an optimal vector $\vec{r^*}$ of ramp metering rates, where:

$$\vec{r^*} = MAX_{\vec{r}} \sum_i U_i(G_i, R_i),$$

and $\vec{r^*}$ will be searched based on an iterative gradient hill-climbing technique [12]. By using real-time traffic data, we can derive the utility gain by adjustments of ramp metering rates, from the throughput-gain functions and the congestion-risk functions. We adjust the ramp metering rate setting step-by-step in each control interval T.

Note that in a large control area, traffic conditions of two distant sections may or may not be significantly correlated. That is, on one freeway stretch it may be unnecessary to consider all the sections in a control area; however, in another stretch we may have to consider the correlation between all the sections. Therefore, we implement our control algorithm in a modular fashion so the system can adapt to different operational environments. In this scheme, the controller in a section only considers the interaction between its K upstream and K downstream adjacent sections when determining metering rates. By adjusting K, we can make the necessary tradeoff between throughput gain and congestion risk according to the local freeway conditions, e.g., geometric structure and behavior of local traffic. Intuitively, if K is small, it implies that the congestion risk is a less significant performance factor; thus,

the control scheme will tend to allow more vehicles to enter the system. Particularly, when $K = 1$ the strategy becomes an isolated control strategy and it considers only local traffic conditions when making control decisions.

We use the hill-climbing algorithm for calculating ramp metering rates in a distributed manner to reduce the computational complexity, and to improve the reliability and scalability of the algorithm. Figure 6 describes the main execution flow of the computational algorithm. First, controllers in the K upstream and downstream adjacent sections exchange traffic information. Then, each controller makes some intermediate computation on their throughput-gain and congestion-risk functions based on state information. Specifically, the intermediate results in each section are some partial derivatives of the local utility functions which are needed by the hill-climbing algorithm. The algorithm computes the final adjustment of r_i for the next control time interval based on the intermediate results received.

We discuss details on utility, throughput-gain, congestion-risk functions, and implementation issues in subsequent subsections.

4.1.1 Throughput Gain: G_i

As discussed above, the throughput gain function reflects the relationship between the freeway throughput and ramp metering rates. Traffic demand on an entrance ramp includes the arriving flow and blocked vehicles on the entrance ramp. At S_i, if the current ramp metering rate r_i already meets the traffic demand on entrance ramp i, there is no need for increasing r_i since U_i cannot increase further. However, if the traffic demand for entrance ramp i is greater than current r_i, U_i will increase with r_i. Here, we propose a throughput gain function, through which we assign a higher value to a ramp with high demand, as:

$$G_i(r_i) = \begin{cases} \frac{\frac{1+Q_i'}{T}+(v_i-r_i')}{r_i'} \cdot r_i, & \text{if } r_i < v_i \\ \frac{\frac{1+Q_i'}{T}+(v_i-r_i')}{r_i'} \cdot v_i, & \text{if } r_i \geq v_i \end{cases}, \tag{1}$$

where Q_i' and r_i' are the queue length and ramp metering rate in the last control interval, respectively, and v_i and r_i are the current arrival rate and the ramp-metering rate on the entrance ramp of S_i. We transform the ramp queue Q_i' into an equivalent arrival flow as $\frac{1+Q_i'}{T}$ in Equation 1, so that we can express different types of demand consistently, and "1" is

a compensation factor to ensure the partial derivative of G_i, with respect to, r_i is non-zero. We normalize the traffic demand $\frac{1+Q_i'}{T} + (v_i - r_i')$ by dividing it by r_i'.

4.1.2 Congestion-Risk Function: R_i

The congestion risk function R_i is a probabilistic measure which describes the likelihood of downstream congestion with respect to an entrance ramp in the near future. Here we propose a congestion risk function based on a general queuing system model in which we assume the probability of congestion increases of traffic flow. Therefore, we first define a *risk factor* with respect to S_i as $d_i(q_i) = 1/(c_i - q_i)$, where c_i is the capacity and q_i is the flow rate of S_i, and d_i satisfies the following attributes: $d_i'(q_i) > 0$, $d_i''(q_i) > 0$ and $\lim_{q_i \to c_i} d_i(q_i) = \infty$, as depicted in Figure 7b.

We can use a more sophisticated expression on d_i to represent more complicated traffic behavior. When we set r_i for S_i, we need to consider its impact on the K downstream sections. Therefore, we can express the congestion risk function for S_i as:

$$R_i = \sum_{j=i}^{i+K} [a_{ij} \cdot d_j(q_j)] = \sum_{j=i}^{i+K} (a_{ij} \cdot \frac{1}{c_j - q_j}), \tag{2}$$

where a_{ij} is the coefficient that reflects the correlation of traffic flows between S_i and S_j; we could derive it from on-line traffic data or using probabilistic estimation. We propose two different models to demonstrate construction of the a_{ij}'s. Model-I is similar to the classical origin-destination model which assumes that a steady state distribution on a source (entrance ramp) to different destinations (exit ramps) exists. Model-II uses only local information to estimate the 'exit-ratio' of each exit ramp based on mainline flow and entrance ramp flow. We explain these two models next.

In model-I, $a_{ij} = p_{ij}$ is directly effected by the percentage of the flow entering from the entrance ramp of S_i that exits at the exit ramp of S_j, as demonstrated in Figure 8. For convenience, we can express p_{ij}'s in a matrix as:

$$P = \begin{vmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ 0 & p_{21} & \cdots & p_{2J} \\ 0 & \cdots & \cdots & \cdots \\ 0 & \cdots & p_{ij} & \cdots \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & p_{nn} \end{vmatrix}$$

where n is the number of sections in the control area. The most accurate, yet most expensive, technique to derive P is via actual survey of motorists. To avoid high survey cost, other more cost-effective modeling techniques have been developed. We will compare some of the most representative ones and show an adaptive technique for estimation of p_{ij}'s.

Once we get p_{ij}'s, we can calculate the congestion risk function contributed by r_i to its K downstream sections as:

$$\begin{aligned} R_i &= \tfrac{1}{c_i - q_i} + (1 - p_{ii+1}) \cdot \tfrac{1}{c_{i+1} - q_{i+1}} \\ &\quad + \cdots + \\ &\quad (1 - \textstyle\sum_{j=i+1}^{i+K} p_{i,j}) \cdot \tfrac{1}{c_{i+K} - q_{i+K}}, \end{aligned} \tag{3}$$

$a_{ij} = 1 - \sum_{l=i+1}^{j} p_{il}$ reflects the percentage of the flow entering from S_i and exiting from S_j. We can express q_i's, r_i's and e_i's as:

$$\begin{aligned} q_i &= q_{i-1} - e_{i-1} + r_i \\ &= q_{i-2} - e_{i-2} + r_{i-1} - e_{i-1} + r_i \\ &= \cdots \\ &= q_{i-K} + \sum_{j=i-K+1}^{i} r_j - \sum_{j=i-K}^{i-1} e_j \end{aligned} \tag{4}$$

By combining Equation 3 and 4, we get R_i as a function of $\vec{r} = [r_{i-K}, \cdots, r_{i-1}, r_i]$, or

$$
\begin{aligned}
R_i \;=\; & \frac{1}{c_i-(q_{i-K}+\sum_{j=i-K+1}^{i} r_j-\sum_{j=i-K}^{i-1} e_j)} \\
& +(1-p_{ii+1})\cdot \frac{1}{c_{i+1}-(q_{i-K+1}+\sum_{j=i-K+2}^{i+1} r_j-\sum_{j=i-K+1}^{i} e_j)} \\
& +\cdots \\
& +(1-\textstyle\sum_{j=i+1}^{i+K} p_{i,j})\cdot \frac{1}{c_{i+K}-(q_i+\sum_{j=i+1}^{i+K} r_j-\sum_{j=i}^{i+K-1} e_j)}
\end{aligned}
\tag{5}
$$

In model II, we base R_i on the notion of *exit ratio* in each section, as depicted in Figure 9, in which q_i denotes the mainline flow rate and L_i denotes the exit ratio of S_i. In this approach, we define L_i as the change rate of exit ramp flow with respect to the mainline flow of S_i. We may justify the exit ratio model by simply observing that if the freeway is to maintain stable flows, the exit rate should be proportional to the mainline flow rate of the section. An example of this model is given in Figure 9b. Let $e_i = \psi(q_i)$ denote a statistical predictor that describes the relation between q_i and e_i, then we can have $L_i = \frac{d\phi(q_i)}{dq_i} \approx \frac{\Delta e_i}{\Delta q_i}$.

Based on the notion of exit ratios, we can estimate the exit ramp flow rates in terms of main flow rates. Assuming that in S_i we can estimate the main flow rate to become q_i^t after a time period of t, we can estimate the exit ramp flow rate after t as $e_i + L_i(q_i^t - q_i)$. Vehicles that enter the entrance ramp of S_i will affect a downstream section S_j only when they actually arrive at S_j. That is, if we can estimate the flow rate of S_j when the flow from ramp i travels to S_j, we can calculate the congestion risk of S_j caused by r_i. We can express equation 2 as:

$$
R_i = \sum_{j=i}^{i+K} d_j(q'_j) = \sum_{j=i}^{i+K} \frac{1}{c_j - q'_j}, \tag{6}
$$

where, q'_j is the expected future mainline flow rate in S_j, and q'_j may be different from q_i which is the current flow rate.

Now we discuss how to estimate q'_j using the ramp metering rates at K upstream sections. Consider the traffic flow from S_i to S_{i+1}. Once the flow passes S_{i+1}, q'_{i+1} may change due to the incoming mainline flow from S_i. Let us denote the estimated flow rate in S_i by q'_i; the flow rate in S_{i+1} can be estimated as:

$$q'_{i+1} = q'_i - [e_i - L_i(q'_i - q_i)] + r_{i+1},$$

where q_i, r_i, and e_i are current flow rates on the mainline, entrance ramp and exit ramp of S_i, respectively. Similarly, if considering the traffic flow from S_{i+1} to S_{i+2}, we have:

$$q'_{i+2} = q'_{i+1} - [e_{i+1} - L_{i+1}(q'_{i+1} - q_{i+1})] + r_{i+2}.$$

Now considering the traffic flow from S_i to S_j, we can estimate q'_j from q_i, r_m's and e_m's, $i \leq m \leq j$, as:

$$\begin{cases} q'_i & = & q_i \\ q'_{i+1} & = & q'_i + r_{i+1} - e_i - L_i(q'_i - q_i) \\ \cdots & & \\ q'_m & = & q'_{m-1} + r_m - e_{m-1} - L_{m-1}(q'_{m-1} - q_{m-1}), \quad i \leq j \leq i + K \\ \cdots & & \\ q'_j & = & q'_{j-1} + r_j - e_{j-1} - L_{j-1}(q'_{j-1} - q_{j-1}) \end{cases} \tag{7}$$

We get R_i based on the q'_j derived from this equation.

4.1.3 Utility Function: U_i

The utility function $U = \sum_i U_i$ is a composite performance index of throughput gain and congestion risk. Recall that in $U_i(G_i, R_i)$, we have

$$G_i = \varphi(r_i)$$

and

$$R_i = \psi(\vec{r}) = \psi(r_{i-K}, \cdots, r_{i-1}, r_i, r_{i+1}, \cdots, r_{i+K}).$$

Two basic attributes of U_i are that $\frac{\partial U_i}{\partial G_i} \geq 0$ and $\frac{\partial U_i}{\partial R_i} \leq 0$. We should define U_i so we can properly reflect the tradeoff between the throughput gain function and the congestion risk function. For example, we can define U_i as:

$$U_i(\varphi(r_i), \psi(\vec{r})) = \frac{\varphi(r_i)}{\psi(\vec{r})}, \tag{8}$$

or

$$U_i(\varphi(r_i), \psi(\vec{r})) = a_i\varphi(r_i) - b_i\psi(\vec{r}), \tag{9}$$

where a_i and b_i are coefficients, and we can express a system-wide optimum of $\vec{r} = [r_1, r_2, \cdots, r_n]$ as:

$$\vec{r^*} = argmax_{\vec{r}} \sum_i U_i(g_i(r_i), R_i(\vec{r})).$$

4.1.4 Implementation Algorithm

As mentioned earlier, ramp metering rates are adjusted periodically for every control interval T. According to the hill-climbing method, the adjustment to r_i is given as:

$$Deltar_i = \varepsilon[\frac{\partial U_i}{\partial G_i}\frac{\partial G_i}{\partial r_i} + \frac{\partial U_i}{\partial R_i}\frac{\partial R_i}{\partial r_i} + \sum_{j=1_{j\neq i}}^{m} \frac{\partial U_j}{\partial R_j}\frac{\partial R_j}{\partial r_i}]. \tag{10}$$

where ε is the adjustment parameter which we can pre-decide from sample data [2, 24]. Together the first two terms reflect the utility gain/loss in S_i by increasing r_i; the remaining terms reflect the utility loss with increased r_i for other sections. The hill-climbing method guarantees that by adjusting control variables, r_i's, the system objective function U will converge to its optimal value [2, 24, 27].

We now show the procedures to solve Equation 10. For the first term $\frac{\partial U_i}{\partial G_i}\frac{\partial G_i}{\partial r_i}$, it is straightforward to get:

$$\frac{\partial U_i}{\partial G_i} = \begin{cases} 1/R_i(\vec{r}) & \text{if Equation 8 is adopted} \\ a_i & \text{if Equation 9 is adopted} \end{cases}$$

and

$$\frac{\partial G_i}{\partial r_i} = \begin{cases} \frac{1+Q_i'}{T} + (v_i - r_i'), & \text{if } r_i < v_i \\ 0, & \text{if } r_i \geq v_i \end{cases}$$

Similarly, $\frac{\partial U_i}{\partial R_i}\frac{\partial R_i}{\partial r_i}$ can be calculated as:

$$\frac{\partial U_i}{\partial R_i} = \begin{cases} -G_i/(R_i(\vec{r}))^2 & \text{if Equation 8 is adopted} \\ -b_i & \text{if Equation 9 is adopted,} \end{cases}$$

where b_i is the adjustment coefficient used in the utility function, i.e., Equation 9.

R_j is a function of r_i's, $i \in \{j-K, j-K+1, \cdots, j+K\}$, and $\frac{\partial R_j}{\partial r_i} = 0$, if $i \notin \{j-K, j-K+1, \cdots, j+K\}$. If we use model-I for the congestion risk functions, we can derive $\frac{\partial R_j}{\partial r_i}$ from Equation 3 and 4 as:

$$\frac{\partial R_j}{\partial r_i} = \sum_{l=j}^{j+K} \frac{-1}{(c_l - q_l'')^2} \cdot [1 - \sum_{m=i+1}^{l-1} p_{im}]^2 \tag{11}$$

.

When we use model-II for the risk functions, we can express $\partial R_j/\partial r_i$ as:

$$\frac{\partial R_j}{\partial r_i} = \sum_{k=i}^{min(j+K, i+K)} \frac{\partial d_k}{\partial r_i}, \tag{12}$$

and we can compute $\partial d_k/\partial r_i$ by Equation 7 as follows:

$$\begin{cases} \frac{\partial q'_{k-K+1}}{\partial r_i} = 0 \\ \cdots \\ \frac{\partial q'_{i-1}}{\partial r_i} = 0 \\ \frac{\partial q'_i}{\partial r_i} = 1 \\ \cdots \\ \frac{\partial q'_{k-1}}{\partial r_i} = (1 - L_{k-2})\frac{\partial q'_{k-2}}{\partial r_i} \\ \frac{\partial q'_k}{\partial r_i} = (1 - L_{k-1})\frac{\partial q'_{k-1}}{\partial r_i} \end{cases}, \text{ where } 0 \leq k - i \leq K, \quad (13)$$

By combining the equations in 13, we get:

$$\frac{\partial d_k}{\partial r_i} = \frac{-1}{(c_k - q_k)^2}\frac{\partial q'_i}{\partial r_k} = \frac{-1}{(c_k - q_k)^2}\Pi_{m=k+1}^{k-1}(1 - L_m),$$

where $0 \leq k - i \leq K$, and thus:

$$\frac{\partial R_j}{\partial r_i} = \sum_{k=i}^{min(j+K,i+K)}[\frac{-1}{(c_k - q_k)^2}\Pi_{m=k+1}^{k-1}(1 - L_m)], \quad (14)$$

where $i - K \leq j \leq i + K$.

At the end of a control interval, each section performs a sequence of operations to set the ramp metering rate for the next control time interval. We summarize the control algorithm at S_i as follows:

(1) Collect mainline flow rate q_i, entrance ramp arrival rate v_i, and exit-ramp flow rate e_i;

(2) Calculate the correlation probability of ramps (O-D trip table or exit ratios);

(3) Send q_i, v_i, e_i, and p_{ij} (or L_i) to K downstream adjacent sections;

(4) Receive q_j, v_j, e_j, and p_{ji} (or L_j) from K upstream adjacent sections;

(5) Calculate $\partial G_i/\partial r_i$ and $\partial R_i/\partial r_j$, $j \in \{i-K, i-K+1, \cdots, i+K\}$;

(6) Send $\partial R_i/\partial r_j$ to S_j, $j \in \{i-K, i-K+1, \cdots, i+K\}$, and receive $\partial R_j/\partial r_i$ from S_j, $j \in \{i-K, i-K+1, \cdots, i+K\}$;

(7) Calculate Δr_i (using Equation 10);

(8) Set the metering rate in S_i as $r_i = r_i + \Delta r_i$;

(9) Go to step 1.

In Steps 5 and 6 the algorithm calculates and exchanges the partial derivatives. In step 7 it uses the derivatives in the calculation of Δr_i's. Note that the algorithm distributes the calculation of partial derivatives among all the controllers.

We now use an example to show how the free flow control algorithm works. Consider a three-lane system consisting of four sections with capacity of 5400 veh/hr each (1800 veh/hr per lane). Figure 10 depicts a snapshot of the traffic system in last control interval T_k, and the ramp metering rates during T_k. Figure 11 shows the measured flow at the end of T_k. Flow rates given in the boxes are the combined demands from the on-ramps and the mainline after T_k. If we do not meter the entrance ramps, then congestion may occur at S_4 because the main flow rate (5460 cars/hr) is higher than the capacity (5400 cars/hr).

Now, we demonstrate how to compute the ramp metering rates based on the utility function given in Equation 9 and model-II. The control parameters used in this example are given as follows:

- The control time interval $T = 3$ minutes;
- Exit-ratio $L_i = 0.25$, $i = 1, 2, 3, 4$;
- The adjustment step $\varepsilon = 1$;
- The utility function coefficient $a_i = 1.0$ (we discuss b_i later).

For simplicity, assume that current queue length on each ramp is 0. Calculate the ramp metering rates for the next control time interval step-by-step as follows, omitting the procedures for information exchange:

1. $\frac{\partial U_i}{\partial G_i} = 1$, $i \in \{1,2,3,4\}$;

2. $\frac{\partial G_1}{\partial r_1} = [(1+0)/(3/60) + (1500-1000)]/1000 = 0.52$,
   $\frac{\partial G_2}{\partial r_2} = [(1+0)/(3/60) + (1000-1000)]/1000 = 0.02$, $\frac{\partial G_3}{\partial r_3} = [(1+0)/(3/60) + (2000-2000)]/2000 = 0.01$,
   $\frac{\partial G_4}{\partial r_4} = [(1+0)/(3/60) + (1500-1000)]/1000 = 0.52$;

3. $\frac{\partial G_i}{\partial r_i} = b_i$, $i \in \{1,2,3,4\}$;

4. $\frac{\partial R_1}{\partial r_1} = 0.07045 \times 10^{-4}$, $\frac{\partial R_1}{\partial r_2} = 0.0871 \times 10^{-4}$, $\frac{\partial R_1}{\partial r_3} = 0.1094 \times 10^{-4}$, $\frac{\partial R_1}{\partial r_4} = 0.0625 \times 10^{-4}$,
   $\frac{\partial R_2}{\partial r_1} = 0.0654 \times 10^{-4}$, $\frac{\partial R_2}{\partial r_2} = 0.0871 \times 10^{-4}$, $\frac{\partial R_2}{\partial r_3} = 0.1094 \times 10^{-4}$, $\frac{\partial R_2}{\partial r_4} = 0.0625 \times 10^{-4}$,
   $\frac{\partial R_3}{\partial r_1} = 0.0615 \times 10^{-4}$, $\frac{\partial R_3}{\partial r_2} = 0.0820 \times 10^{-4}$, $\frac{\partial R_3}{\partial r_3} = 0.1094 \times 10^{-4}$, $\frac{\partial R_3}{\partial r_4} = 0.0625 \times 10^{-4}$,
   $\frac{\partial R_4}{\partial r_1} = 0.0264 \times 10^{-4}$, $\frac{\partial R_4}{\partial r_2} = 0.0352 \times 10^{-4}$, $\frac{\partial R_4}{\partial r_3} = 0.0469 \times 10^{-4}$, $\frac{\partial R_4}{\partial r_4} = 0.0625 \times 10^{-4}$;

5. $\Delta r_1 = \varepsilon[\frac{\partial U_1}{\partial G_1}\frac{\partial G_1}{\partial r_1} + \frac{\partial U_1}{\partial R_1}\frac{\partial R_1}{\partial r_1} + \frac{\partial U_2}{\partial R_2}\frac{\partial R_2}{\partial r_1} + \frac{\partial U_3}{\partial R_3}\frac{\partial R_3}{\partial r_1} + \frac{\partial U_4}{\partial R_4}\frac{\partial R_4}{\partial r_1}]$,
   $= 1000 \times (0.52 - b_1 \times 0.2237 \times 10^{-4})$,
   $\Delta r_2 = \varepsilon[\frac{\partial U_2}{\partial G_2}\frac{\partial G_2}{\partial r_2} + \frac{\partial U_1}{\partial R_1}\frac{\partial R_1}{\partial r_2} + \frac{\partial U_2}{\partial R_2}\frac{\partial R_2}{\partial r_2} + \frac{\partial U_3}{\partial R_3}\frac{\partial R_3}{\partial r_2} + \frac{\partial U_4}{\partial R_4}\frac{\partial R_4}{\partial r_2}]$,
   $= 1000 \times (0.02 - b_2 \times 10^{-4} \times 0.2915)$,
   $\Delta r_3 = \varepsilon[\frac{\partial U_3}{\partial G_3}\frac{\partial G_3}{\partial r_3} + \frac{\partial U_1}{\partial R_1}\frac{\partial R_1}{\partial r_3} + \frac{\partial U_2}{\partial R_2}\frac{\partial R_2}{\partial r_3} + \frac{\partial U_3}{\partial R_3}\frac{\partial R_3}{\partial r_3} + \frac{\partial U_4}{\partial R_4}\frac{\partial R_4}{\partial r_3}]$,
   $= 1000 \times (0.01 - b_3 \times 10^{-4} \times 0.375)$,
   $\Delta r_4 = \varepsilon[\frac{\partial U_4}{\partial G_4}\frac{\partial G_4}{\partial r_4} + \frac{\partial U_1}{\partial R_1}\frac{\partial R_1}{\partial r_4} + \frac{\partial U_2}{\partial R_2}\frac{\partial R_2}{\partial r_4} + \frac{\partial U_3}{\partial R_3}\frac{\partial R_3}{\partial r_4} + \frac{\partial U_4}{\partial R_4}\frac{\partial R_4}{\partial r_4}]$,
   $= 1000 \times (0.52 - b_4 \times 10^{-4} \times 0.25)$.

Note that the value of b_i reflects the relative significance of congestion risk with respect to throughput gain. Therefore, selection of b_i value should reflect conditions that may affect the freeway capacity. For example, in a lightly loaded condition, we can use a small b_i value since the chance of flow breakdown may be relatively smaller than a heavily loaded condition. Moreover, congestion is more likely to occur in bad weather; thus, we should use a larger b_i value. Therefore, we need to adjust b_i based on historical data, just as in congestion prediction. We may also adjust the b_i value adaptively in connection with our congestion prediction algorithm. That is, if we positively predict congestion by several sections, then we should increase b_i to reflect the increasing congestion risk. Otherwise, we can decrease b_i if we detect congestion rarely. As an illustration, Figures 12, 13 and 14 use $b_i = 10 \times 10^3$, $b_i = 8 \times 10^3$ and $b_i = 6 \times 10^3$, $i \in \{1,2,3,4\}$, respectively, to compute ramp metering rates and the resulting freeway flows in the next control time interval.

4.1.5 Origin-Destination Trip Table

The effectiveness of the control method depends on accuracy of O-D probabilities. Here we briefly review some existing techniques to estimate the O-D trip tables [18, 32] and then present our method. Let O_i denote the total number of trips entering the entrance ramp of S_i, and D_j denote the total number of trips departing the exit ramp of S_j within a time period. The portion of trips from S_i to S_j, denoted as T_{ij}, satisfies the relation $T_{ij} \propto O_i \cdot D_j$. Based on this relation, the most popular O-D model is the *proportional distribution method* in which we can distribute the flow rates of exit ramps among origins in proportion to their entrance flows. That is, the basic assumption of this model is that the number of vehicles entering from entrance ramps to an exit ramp is proportional to traffic volume at these entrance points [32]. This method is very simple since we can estimate the O-D tables by observing traffic flows at entrance and exit ramps.

The proportional distribution model, considers neither distance nor travel time. In fact, it has been observed that vehicles with a very short or very long travel distance have a very small probability to exit the freeway, no matter how large the entrance flow is. To remedy this deficiency, Nihan [18] proposed the Gamma distribution to model the travel distances of freeway motorists. Geometric characteristics such as shopping malls or sports stadiums may affect the accuracy of this method.

Willis and May [32] developed another O-D model in which they determined the likelihood of a driver departing from a freeway by the time difference between traveling along freeway or the frontage road to reflect physical conditions of the frontage road and the freeway system. It expresses T_{ij} as:

$$T_{ij} = \frac{D_j G_{ij}}{\sum_{m=1}^{z} D_m G_{im}} O_i,$$

where z is the total number of downstream entrance ramps of entrance ramp i, and G_{ij} is the relative attraction factor of exit ramp j with respect to entrance ramp i. This method defines G_{ij} as the product of two behavioral factors E_{ij} and L_{ij}, where E_{ij} is a relative travel time factor denoting the reduced travel time by driving along the freeway rather than along the best alternative surface street; and L_{ij} denotes the intervening opportunity and separating factors that measure the tendency of travelers to select an exit. In practice, we

can determine E_{ij} by the diversion curves published by Bureau of Public Road (BPR), and L_{ij} is calculated as:

$$L_{ij} = \prod_{k=0}^{j-1} P_k \text{ and } P_K = 1 - \frac{E_{ik} D_k}{\sum_{m=1}^{z} E_{im} D_m}$$

where k denotes an exit ramp between the entrance ramp i and exit ramp j, and D_k is the flow rate of the exit ramp k.

Cremer and Keller [32] proposed an algorithm which can dynamically produce the transition probabilities based on measured deviations of traffic volume on both entrance ramps and exit ramps. The basic relation used in their method is

$$D^T(k) = O^T(k)B(k),$$

and

$$\Delta D^T(k) = \Delta O^T(k)B(k),$$

where $D^T(k)$ is the row vector of observed exit ramp flow rates, $O^T(k)$ is the row vector of observed entrance ramp flow rates, and $B(k)$ is the matrix of the unknown transition probabilities during time interval k. They considered two cases in their work: the transition probabilities are either very stable over time or not stable. For the first case, the deviations of traffic volumes in different time periods are collected, taking into account the travel time between the entrance and exit ramps. Then, by taking n sets of deviations of traffic volumes, where n is the number of entrance ramps, one can calculate the new transition probabilities by solving n linear equations. In the second case they employ a recursive estimation algorithm to calculate the unknown transition probabilities. At each time interval, measured deviations of exit ramp volumes are compared against estimated deviations based on the measured deviations of entrance ramps and the transition probabilities of the last time interval. Then they use the differences between these two deviations to adjust the transition probabilities iteratively until the probabilities they obtain become stable. They showed that this method can track time varying transition probabilities, provided all vehicles have an identical speed.

Here, we propose a dynamic estimation technique, which can reflect the effect of travel time through normalization of traffic data for dynamic estimation of transition probabilities. First, we normalize the input volume O_i and output volume D_j by their total number and let $P_{O_i} = O_i / \sum_j O_j$ and $P_{D_j} = D_j / \sum_i D_i$. By doing so, we can treat P_{O_i} and P_{D_j} as probabilistic distributions of traffic volume at entrance and exit ramps, respectively. The

transition probability matrix is, thus, a mapping between the input to the output probabilistic distributions. That is:

$$\mathbf{P}_D = \mathbf{P}_O \mathbf{P},$$

where, $\mathbf{P}_D = \{P_{D_1}, P_{D_2}, \cdots\}$, $\mathbf{P}_O = \{P_{O_1}, P_{O_2}, \cdots\}$, and $\mathbf{P} = [p_{ij}]$. After taking the gradient of $\mathbf{P}$, we have:

$$p_{ij} = \overline{p}_{ij} + y P_{O_i}(P_{D_j} - \overline{P}_{D_j}),$$

where $\overline{p}_{ij}$ is the transition probability in last iteration, P_{O_i} and P_{D_j} are the current entrance and exit ramp volumes, respectively, and y is a positive coefficient. We calculate $\overline{P}_{D_j}$ from the transition probability of the last iteration and current entrance ramp volumes as $\overline{\mathbf{P}}_D = \mathbf{P}_O \overline{\mathbf{P}}$, where $\overline{\mathbf{P}}$ is the matrix of transition probabilities in last iteration. We repetitively use the above formula to evaluate p_{ij}'s by plugging the "current" transition probabilities derived in the last iteration into the "old" transition probabilities. The process stops when the difference between the calculated probabilities in two consecutive iterations is less than predetermined error tolerance values.

We compare the acurracy of our adaptive method for O-D probability calculation with that of proportional distribution method under a simulated freeway environment.

In our simulation, vehicle arrivals at each entrance ramp follow a Poisson distribution whose the average arrival rate is based on field survey data [1] for 3 hours of morning peak traffic. In Table 1 we show inter-arrival rates of entrance ramps where these rates can increase or decrease randomly with time.

The speed of a vehicle follows the normal distribution. We decide a vehicle's exit ramp randomly using a given O-D probability table which may randomly change with time. The O-D probability table for the survey traffic data is given in Table 2. In each time window, an O-D probability calculation algorithm updates its own synthetic O-D probability table by using real-time volume data from entrance and exit ramps. Tables 3 and 4 show two examples of generated volumes at entrance and exit ramps.

Simulation Results

Our experiments consisted of two parts based on how to randomly change a given O-D probability table. We first changed the distribution probabilities of only one origin to

exclude the effect of other origins, then we changed distribution probabilities of several origins simultaneously.

The simulation tested each run for 200 minutes, and the length of each time window was 10 minutes. For each window, we estimated O-D probabilities by our algorithm from real-time volume data. For comparison, synthesized O-D probabilities calculated by the proportional distribution method (implemented in SYNODM) are also presented. Overall, simulation results, which we plot in Figures 15 to 24, show that our method adapts to new traffic conditions faster than the proportional distribution method.

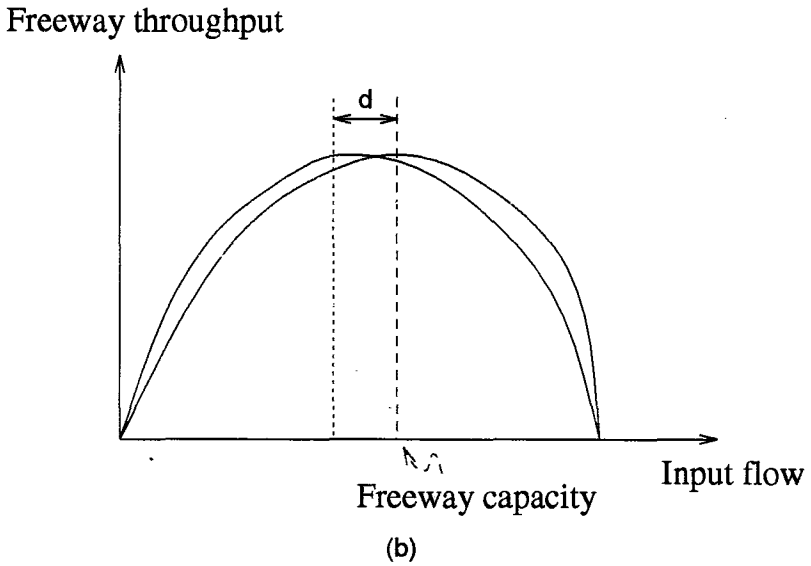


Figure 5: The relation between freeway input flow and throughput

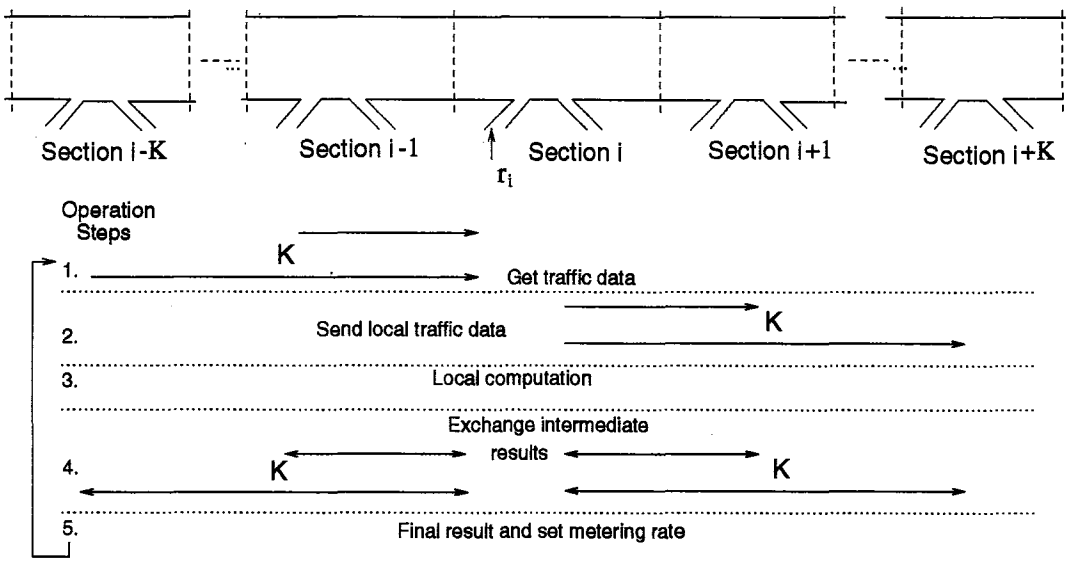


Figure 6: The outline of the free-flow control algorithm

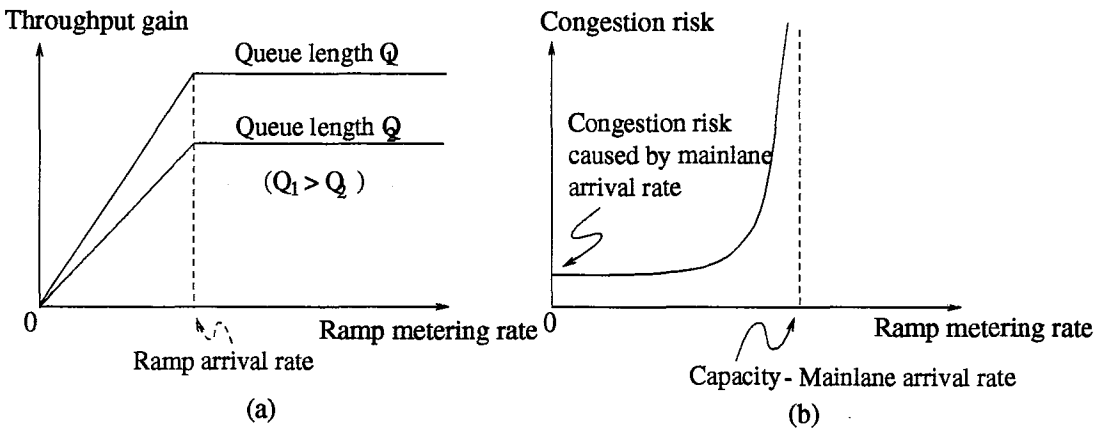


Figure 7: Characteristics of (a) throughput gain function, and (b) risk function

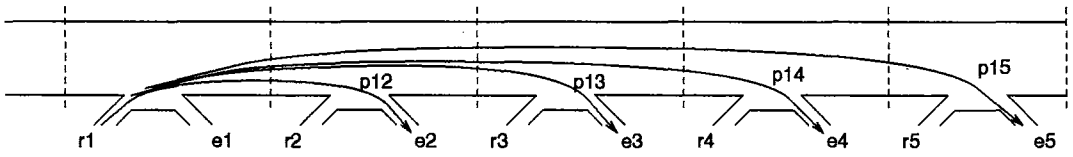


Figure 8: The O-D distribution between an entrance ramp and its downstream exit ramps

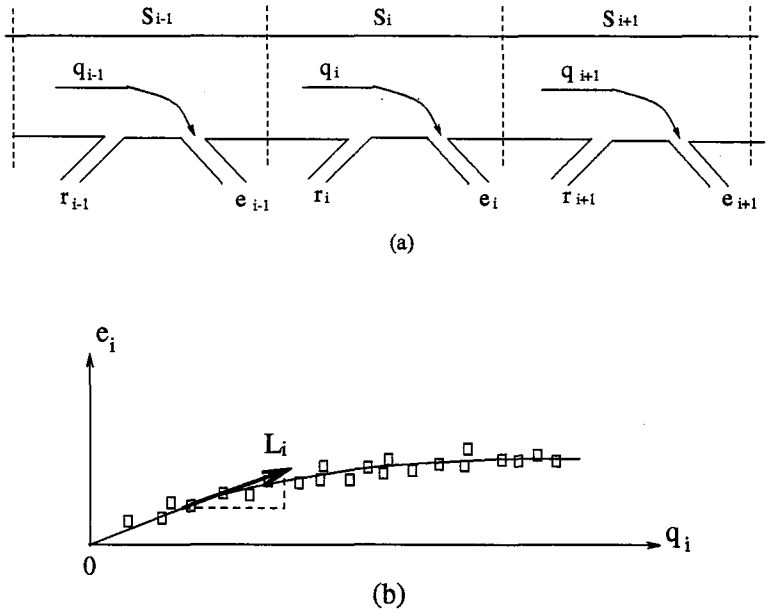


Figure 9: (a) The exit-ratio model for a freeway system, and (b) relationship between exit ramp and mainline flow rates

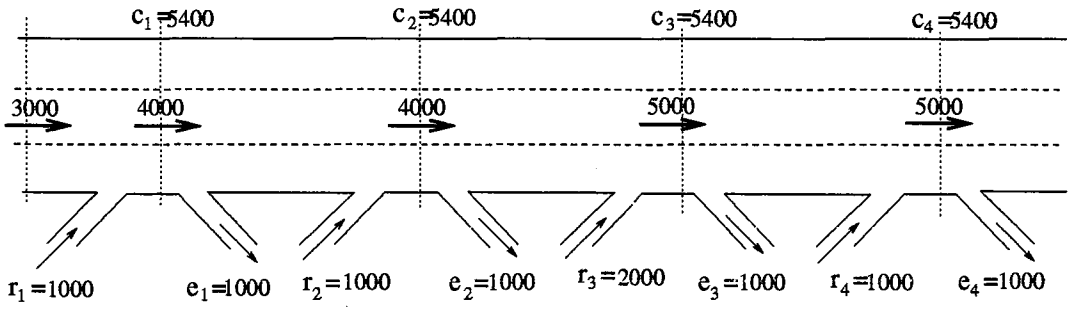


Figure 10: The ramp metering rates in control time interval T_K

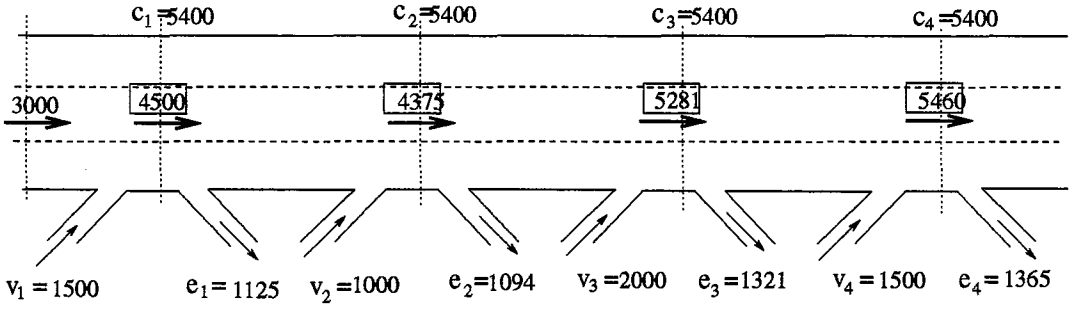


Figure 11: Current traffic demand in T_K

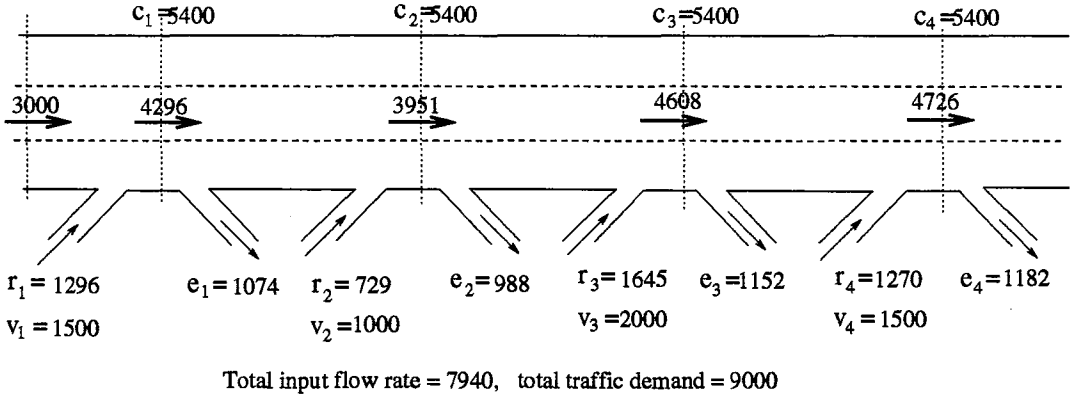
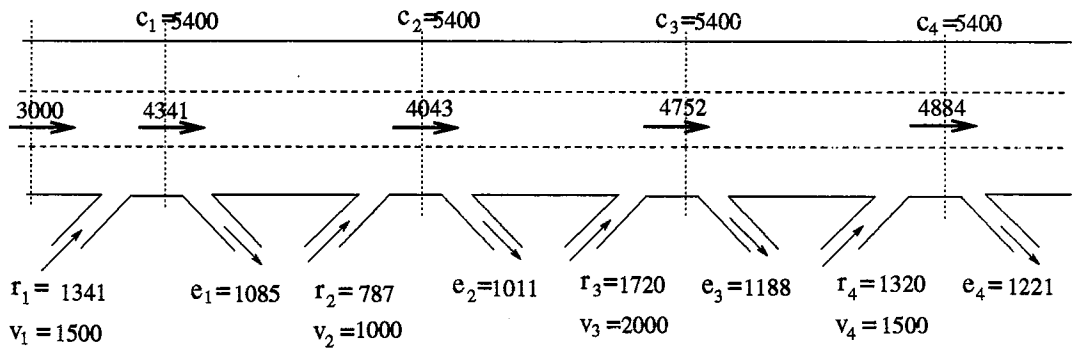
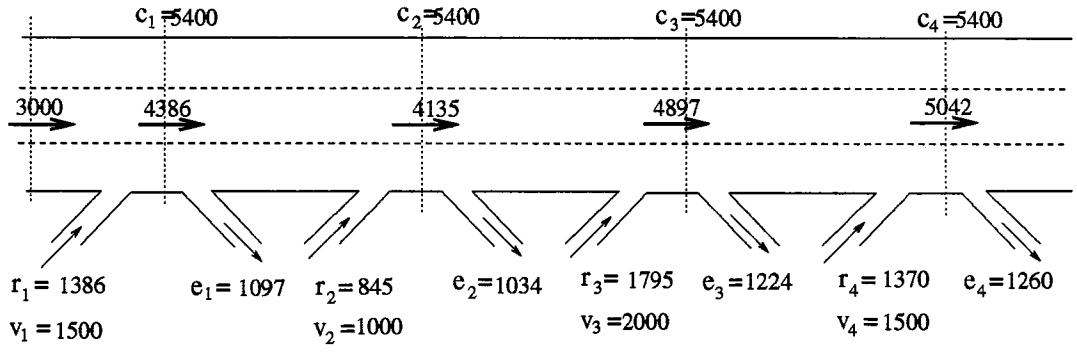


Figure 12: The ramp metering when $b_i = 10 \times 10^3$



total input flow rates = 8168, total traffic demand= 9000

Figure 13: The ramp metering when $b_i = 8 \times 10^3$



total input flow rates = 8395, total traffic demand = 9000

Figure 14: The ramp metering when $b_i = 6 \times 10^3$

Table 1: Average arrival rate and average inter-arrival time at entrance ramps

Entrance ramp	*Average arrival rate*	*Average inter-arrival time*
1	4730	2.283
2	894	12.081
3	905	11.934
4	396	27.273
5	475	22.737
6	1501	7.195
7	958	11.273
8	533	20.263
9	1157	9.334
10	568	19.014

Table 2: An O-D probability table for distributing vehicles

	1	2	3	4	5	6	7	8	9	10
1	0.0170	0.0780	0.0190	0.1450	0.0410	0.1190	0.0660	0.0460	0.0220	0.4480
2		0.0790	0.0190	0.1480	0.0410	0.1210	0.0670	0.0460	0.0220	0.4560
3			0.0210	0.1600	0.0450	0.1310	0.0730	0.0510	0.0240	0.4940
4				0.1650	0.0460	0.1340	0.0730	0.0510	0.0250	0.5060
5					0.0550	0.1600	0.0890	0.0610	0.0300	0.6050
6						0.1690	0.0940	0.0650	0.0310	0.6400
7							0.1140	0.0780	0.0380	0.7700
8								0.0880	0.0430	0.8960
9									0.0470	0.9530
10										1.0000

Table 3: The traffic volume generated at an entrance ramp for each time frame

Window	*Ent. 1*	*Ent. 2*	*Ent 3.*	*Ent 4.*	*Ent 5.*	*Ent 6.*	*Ent 7.*	*Ent 8.*	*Ent 9.*	*Ent 10.*
1	448	80	84	33	50	142	92	38	112	58
2	443	76	96	44	31	143	97	57	109	50
3	431	86	78	39	38	157	88	54	109	62
4	420	88	97	18	35	151	88	46	110	38
5	464	91	90	40	50	127	111	42	112	55
6	402	84	78	42	43	145	88	57	103	53
7	450	84	80	43	48	140	97	52	105	46
8	451	88	80	26	43	146	90	52	115	51
9	415	96	84	38	36	137	91	51	105	50
10	477	87	95	32	45	141	66	58	125	51

Table 4: A volume for exit ramp for each time frame

Window	*Ext. 1*	*Ext. 2*	*Ext. 3*	*Ext. 4*	*Ext. 5*	*Ext. 6*	*Ext. 7*	*Ext. 8*	*Ext. 9*	*Ext. 10*
1	8	41	5	69	28	83	67	30	22	468
2	10	35	14	112	19	114	70	46	28	656
3	4	25	18	94	16	118	75	43	29	715
4	10	37	15	90	24	112	74	61	40	676
5	7	51	11	94	45	110	72	42	22	668
6	5	29	12	99	31	98	64	45	24	672
7	7	44	8	100	25	112	66	57	19	717
8	10	50	13	101	29	115	78	48	27	693
9	7	43	12	91	33	115	72	48	35	661
10	1	35	8	90	28	109	59	51	30	714

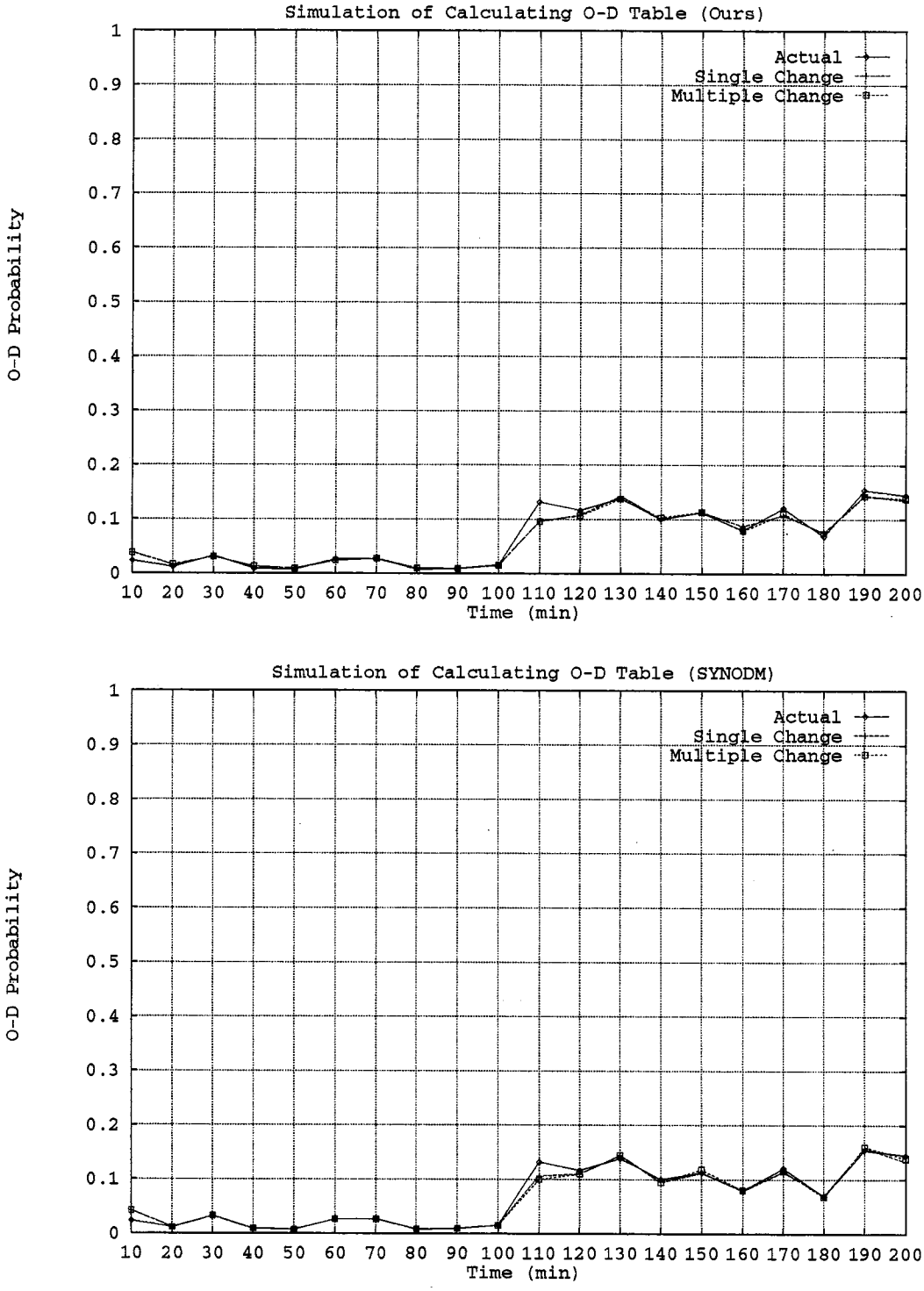


Figure 15: Synthesized O-D probabilities of O1-D1

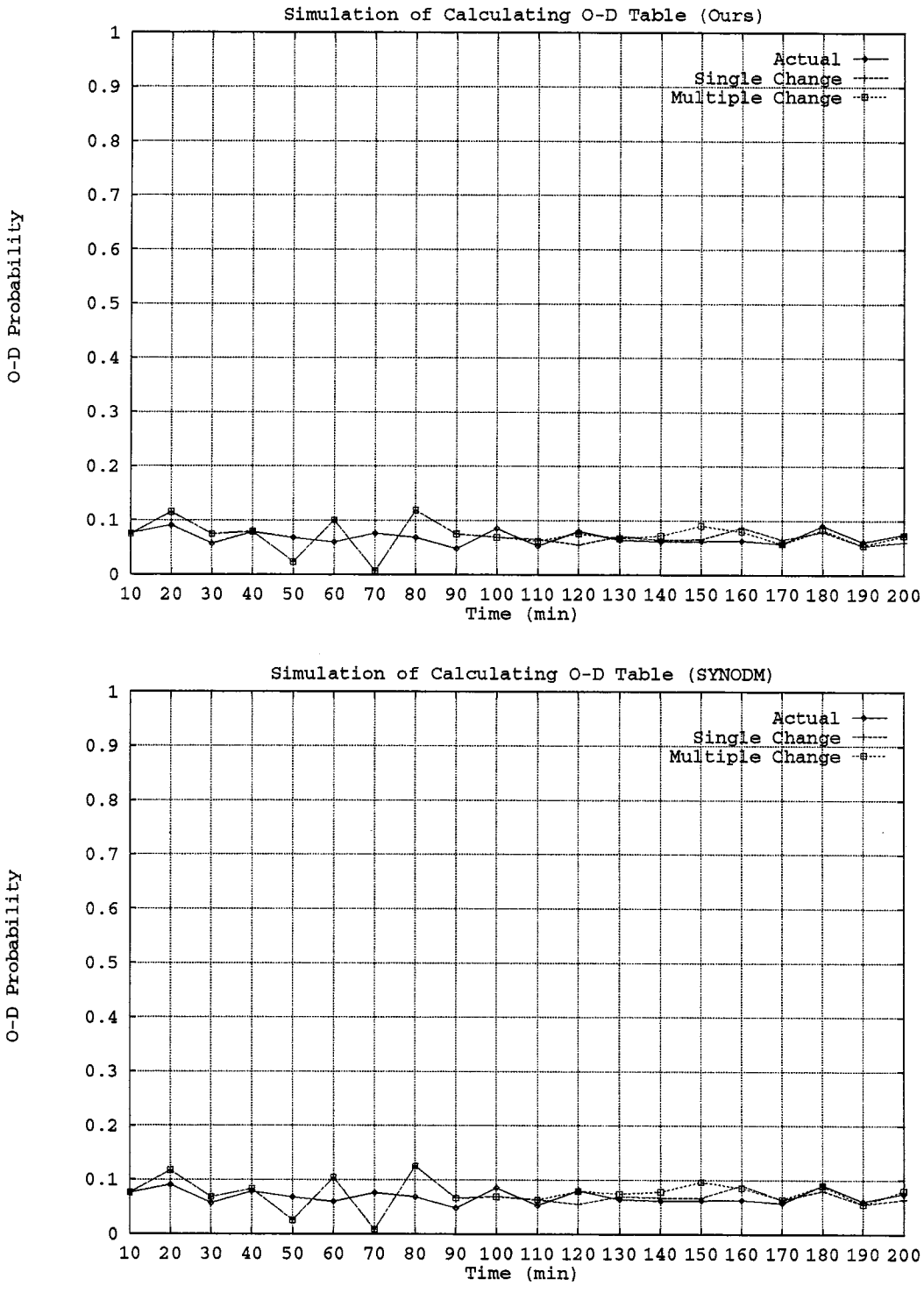


Figure 16: Synthesized O-D probabilities of O1-D2

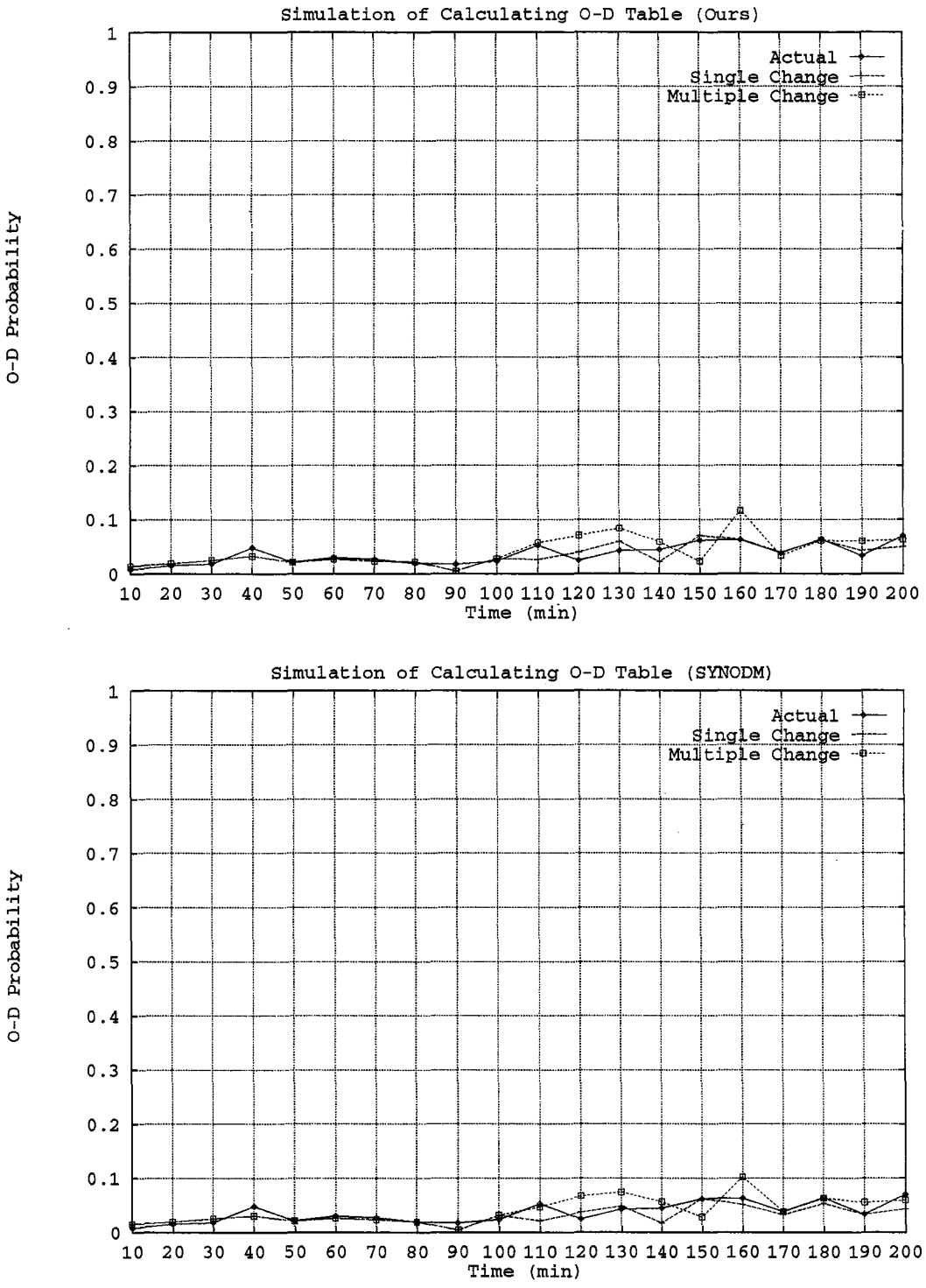


Figure 17: Synthesized O-D probabilities of O1-D3

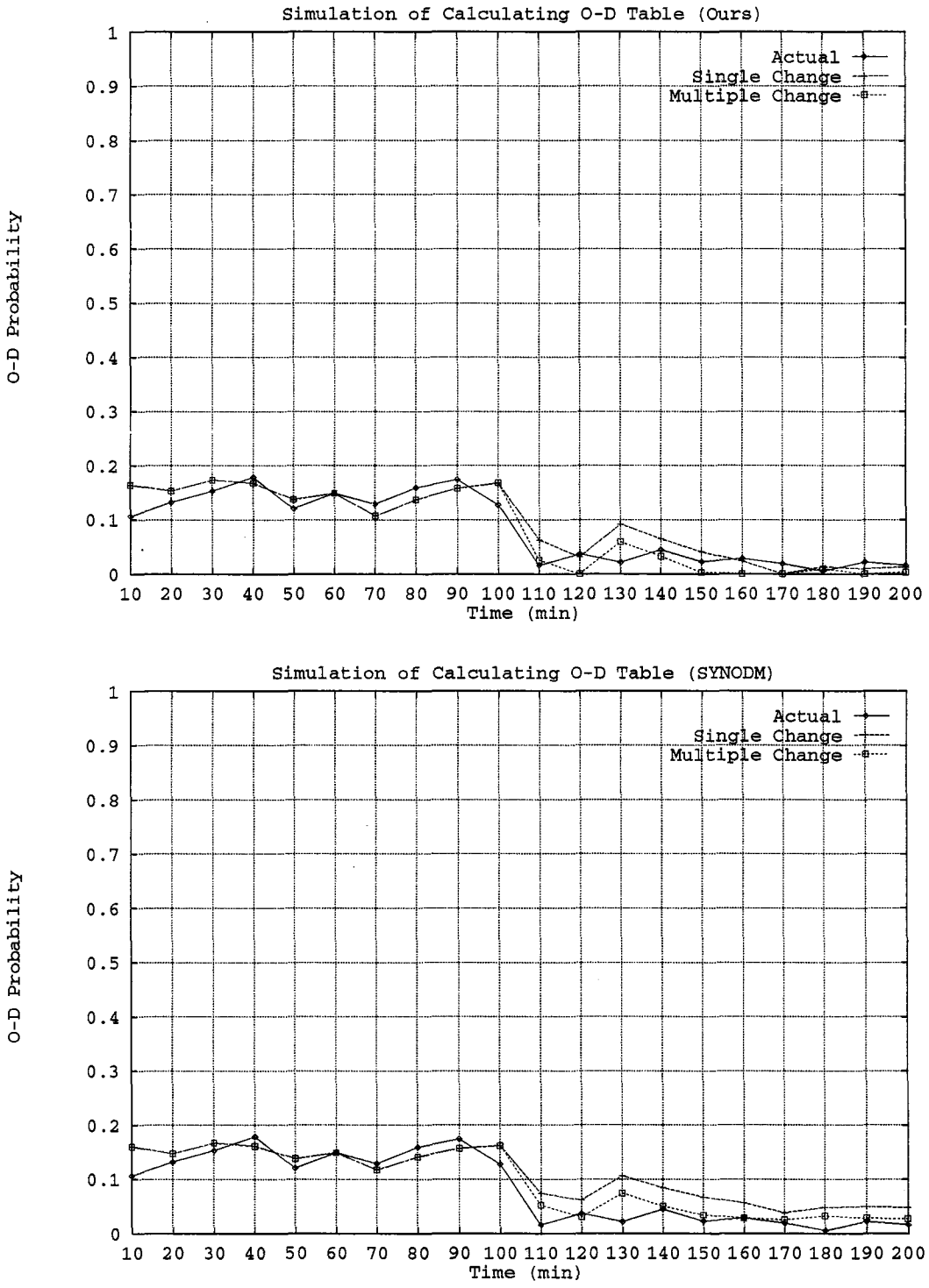


Figure 18: Synthesized O-D probabilities of O1-D4

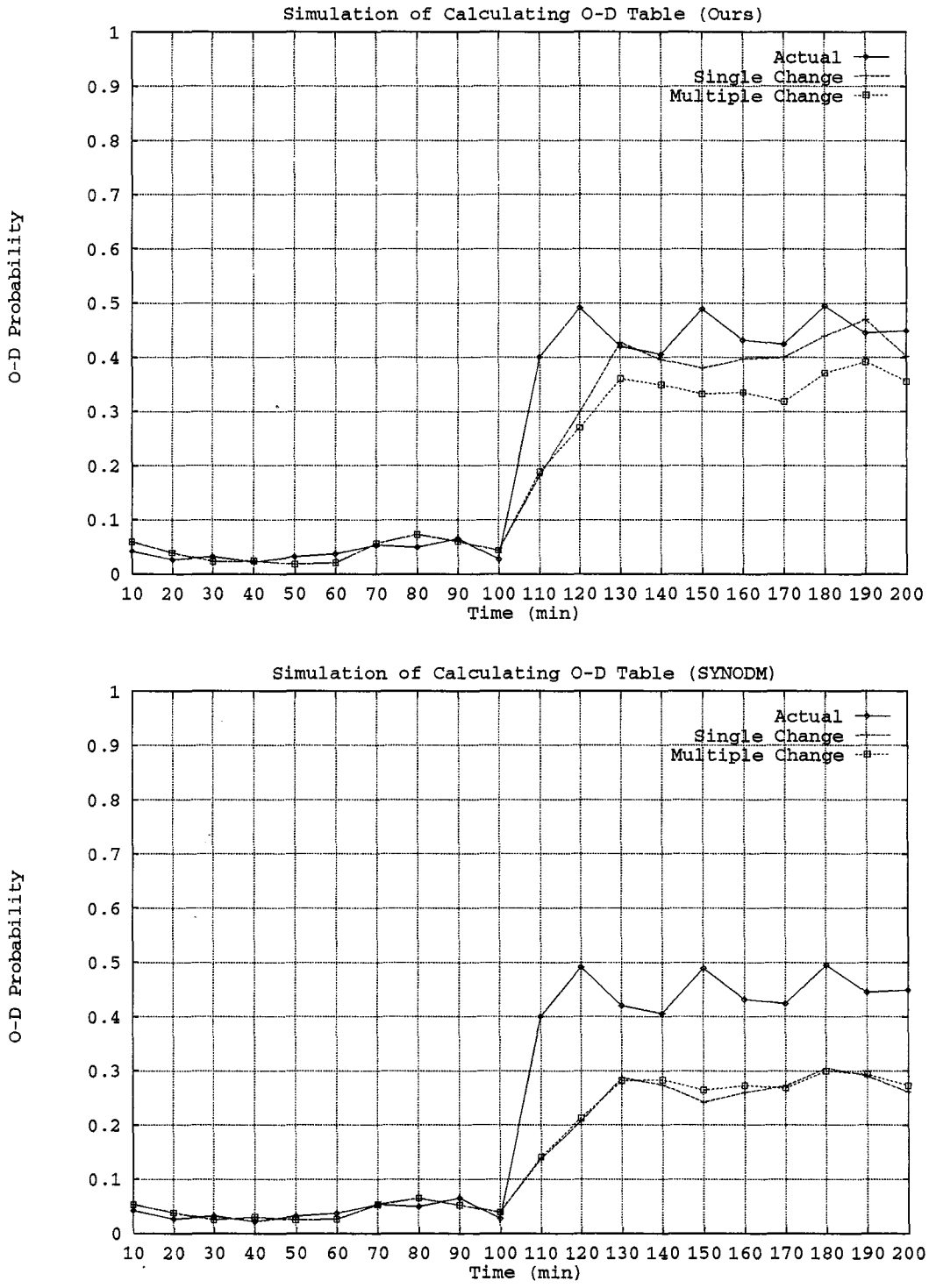


Figure 19: Synthesized O-D probabilities of O1-D5

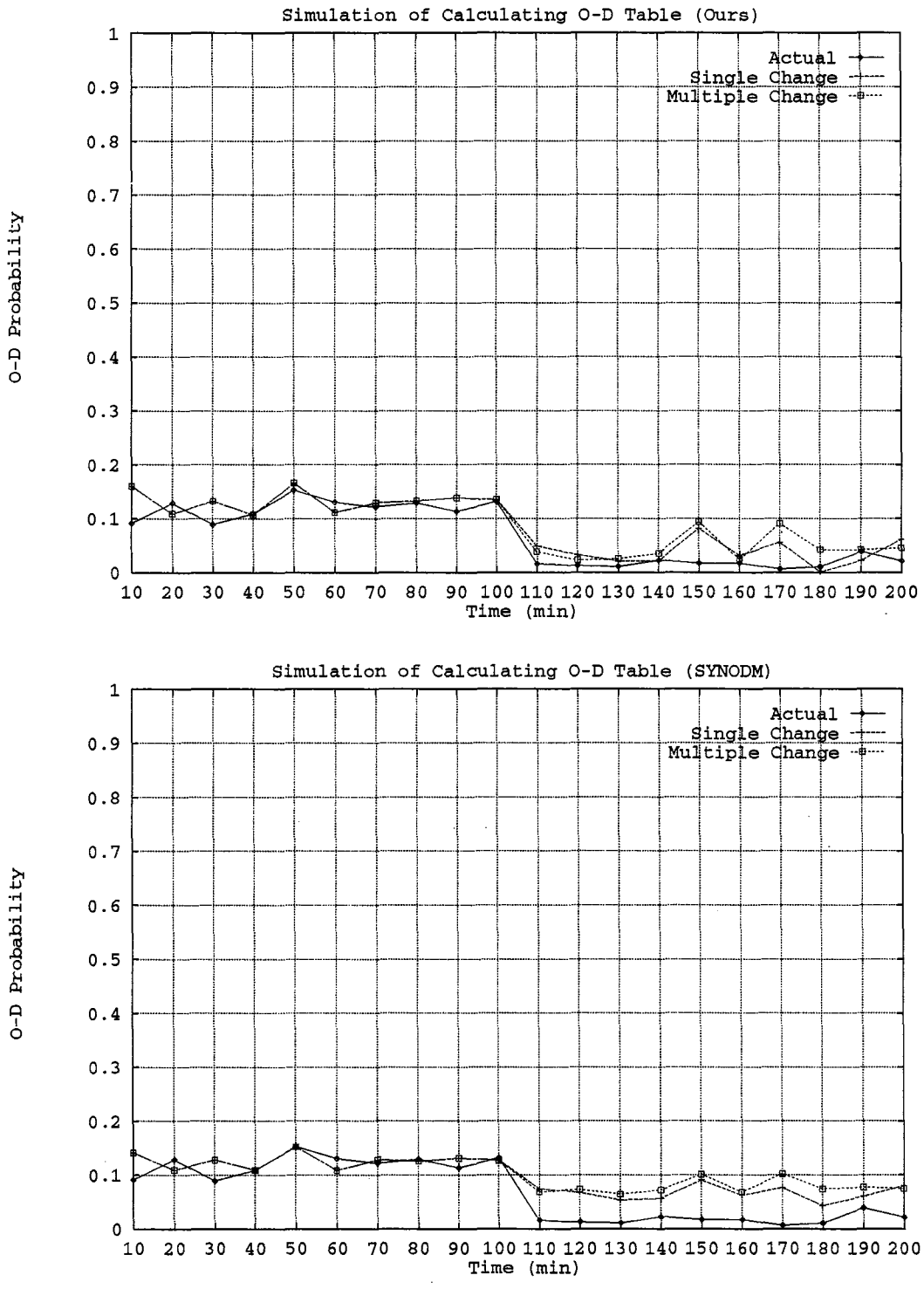


Figure 20: Synthesized O-D probabilities of O1-D6

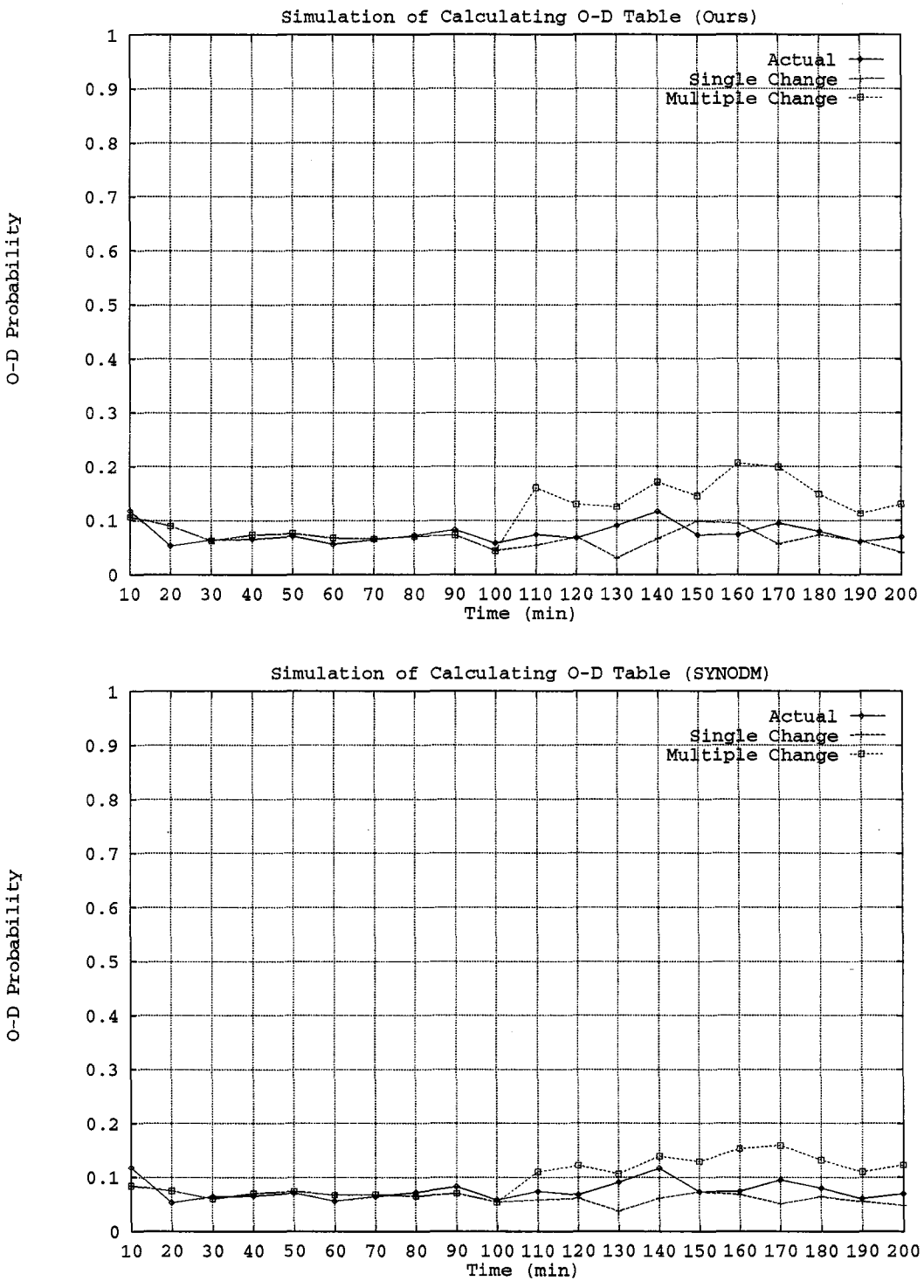


Figure 21: Synthesized O-D probabilities of O1-D7

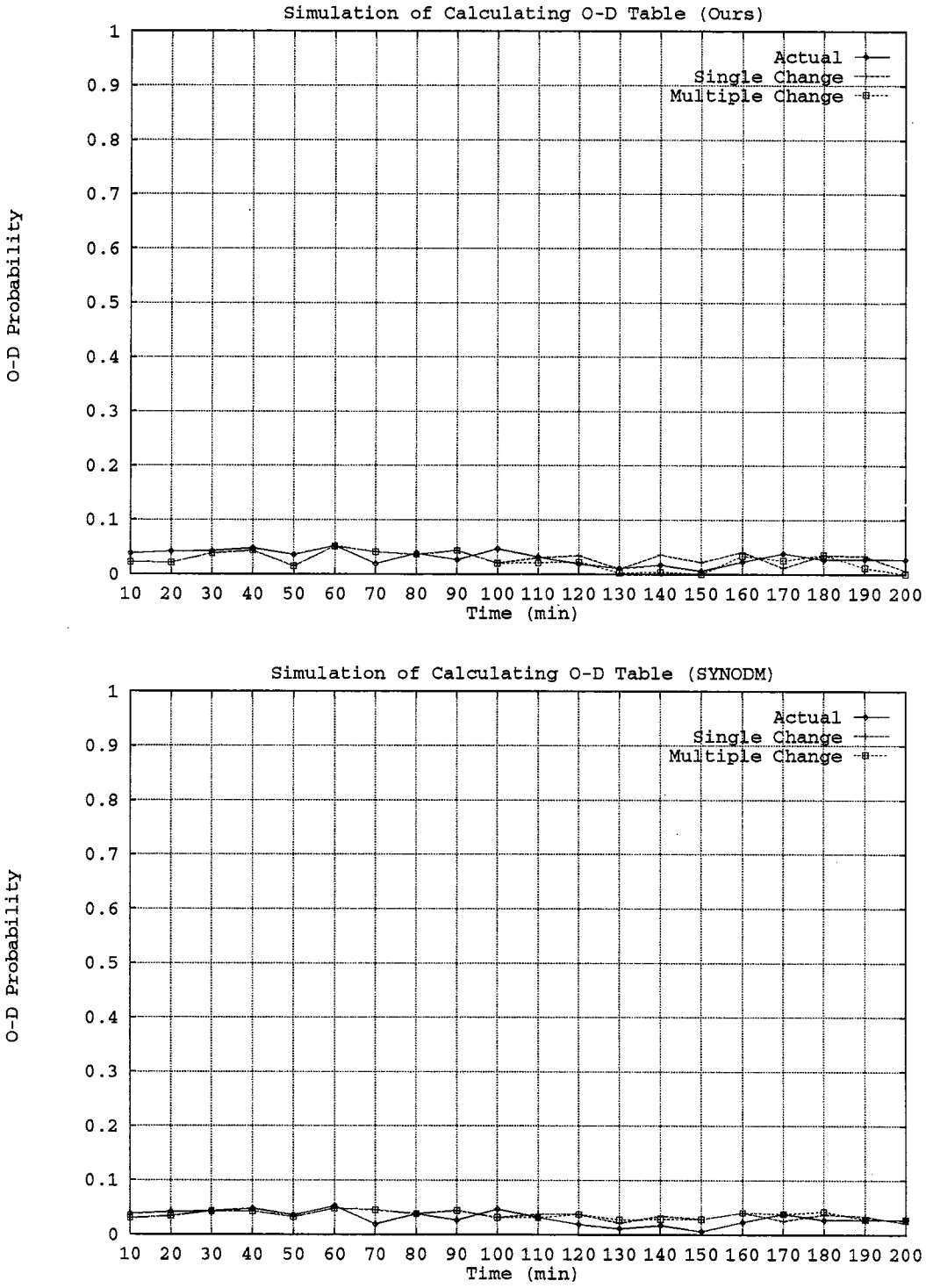


Figure 22: Synthesized O-D probabilities of O1-D8

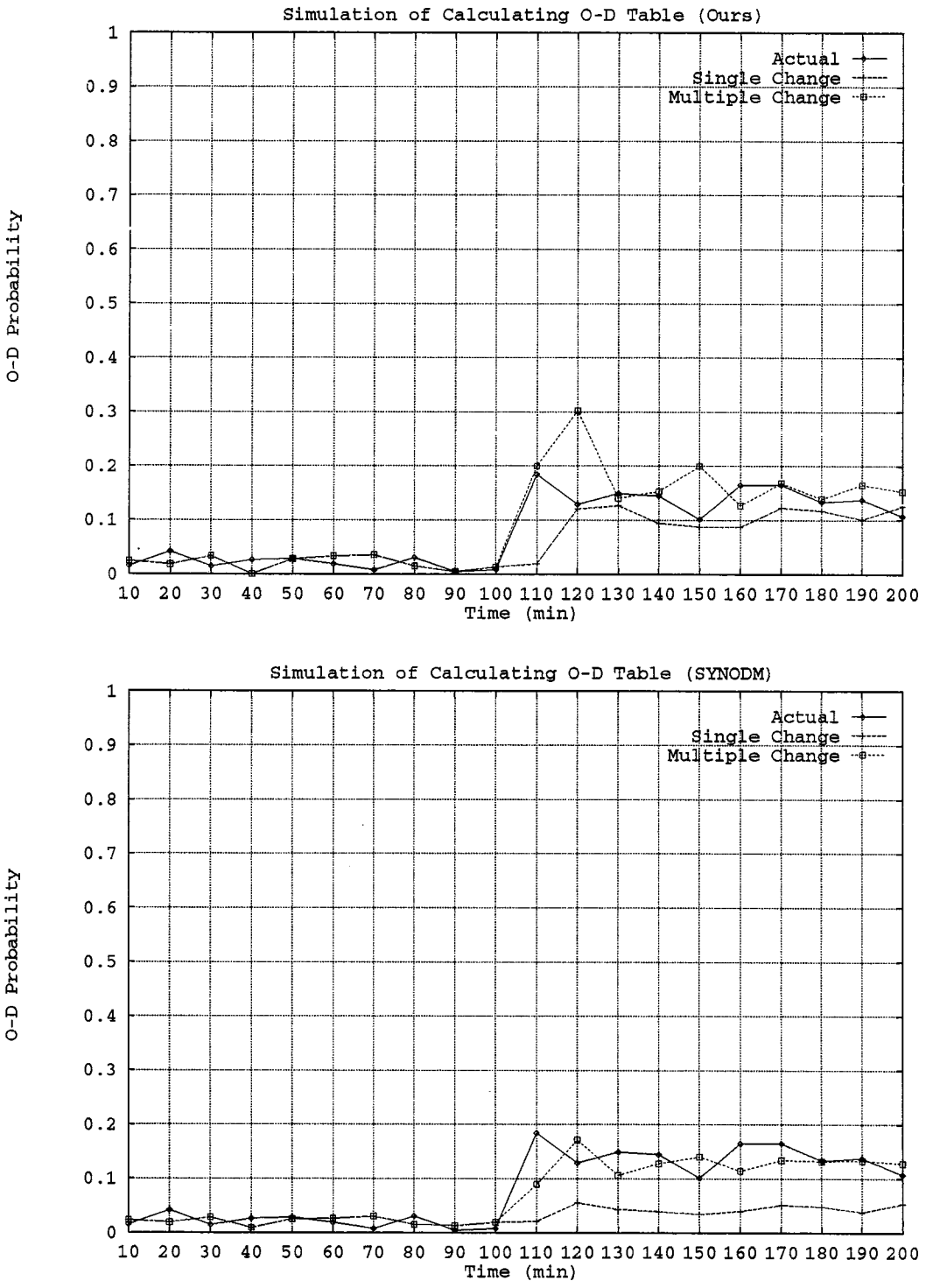


Figure 23: Synthesized O-D probabilities of O1-D9

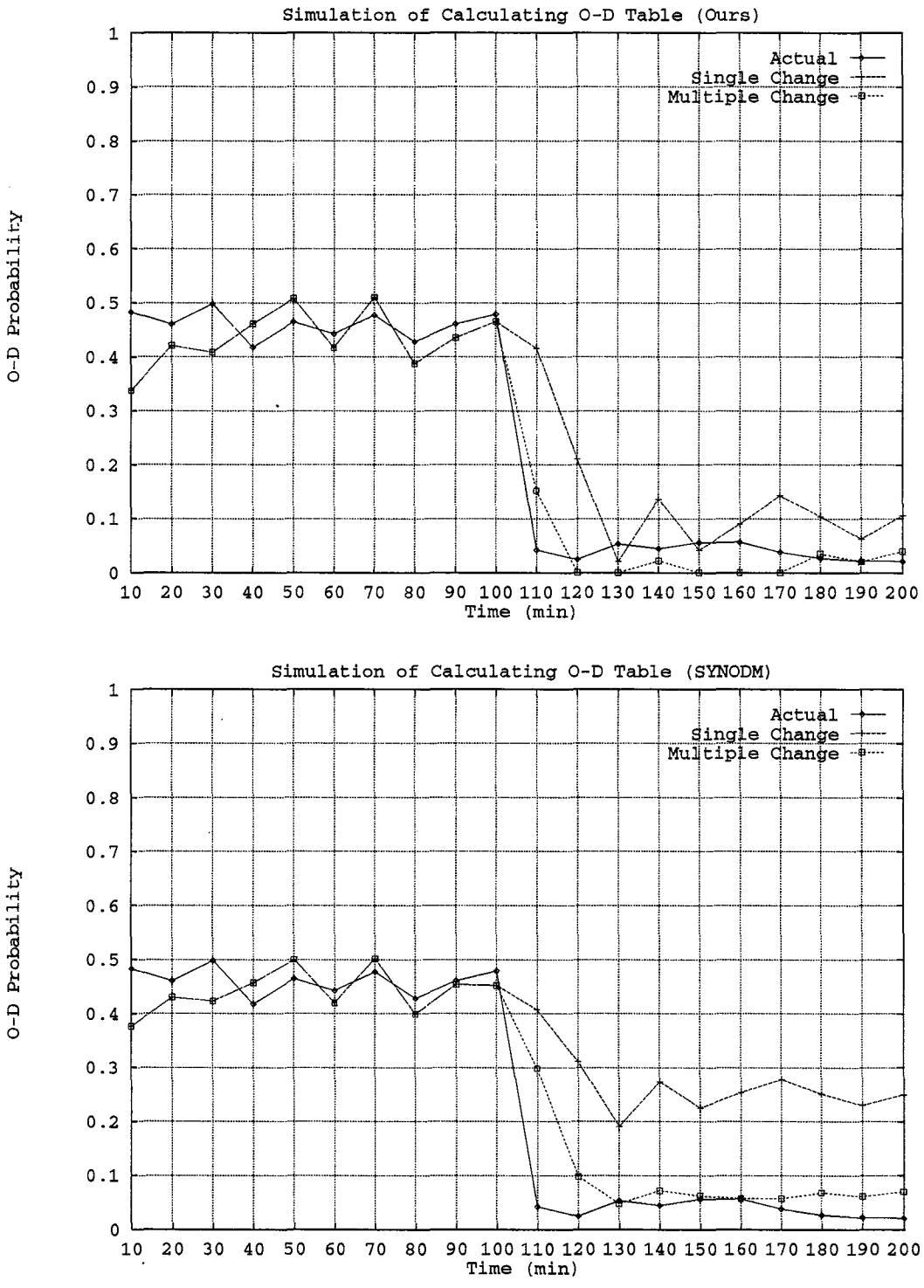


Figure 24: Synthesized O-D probabilities of O1-D10

4.2 BREAKDOWN PREDICTION AND PREVENTION

In the free-flow control algorithm we treat traffic as a smooth flow. However, in a small time window, traffic flow fluctuates due to individual driver behavior or presence of high density platoons. When traffic flow approaches freeway capacity, the chance of flow breakdown due to the short term fluctuation increases quickly. To prevent congestion due to short-term traffic fluctuation, we developed a self-learning flow-breakdown predictor to determine flow breakdown before it actually occurs. Once we predict flow breakdown, the controller can reduce the metering rate of upstream entrance ramps to prevent congestion.

We based the breakdown predictor on the observation that certain traffic patterns have a high probability of causing subsequent traffic-flow breakdowns. These patterns are probabilistic in nature because of factors such as driver behavior, weather conditions, geometric conditions, etc. Therefore, it is very difficult to develop a set of deterministic rules for effective prediction of flow breakdown. Instead, we propose a two-phase procedure for dynamic breakdown prediction. In phase I, we describe historical flow-patterns that actually led to flow-breakdown by certain congestion prediction rules. Then, in phase II, we examine the real-time traffic patterns for prediction of traffic flow-breakdown based on the breakdown prediction rules generated in phase I. Note that flow-breakdown patterns may have quite different behavior in different freeway segments. Therefore, it is highly desirable to combine phases I and II, so the prediction results in phase II can be used to improve the accuracy of the breakdown prediction rules. We shall present a scheme for this in the subsequent discussion.

4.2.1 Fundamentals of the statistical pattern recognition

Let $E = \{e_1, e_2, \cdots, e_n\}$ denote a set of events in a physical system and $S_X = \{X_1, X_2, \cdots, X_m\}$ denote the set of parameters obtained from the system. The pattern recognition problem is to find a distribution of X_i's which can best estimate e_k, $k \in \{1, 2, \cdots, n\}$.

The pattern recognition algorithm consists of three phases: *training, recognition*, and *self-learning.* In the training phase, the relations, or "patterns," between S_X and E are established using knowledge obtained from data collected *a priori.* In the recognition phase,

we compare sampled data to the established patterns to predict events by the *discrimination function.* In the meantime, we use the self-learning process to adjust the patterns so that we can improve recognition results over time.

The first step in the design of such a pattern recognition algorithm is to select the *features*, a parameter set S_x, that can best characterize events of interests in E. For example, we can select occupancy, traffic volume, density, and vehicular speed as features of ramp metering control. The features we select must be able to distinguish different events in the event set with a high probability. We can study the discrimination quality of a feature through repeated experiments using data collected *a priori* [11]. We can measure the quality of features by the degree of correlation between the events and the features. Figure 25 demonstrates the three features X_1, X_2, and X_3. The three different curves in this figure represent the density functions of X_1, X_2, and X_3 corresponding to events e_1 and e_2. It can be seen from the figure that X_1 is a good feature; but, X_2 and X_3 are not because the overlap of the two probability distributions $P(X_1|e_1)$ and $P(X_1|e_2)$ is very small. This is not the case for X_2 and X_3.

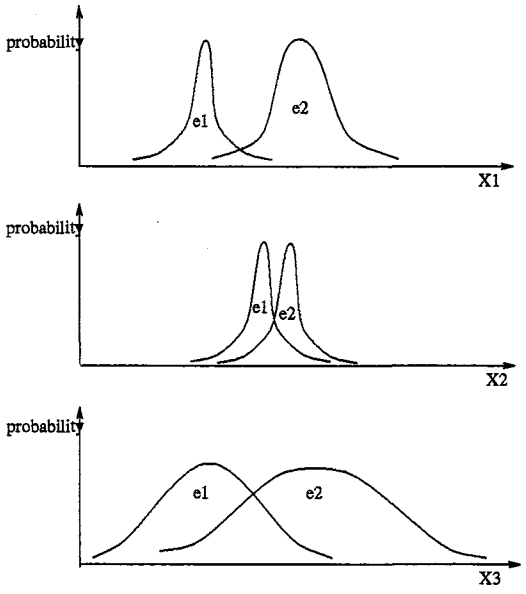


Figure 25: Examples of data distributions experiments for feature selection

Let us denote the domain of a feature X_i as $[a_i, b_i]$. When we use m features to establish the pattern, an m-dimensional feature space represents the complete collection of relevant parameters. Each snapshot at the system state is a point in the feature space. Construction of the discrimination functions for different events are based on the notion of

distance between the sampled features. We can calculate the distance between two points $X_1 = (x_1^{(1)}, x_2^{(1)}, \cdots, x_m^{(1)})$ and $X_2 = (x_1^{(2)}, x_2^{(2)}, \cdots, x_m^{(2)})$ in the feature space in different ways. The two most popular methods are the

Minsky Distance Measure: $D_m = \sum_{i=1}^{m} w_i |x_i^{(1)} - x_i^{(2)}|$,

and:

Euler Distance Measure: $D_e = \sqrt{\sum_{i=1}^{m} (w_1 (x_i^{(1)} - x_i^{(2)}))^2}$,

where $w_i > 0$ are weights normalized to be $\sum_{i=1}^{m} w_i = 1$.

The Euler distance is a geometric measurement of two points in the feature space. The Minsky distance gives the arithmetic difference between two vectors. Once we select the distance measurement, we can define discrimination functions to divide the feature space into several regions for the different events; we describe the boundary between any two events as their discrimination boundary. The regions can be disjoint or overlapping in a probabilistic sense. We can define a typical discrimination function f_i as:

$$\begin{cases} f_i(x_1, x_2, \cdots, x_n) > 0 & \rightarrow e_i \ true \\ f_i(x_1, x_2, \cdots, x_n) < 0 & \rightarrow e_i \ false \end{cases}$$

Figure 26 shows an example which has an event space $E = \{e_1, e_2, e_3\}$ and features X_1 and X_2.

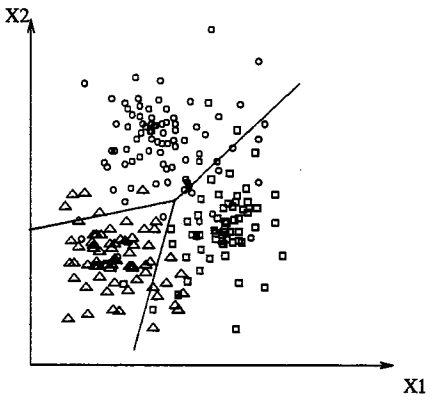


Figure 26: Pattern space

We represent discrimination functions the three curves which divide the space (X_1, X_2) into three regions corresponding to e_1, e_2 and e_3. In the self-learning process, we re-evaluate

the discrimination boundaries based on the new prediction results we obtained through on-line operation.

We use the prediction algorithm proposed by Nihan and Berg to demonstrate the application of pattern recognition technique to predict breakdown. In their algorithm, the event set consists of two exclusive events: *bottleneck* and *non-bottleneck*. A bottleneck occurs in a highway section which satisfies the breakdown conditions. To predict the flow-breakdown in section S_i as shown in Figure 27, their algorithm continuously examines the traffic pattern in sections S_{i-1}, S_i, and S_{i+1} preceding breakdown in S_i. Two parameters, X_1 and X_2, are selected as features, where X_1 denotes the loop occupancy averaged over the past minute for S_i, and X_2 denotes the difference between the input volume to S_{i-1} minus the output volume from S_{i-1}, lagged two minutes.

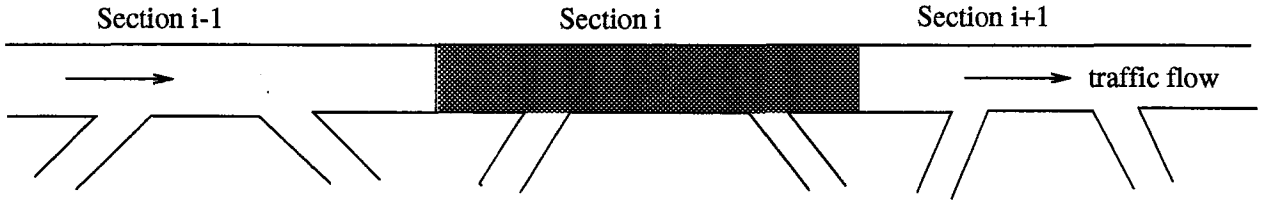


Figure 27: Breakdown prediction in Nihan's approach

We construct patterns based on n groups of historical traffic data that we know to cause flow-breakdown (see Figure 28). Data in S_i is the occupancy value over the control interval (1 min) and data in S_{i-1} is the I/O difference. We select n data sets from the historical traffic data as follows. In the historical data of S_i, we consider flow breakdown to occur at the control time interval $[t, t_c + 1]$ if the occupancy of the section is greater than 18% and the I/O difference is positive. Although the two parameters X_1, and X_2, are dependent on each other they assume that both X_1 and X_2 follow the normal distribution with their mean and standard deviation μ_1, μ_2, and σ_1, σ_2, respectively.

In the next step, they define a discrimination function to predict the two events: bottleneck or non-bottleneck. That is, if

$$\int_0^{x_1} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\pi\sigma_1}} dx_1 < 75\%$$

and

$$\int_0^{x_2} \frac{1}{\sqrt{2\pi}\sigma_{X_2}} e^{-\frac{(x_2-\mu_{X_2})^2}{2\pi\sigma_{X_2}}} dx_2 < 75\%,$$

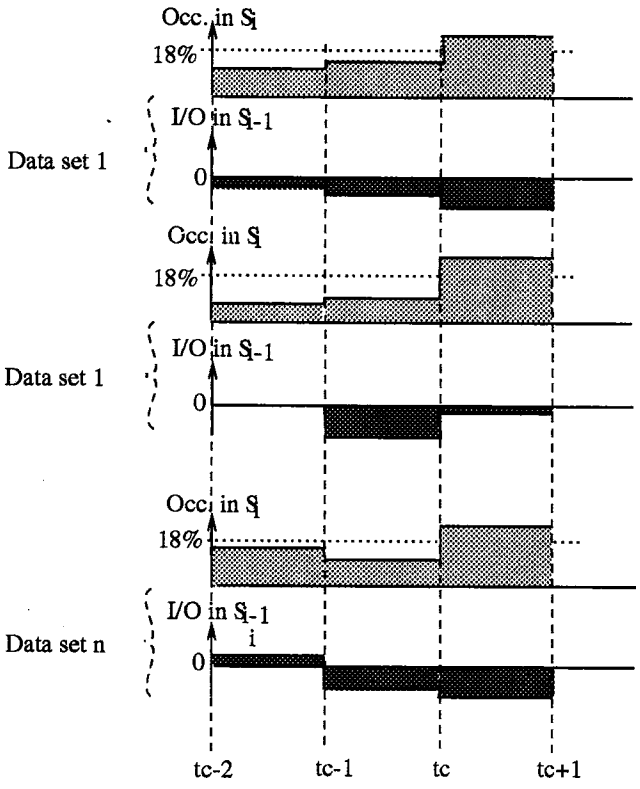


Figure 28: Traffic patterns defining bottleneck

we predict non-bottleneck. Otherwise, we predict a bottleneck, where x_1 and x_2 are the measured values of X_1 and X_2. In other words, the prediction rule is:

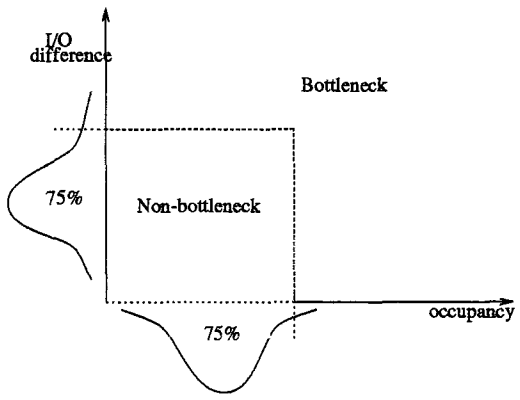


Figure 29: Discrimination boundary between a bottleneck and non-bottleneck

if *the measurement of either occupancy value or the I/O difference exceeds 75% of the bottleneck distribution,*

then *we predict a bottleneck*

Figure 29 illustrates this rule. This prediction rule reportedly has a good accuracy in heavily loaded situations in their experiment. One of the potential deficiencies of this algorithm is that it makes predictions based on snapshots. Because of the dynamic nature of traffic flow, snapshots may not constitute reliable patterns for breakdown prediction. Figure 30 illustrates two cases both of which the bottleneck algorithm predicts as bottlenecks. However, flow-breakdown is less likely to occur in the second case, and thus, it may not be beneficial to make a positive prediction.

To overcome the deficiency of the above approach, we propose to model the traffic flow as a stochastic process for flow-breakdown prediction. The stochastic process-based model can accommodate dynamic fluctuation of traffic flow for more reliable prediction of flow-breakdown. This is so because we can distinguish case (b) from case (a) of Figure 30, so that, only in case (a) we will make a positive prediction.

In our scheme, we characterize the prediction pattern for a particular event, e.g., flow-breakdown, by a *stochastic finite state machine* (SSM), which Figure 31 illustrates. This SSM consists of three states $\{S_0, S_1, S_2\}$, where S_0 corresponds to the lightly loaded situation, S_1 denotes the moderately loaded situation and, S_2 represents the heavily loaded situation. The transition from S_i to S_j represents the probability of a state change in the traffic situation.

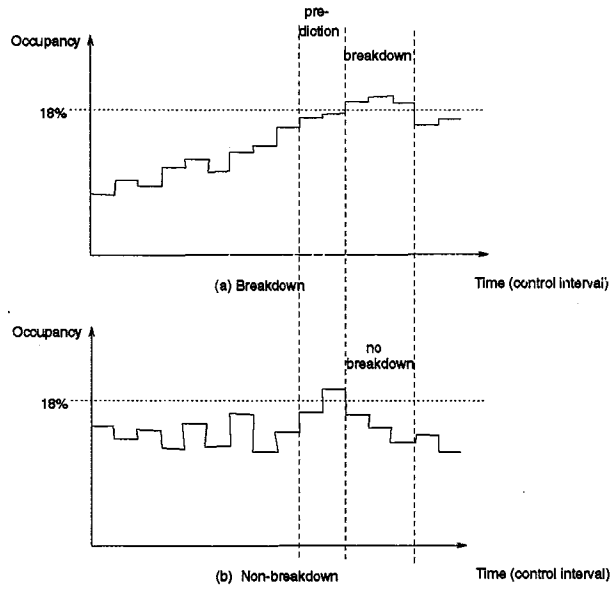


Figure 30: Nihan and Berg's algorithm may predict a bottleneck for both situations

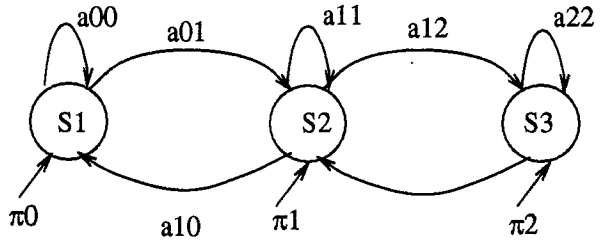


Figure 31: Transition diagram between states

Therefore, no deterministic boundary exists between S_i and S_j since a transition between them occurs probabilistically.

With regard to the SSM, we depict the basic structure of a self-learning breakdown prediction system in Figure 32. In this system, SSM_1 and SSM_2 predict the two events: *breakdown* e_1 and *non-breakdown* e_2. After the SSM is trained, it takes the pre-processed real-time data as inputs for on-line prediction of events e_1 and e_2. The two SSMs calculate the probability of the two events and pass the results to a central decision-making algorithm. The algorithm makes the final judgment on the likelihood of the two events. For self-learning purposes, we keep the prediction results and the associated traffic data for further refinement of the SSMs.

Assume that we use the traffic data in W consecutive control intervals, called a *frame*, for

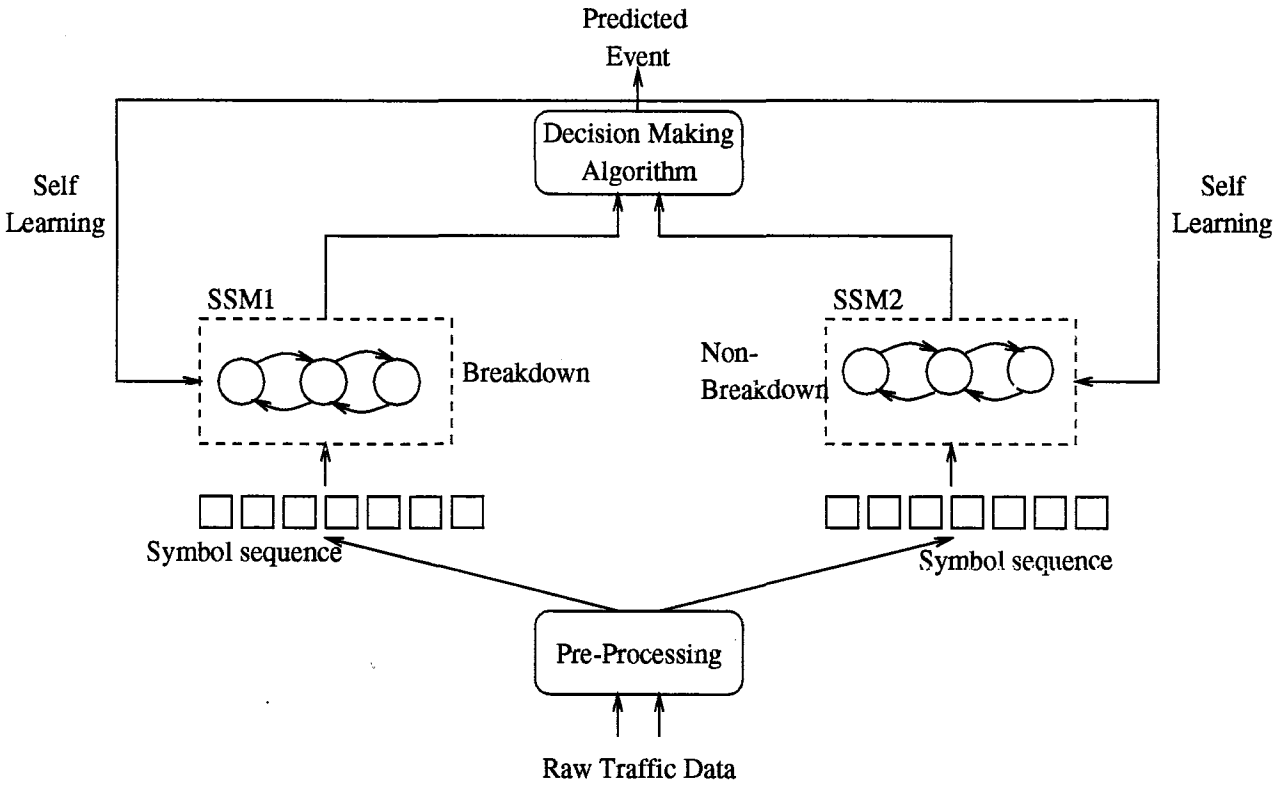


Figure 32: System diagram of a breakdown predictor

congestion prediction, where we determine W from traffic characteristics. If W is excessively long, the congestion prediction may become too slow. On the other hand, if W is too short, the accuracy of prediction may be susceptible to short-term traffic fluctuations.

In the training phase of SSM_i, a set of data frames known to lead to the occurrence of e_i are used to generate (train) the transition i model for SSM_i. We can characterize SSM_k, for $k = 1, 2$ (see Figure 31) by three parameter sets: $\pi = \{\pi_0, \pi_1, \pi_2\}$, $A = [a_{ij}]_{3\times3}$, and $B = \{B_1, B_2, B_3\}$. π_i is the probability that the initial state is i, a_{ij} is the state transition probability from S_i to S_j, and B_j is a probability distribution function of features. The training process optimizes these parameters for a given set of data frames by maximizing the probability of correct prediction for each event. The traffic data in a frame consists of six parameters: the normalized flow rates and occupancies at the upstream and downstream detectors, and the normalized entrance and exit flow rates. Let $\mathbf{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ depict the set of the six parameters, where $x_1 = \frac{q_{upstream}}{c}$, $x_2 = \frac{q_{downstream}}{c}$, $x_3 = \frac{q_{entrance\ ramp}}{c_{ramp}}$, $x_4 = \frac{q_{entrance\ ramp}}{c_{ramp}}$, $x_5 = OCC_{upstream}$, and $x_6 = OCC_{downstream}$. The domains of these six

parameters constitute a feature space and $x \in \mathbf{X}$ depicts a point in the feature space. Note that $\mathbf{X}$ is defined for illustration purposes. For actual implementation we use parameters for $\mathbf{X}$ selected otherwise.

It is inconvenient to use traffic data directly from the detector for construction and manipulation of SSMs since they have continuous values. To simplify computation, we first quantize the traffic data so we can assign to each state in SSMs a discrete probability distribution function instead of a continuous one. *Vector quantization* is a popular quantization technique. In this technique, we divide the six-dimensional space into n regions, where n is a predetermined integer reflecting the granularity of quantization. The centers of the n regions are called the *codebook* of the pattern. We present next the basic steps to generate the optimal codebook for the n regions from m, $m \gg n$, historical data vectors.

1. Initialize the n center vectors in the codebook as evenly distributed in the feature space.
2. Classify each of the m data vectors into the region whose center has the shortest distance to the data vector. Derive the quantization error.
3. Define the new center of each region as the average of all data vectors in the region.
4. Repeat steps 2 and 3 until the quantization error decreases further.

After we divide the feature space into n regions, we denote each one of the regions by one of the n symbols $\{\xi_1, \xi_2, \cdots, \xi_n\}$. A probability mass function $B_j = \{b_j(k), k = 1, 2, \cdots, n\}$ characterizes each state S_j $(j = 1, 2, 3)$ in an SSM. $b_j(k)$ denotes the probability that the SSM is at state S_j when the input is ξ_k, and $\sum_{k=1}^{n} b_j(k) = 1$. During on-line operation, we convert all detectors' data into one of the n symbols. We can, thus, represent the real-time traffic data measured in W consecutive intervals by a sequence of symbols $\mathbf{O} = O_1 O_2 \cdots O_W$, where $O_i \in \{\xi_1, \xi_2, \cdots, \xi_n\}$. For X_i sampled at the ith interval, O_i is the symbol whose center has the smallest distance to X_i than any other centers. From a probabilistic viewpoint, two sequences with a small difference should produce the same event prediction. This is a unique capability offered by the SSM model.

In the training process, we need to estimate $\{\pi, A, B\}$ based on the traffic data collected *a priori*. Given a set of quantized historical data $\mathbf{H}_{e_1} = \{\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_{\mathcal{N}}\}$, $\mathbf{O}_i = O_1^i O_2^i \cdots O_W^i$,

where each sequence in $\mathbf{H}_{e_1}$ led to flow-breakdown, we show the process to derive parameters of SSM_1. We can apply the same process to derive parameters in SSM_2 with some minor modifications. The main optimization criterion of SSM_1 is the *maximum likelihood* principle. That is, we determine the parameters in SSM_1 such that for the given historical data set $\mathbf{H}_{e_1}$, we get:

$$SSM_1^* = \arg \max_{SSM_1} \prod_{\ell=1}^{\mathcal{N}} P(\mathbf{O}_\ell | SSM_1),$$

where $P(\mathbf{O}_\ell | SSM_1)$ is the probability for a particular configuration of SSM_1 to generate the symbol sequence $\mathbf{O}_\ell$. we determine $P(\mathbf{O}_\ell | SSM_1)$ by the initial state, the state transition probability and the probabilistic distribution of each state in SSM_1. That is,

$$P(\mathbf{O}_\ell | SSM_1) = \sum_{i_1 i_2 \cdots i_W} \pi_{i_1} P_{i_1}(O_1^\ell) a_{i_1 i_2} P_{i_2}(O_2^\ell) \cdots a_{i_{W-1} i_W} P_{i_W}(O_W^\ell). \tag{15}$$

A straightforward approach to estimation of π_i, a_{ij}, and $b_j(k)$; $i, j = 1, 2, 3$; $k = 1, 2, \cdots n$, is an exhaustive matching of the symbols and the probability distributions of each state. That is, for $\mathbf{O}_\ell \in \mathbf{H}_{e_1}$, we need to first find a globally optimal matching between each O_j^ℓ and one of the three states in the model. Then, we derive the different probability measurements by the following counting routine. Let $C_j(k)$ denote the total number of times ξ_k appears in state S_j for $\mathbf{O}_\ell$, $\ell \in \{1, 2, \cdots, \mathcal{N}\}$, we get $b_j(k) = \frac{C_k}{\sum_k C_k}$. To derive the state transition probability of SSM_1, we count the total number of state transitions between two adjacent states in each sequence $\mathbf{O}_\ell$, $(\ell \in \{1, 2, \cdots, \mathcal{N}\})$. Let C_{ij} denote the total number of state transitions from S_i to S_j, we define $a_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}$. Similarly, let F_i denote the number of times that S_i is the first state in the different sequences, we get $\pi_i = \frac{F_i}{\sum_j F_j}$.

The approach mentioned above is straightforward; however, its time complexity is prohibitively high. To get the optimal solution at a reduced time complexity, we use the following approach. We can express the probability that a given sequence $\mathbf{O}_\ell$ is produced by SSM_1 in terms of the probability of the t_{th} symbol in $\mathbf{O}_\ell$ as:

$$P(\mathbf{O}_\ell | SSM_1) = \sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t \in \{1, 2, \cdots, T\}, \tag{16}$$

where a) $\alpha_t(i)$ denotes the *forward probability* of generating the realized sequence in $\mathbf{O}_\ell$ up to step t, when SSM_1 is at state S_i at t; b) $b_j(O_{t+1}^\ell)$ is the probability that at step $t+1$,

the state becomes S_j and we observe symbol O_{t+1}^{ℓ} is observed at $t+1$; and c) $\beta_t(i)$ is the *backward probability* that the subsequently realized sequence is in $\mathbf{O}_\ell$. Note that both $\alpha_t(i)$ and $\beta_t(j)$ in Equations 17 and 18 have a recursive form, thus, we can express them as:

$$\begin{cases} \alpha_1(i) = \pi_i b_i(O_1) & (i = 1, 2, \cdots, 3) \\ \alpha_{t+1}(j) = \sum_{i=1}^{3} \alpha_t(i) a_{ij} b_j(O_{t+1}) & (t = 1, 2, \cdots T-1) \text{ and } 1 \leq j \leq 3, \end{cases} \tag{17}$$

$$\begin{cases} \beta_t(i) = 1 & (i = 1, 2, \cdots 3) \\ \beta_t(i) = \sum_{j=1}^{3} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) & (t = T-1, T-2, \cdots, 1),\ 1 \leq i \leq 3. \end{cases} \tag{18}$$

Note that our ultimate goal is to maximize $\prod_\ell P(\mathbf{O}_\ell | SSM_1)$ by finding the optimal A, B, and π. By the total probability property, we have $\sum_{j=1}^{3} a_{ij} = 1$, $\sum_{k=1}^{n} b_j(\xi_k) = 1$, and $\sum_i \pi_i = 1$. We can model the derivation of the optimal A^*, B^*, and π^* as a nonlinear optimization problem by utilizing the following two lemmas to obtain the solution:

Lemma 1: If $P = F(x_1, x_2, \cdots, x_K)$, $x_i \geq 0$ is a real function, and $\sum_{i=1}^{K} x_i = 1$, then we can compute the optimal point $\bar{X} = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_K\}$ of P iteratively by:

$$x_j^{(n+1)} = \frac{x_j^{(n)} \frac{\partial P}{\partial x_j}|_{X=X^{(n)}}}{\sum_{i=1}^{K} x_i^{(n)} \frac{\partial P}{\partial x_i}|_{X=X^{(n)}}},$$

where $X = (x_1, x_2, \cdots, x_K)$ and $\bar{X} = \lim_{n \to \infty} X^n$.

Proof: This Lemma can be easily proven by using the Lagrangian Multiplier method [29].

Let $G(x_1, x_2, \cdots, x_K) = F(x_1, x_2, \cdots, x_K) + \lambda(1 - (x_1 + x_2 + \cdots, + x_K))$ where λ is a multiplier. Thus G and F have the same optimal solution since $\sum_{i=1}^{K} x_i = 1$. Take the partial derivative of G with respect to x_i $(i = 1, 2 \cdots, K,)$ we have

$$\frac{\partial G}{\partial x_i} = \frac{\partial F}{\partial x_i} - \lambda = 0.$$

That is,

$$\frac{\partial F}{\partial x_i} = \lambda, \ \forall i.$$

Hence

$$\sum_{i=1}^{K} x_i \frac{\partial F}{\partial x_i} = \lambda \sum_{i=1}^{K} x_i = \lambda.$$

We can thus get the recursive formula in the Lemma.

Lemma 2: If $P = \prod_{l=1}^{\mathcal{N}} P_l$, $P_l = F_l(x_1, x_2, \cdots, x_K)$, and $\sum_{i=1}^{K} x_i = 1$, $x_i \geq 0$, we can calculate the optimal point $\bar{X}$ of P iteratively as:

$$x_j^{(n+1)} = \frac{\sum_{l=1}^{\mathcal{N}} \frac{1}{P_l} x_j^{(n)} \frac{\partial P_l}{\partial x_j}|_{X=X^{(n)}}}{\sum_{l=1}^{\mathcal{N}} \frac{1}{P_l} \sum_{i=1}^{K} x_i^{(n)} \frac{\partial P_l}{\partial x_i}|_{X=X^{(n)}}},$$

where $\bar{X} = \lim_{n \to \infty} X^n$.

Proof: Since $P_l > 0$, by taking the logarithmic value of both sides of the equation we get $P = \prod_{l=1}^{\mathcal{N}} P_l$, and we get

$$\ln(P) = \sum_{l=1}^{\mathcal{N}} \ln(P_l). \tag{19}$$

Since $\ln(P)$ is a non-decreasing function with respect to P, P and $\ln(P)$ have the same optimal solutions. After taking the differentiation of the both sides of Equation 19, we have $\frac{\partial P}{\partial x_i} = P \sum_{l=1}^{\mathcal{N}} \frac{1}{P_l} \frac{\partial P_l}{\partial x_i}$. By plugging this equation into the recursive formula of Lemma 1, Lemma 2 is proven.

Note that $\Pi_\ell P(\mathbf{O}_\ell | SSM_1)$ has an identical format as the expressions in Lemma 2. Therefore, based on the two lemmas, and by calculating $\partial P(\mathbf{O}_\ell | SSM_1)/\partial x$, where x can be any one of π_i, a_{ij}, or $b_j(k)$ in Equations 16, 17, and 18, we get:

$$a_{ij} = \frac{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell | SSM_1)} \sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell | SSM_1)} \sum_{j=1}^{3} \sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}, \tag{20}$$

$$b_j(k) = \frac{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell|SSM_1)} \sum_{\substack{t=1 \\ O_t=k}}^{T} \alpha_t(j)\beta_t(j)}{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell|SSM_1)} \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}, \tag{21}$$

and

$$\pi_i = \frac{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell|SSM_1)} \pi_i b_i(O_1)\beta_1(i)}{\sum_{\ell=1}^{\mathcal{N}} \frac{1}{P(\mathbf{O}_\ell|SSM_1)} \sum_{i=1}^{3} \pi_i b_i(O_1)\beta_1(i)}. \tag{22}$$

Through a similar process, we can create A, B, and Π for SSM_2 based on another set of data. Once we obtain a_{ij}'s, π_i's, and $b_j(k)$'s for SSM_1 and SSM_2, we are ready to make on-line flow breakdown prediction. During on-line operation, SSM_1 and SSM_2 both accept pre-processed data to calculate probabilities of two traffic events: breakdown and non-breakdown, based on Equation 16. In each control period, we use the current traffic data $\mathbf{x}$ and the traffic data in previous $W-1$ control intervals to calculate probabilities $P(\mathbf{O}|SSM_1)$ and $P(\mathbf{O}|SSM_2)$ by SSM_1 and SSM_2, respectively. We predict a congestion if $P(\mathbf{O}|SSM_1) > P(\mathbf{O}|SSM_2)$, otherwise we predict no-breakdown. In the meantime, we collect the prediction results for further improvement of the prediction patterns. We can implement this self-learning process on-line or off-line similar to the training process.

Now that we know how to make predictions on flow-breakdowns, we design the controller as follows. If we predict that flow breakdown in a freeway section, we upgrade the level-of-service of the section. For example, if the current traffic flow in a section is v_1 which is at the level-of-service E, we try to decrease the flow rates so it will be in the level-of-service D which corresponds to flow rate v_2 ($v_2 < v_1$). This means that the flow coming from upstream should decrease by $v = v_1 - v_2$. We realize this control decision by decreasing the metering rates of up to h upstream ramps, so the total rates decrease is equal to v. To reflect distance effects of entrance ramps, we assign a weighted factor to each section, such that, the rate reduction in the nearest ramp is maximum.

4.3 CONGESTION RESOLUTION

4.3.1 Background

In order to define the traffic condition as congested or non-congested, most of the existing algorithms employ occupancy as the prime control variable [3, 13]. If the current measurement of occupancy does not satisfy the pre-defined set of non-congested traffic conditions, then we say that congestion occurs [22, 23]. Cook and Cleveland [4] carried out an analysis of thirteen control variables for congestion detection, and they found that the most effective ways to detect congestion are either volume, occupancy, or station discontinuity of kinetic energies in neighboring stations.

When a flow breakdown occurs, congestion may propagate in the form of a shock-wave towards downstream, upstream, or both directions, depending on the traffic conditions [15]. In order to resolve congestion, we need to consider the traffic conditions both on freeway and surface streets. If we assign a low metering rate to related entrance ramps, then we may resolve the congestion in a shorter time, but the surface street traffic may suffer from long blocking time at the entrance ramps. On the other hand, if we set metering rates too high, we may aggravate the congestion on the freeway. Therefore, we need to balance the needs for congestion dissipation and surface street service in the congestion resolution strategy.

Not many publications consider entrance ramp control for congestion dissipation. Moreover, most existing solutions do not consider balancing of congestion dissipation time and surface street service. In Chen and May's study [3], an operator takes over the control of traffic manually with the objectives: 1) to maximize overall throughput, 2) to maintain safe conditions in the congested area, and 3) to minimize adverse impact on local streets. Ritchie and Prosser [26] present a simplified version of the algorithm developed in Seattle, Washington [10] for the resolution of incidents. In their scheme, if the degree of severity of an incident is higher than a specified threshold, they close all the upstream ramps within a predetermined distance from the incident site. An operator needs to decide the threshold and upstream distance values. Another proposed system [6] also needs assistance from the operator. It uses an *advisory metering* [28] which informs drivers of the traffic condition and the availability of less congested alternate routes through traffic signs.

Some algorithms only take the current congestion condition into account without considering the condition of upstream sections of the congested area. Gartner and Reiss proposed a scheme [7] in which they calculate metering rates based on the predicted demands at the entrance ramps and the downstream capacities. In the gap-acceptance approach [5, 14, 8], the entrance ramp allows vehicles to enter the freeway when its detectors find gaps in the mainline traffic stream. The major disadvantage of this approach is that they control entrance ramps independently ignoring interdependencies with other ramps and, thus, may not be effective in dissipating congestion.

Wattleworth presented a scheme [31] in which he proposed to solve the congestion dissipation problem using a linear programming model. He employs this model both in congested and non-congested situations. In congested situations, a revised capacity flow rate replaces the free-flow capacity in the linear programming model; then, new desired merging rates are calculated. This is feasible only under the assumption that steady-state conditions exist in the congested area. Hence, when the traffic exhibits a transient condition, such as a transition from a non-congested state to a congested state or vice-versa, this scheme may become ineffective. Besides all the problems discussed above, the notion of control area is not well-defined in the literature. Chen and May [3], for example, presented a scheme in which they employ a fuzzy definition of control area. They do not control congestion occurring outside the boundary of control area until the impact of the congestion propagates to the control area.

The bottleneck algorithm [19] determines the control area dynamically according to the severity of congestion. This algorithm determines ramp metering rates solely by detection (or prediction) of bottlenecks, implying that the system meters a ramp only when it detects congestion. It resolves congestion by reducing the upstream entrance ramp rates until the I/O difference of the congested area reduces to zero. To determine the metering rates, the algorithm uses weighting factors adjusted by an operator for each entrance ramp. Hence, the control area includes all upstream entrance ramps whose rates are deducted. The bottleneck algorithm does not consider reducing volumes in the congested area, but tries to maintain the current volume since it deducts only the amount of I/O difference of the congested area from the upstream entrance ramp rates. Thus, the vehicles already stored in the area are not considered. This approach does not take into account the surface street conditions, either. Thus, the system may generate excessively restricted metering rates which may cause severe

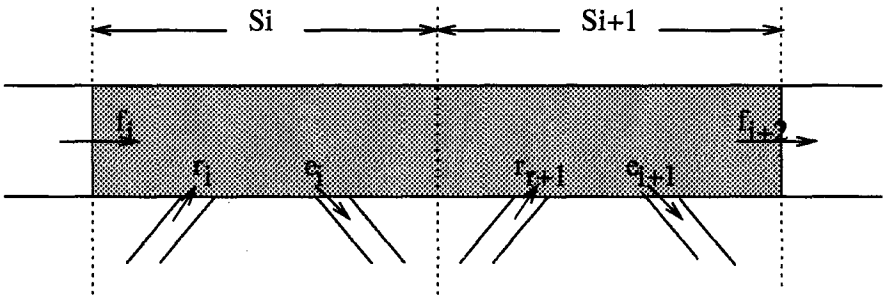


Figure 33: Illustration of F_I and F_O

congestion on the surface streets.

In summary, the problems of most existing congestion or incident resolution algorithms include: (1) they require operator assistance for efficient operation which is quite expensive and unreliable, (2) they do not consider balancing between the congestion dissipation time and the adverse effects on the surface streets, (3) the control area is not clearly defined, and the upstream or downstream traffic conditions are not considered. We present next a congestion resolution algorithm to overcome the aforementioned deficiencies.

4.3.2 Congestion Resolution Algorithm

For simplicity, we define one freeway section as the atomic unit in our congestion resolution algorithm, in which we define a congestion area as the set of sections whose midstream detectors report that the occupancy is higher than the threshold.

We execute the congestion resolution algorithm after detecting a congested freeway section. The first step in the resolution of congestion is to determine the nature of the congestion. Let F_I denote the total input flow to the congested area from the entrance ramps and the freeway mainline, and F_O denote the output flow from the congested area. For example, in Figure 33, $F_I = f_i + r_i + r_{i+1}$ and $F_O = e_i + e_{i+1} + f_{i+2}$, where the shaded area represents the two congested sections. Let us denote the estimated number of vehicles in the congested area which we obtained from traffic detectors as Q. We call the congestion condition as the *shrinking congestion* if $F_O > F_I$, whose congestion duration T_D can be simply estimated as $\frac{Q}{F_O - F_I}$. We call the congestion condition as *growing congestion*, if $F_I > F_O$. Note that we cannot determine the congestion duration before F_I becomes smaller than F_O. Due to

their different characteristics we need two different resolution strategies for the growing and shrinking congestion conditions. In order to strike a balance between congestion dissipation time and the adverse effects on the surface streets, we use a *utility function* which reflects these two factors.

It is relatively easier to cope with the shrinking congestion than the growing one. In the first step, we determine the control area for congestion resolution. Ramp metering controllers in the control area then determine and distribute their metering rates cooperatively. On the other hand, to cope with growing congestion we determine dynamically a utility function based on the tradeoff between the congestion reduction rate and queue lengths of entrance ramps. Then we determine the optimal control area and the metering rate of each entrance ramp to meet the control objective.

Let us first discuss resolution of shrinking congestion. We consider an entrance ramp controller to be within the control area if the congested area will block the vehicle at the head of its waiting queue. To determine whether or not an entrance ramp is in the control area, we need to estimate how long the current congestion will last. Let T_t denote the time required for a vehicle to pass one section, and T_D denote the congestion duration. The control area consists of the already congested sections plus its $\frac{T_D}{T_t}$ upstream sections where vehicles in the entrance ramps of these sections are expected to block.

Now we describe how to decide the total metering rate R of entrance ramps in the control area. Recall that we need a utility function to balance congestion duration T_D and the total ramp queue length Q_R within the control area. Also, F_I is equal to the mainline input flow M_I entering the control area, and the total entrance ramp input flow in the control area is R. Similarly, F_O is equal to the mainline output flow M_O, and the total output flow $\sum e_i$ from all the exit ramps of the control area. Therefore, we can re-express T_D as $\frac{Q}{M_O+\Sigma e_i-(M_I+R)}$. The total queue length Q_R of entrance ramps within the control area is equal to $Q'_R+T(\Sigma v_i-R)$, where Q'_R denotes the current total queue length, v_i denotes the vehicle arrival rate on ramp i, and T denotes the control interval.

To reflect the trade-off between the congestion dissipation time and the adverse effect of surface streets, we define $T_D + wQ_R$ as our utility function for simplicity, where w is a weighting factor. We can define as needed other more sophisticated utility functions.

For efficient resolution of congestion, it is important to maintain the congestion condition as shrinking. Therefore, we consider the derived ramp metering rate as valid only if the congestion is being progressively reduced. That is, we need to ensure that $T_D < T'_D$, where T'_D and T_D are the durations for the current and next control time intervals, respectively. Therefore, R has to be less than $M_O + \Sigma e_i - MI - \frac{Q}{T'_D}$. Moreover, R should not be larger than the total entrance rate of traffic to the entrance ramps, Σv_i. As a result, we can formally express the optimization problem as:

$$\begin{aligned} R^* &= \arg \max (T_D + wQ_R), \\ \text{s.t.} \quad & R^* < \min(M_O + \Sigma e_i - MI - \frac{Q}{T'_D}, \Sigma v_i). \end{aligned}$$

Since $T_D = \frac{Q}{M_O + \Sigma e_i - MI - R}$, and $Q_R = (Q'_R + \Sigma v_i - R)$, it is not difficult to get $R^* = M_O + \Sigma e_i - MI - \sqrt{Q}$ which will minimize the utility function if it satisfies the optimization constraint.

After we determine the total ramp metering rate, we next try to distribute it to entrance ramps in the control area A_k as follows. For the efficient distribution of the total reduction rate we define a cost function for each entrance ramp to reflect its local queue length. Note that in current traffic management regulations vehicles on an entrance ramp should not overflow to the surface street. Thus, we reflect this requirement in our cost function as follows.

Let $C = C_i, i \in A_k$, denote the capacity of every entrance ramp. The current remaining capacity on entrance ramp i is $CA_{current} = C - Q'_i$. If the ramp metering rate is r_i for ramp i, then the remaining capacity for the next control interval is equal to $CA_{next} = C - (Q'_i + (v_i - r_i)T)$. For mathematical tractability, we define the cost function f as the normalized, reciprocal form of CA_{next}, as:

$$f = \sum_{i \in A_k} \frac{1}{(C - (Q'_i + (v_i - r_i)T))/T} = \sum_{i \in A_k} \frac{1}{r_i - a_i},$$

where $a_i = \frac{Q'_i + v_i T - C}{T}$. Based on this cost function, we can directly employ the Lagrange

multiplier method to derive r_i^*, $i \in A_k$, for the minimization of f. Recall that $R^* = \sum_{i \in A_k} r_i^*$, and let $F = f + \sigma R^*$, where σ is the Lagrangian coefficient, we get $r_i^* = a_i + \frac{R^* - \Sigma a_i}{n}$ where n is the number of entrance ramps under control. The derived ramp metering rate r_i^* is feasible only if it satisfies the conditions that $r_i^* > a_i$, and $0 < r_i^* \leq c_{m_i}$, where c_{m_i} is the merging lane capacity.

We now discuss the growing congestion. Resolving growing congestion may not always be feasible, even if we include all upstream sections in the system. To cope with this situation our design philosophy is to gradually resolve the congestion, provided that surface streets are not seriously affected, until it begins to shrink. We can then apply the shrinking congestion control algorithm to resolve the congestion.

To implement this control philosophy, we use factors ΔF_I and ΔQ_R to define the utility function, where $\Delta F_I = F_I$(current interval)$-F_I$(next interval) denotes the change of the in-bound flow, and $\Delta Q_R = Q_R$(current interval)$-Q_R$(next interval) denotes the change of the total ramp queue length of the congested area. For a given R, ΔF_I reflects how fast the growing congestion reduces to a shrinking one, and ΔQ_R reflects how fast the queue will grow.

Figure 34 shows the correlation between ΔF_I and ΔQ_R. Note that F_I increases with R but, ΔF_I with R. On the other hand, Q_R reduces as R increases, but ΔQ_R increases with R. To reduce adverse effects on surface street traffic, we want to increase ΔQ_R, and thus, R.

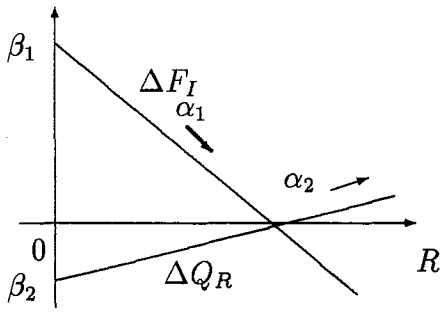


Figure 34: The correlation between ΔF_I, ΔQ_R, and R

On the other hand, to speed up dissipation of congestion, we need to maximize ΔF_I,

thereby, minimizing R. It is not difficult to see in Figure 34 the slope of ΔF_I, α_1, is equal to -1, since the change in total ramp rate is linearly proportional to ΔF_I. Also, the slope of ΔQ_R, α_2, is equal to the value of the control time interval. β_1 is the ΔF_I value when all entrance ramps in the system are totally blocked at the next control interval, and is equal to the current total entrance ramp rate. Likewise, β_2 is equal to $-Q$, where Q is the queue length added by blocking all the entrance ramps.

Based on the above discussion, we define the utility function as:

$$\begin{aligned} utility &= (\Delta F_I)(\Delta Q_R) \\ &= (\alpha_1 R + \beta_1)(\alpha_2 R + \beta_2). \end{aligned}$$

The optimal total ramp metering rate in the system which maximizes the utility function is:

$$R^* = -\frac{\alpha_1\beta_2 + \beta_1\alpha_2}{2\alpha_1\alpha_2}.$$

After we determine the total ramp metering rate, the control area and the metering rate of each entrance ramp within the control area are determined in a different, though very similar, way as in the shrinking congestion case. That is, we define the cost function in the same way as in the shrinking congestion case, but the control area is unknown here. By applying the Lagrange multiplier method to derive f for different values of A_k, the optimal control area A_k is the one with the minimum f value.

We can implement our algorithm in a distributed or centralized way. In the centralized approach, we can designate a controller in the control area as the central controller. The other controllers deliver information, such as, queue length, entrance ramp rate, input flow, output flow, etc. to the central controller. Also, after completing the computation to decide the individual ramp rate at the central controller, the result is sent back to each controller. Therefore, the centralized scheme has a low communication overhead. In the distributed approach, each controller computes its own ramp rate following the same procedure, based on the traffic data gathered from other controllers. However, according to our algorithm, a particular controller within the control area has to obtain the required traffic data from

all the other controllers through communication networks. This scheme is more resilient to failures than the centralized scheme; but, it has a high communication and computation overhead.

4.3.3 An Example for the Congestion Resolution Algorithm

Now we illustrate the congestion resolution algorithm through an example, in which each section is one mile long and has one entrance and exit ramp. We assume that the control time interval T is one minute long, and the average travel speed of a vehicle is 40 miles/hour. At the current time instant, two sections are identified as congested; the number of vehicles stored in the congested area is Q =140.

Figure 35 depicts a snapshot of this traffic system with the shaded area depicting the two congested sections. Numbers in entrance ramp boxes represent the current queue lengths. For simplicity, we assume that all the upstream sections of S_{i-3}, including S_{i-3}, have the

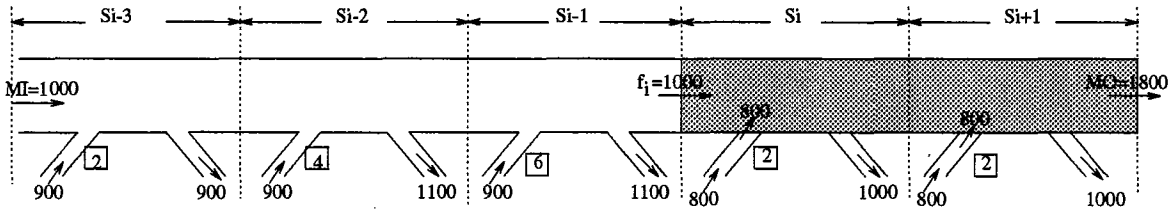


Figure 35: An example of the shrinking congestion

same measurement on entrance and exit ramp rate and ramp queue length. If the congestion is a shrinking type, then we can calculate the ramp metering rates for the next control time interval as follows:

1. $F_I = f_i + r'_i + r'_{i+1} = 2600$, $F_O = M_O + e_i + e_{i+1} = 3800$.
2. $T_D = \frac{140}{1200} \approx 7$ minutes, thus, the control area is from S_{i-5} to S_{i+1}.
3. Although we calculate the optimal total ramp metering rate as $R = M_O + \Sigma_{j=i-5}^{i+1} e_j - M_I - \sqrt{Q} \approx 7688$, R must be smaller than $<$ minimum(7688,6100). Therefore, we set R^* at 6100.

4. $a_{i+1} = 920 - 60C$, $a_i = 920 - 60C$, $a_{i-1} = 1260 - 60C$,
   $a_{i-2} = 1140 - 60C$, $a_k = 1020 - 60C$ for $i - 5 \leq k \leq i - 3$.
   If C=6, then we can get $\Sigma_{j=i-5}^{i+1} a_j = 4780$.

5. The ramp metering rates are :
   $r_{i+1}^* \approx 749$, $r_i^* \approx 749$, $r_{i-1}^* \approx 1089$, $r_{i-2}^* \approx 969$, $r_k^* \approx 849$ for $i - 5 \leq k \leq i - 3$.

Figure 36 depicts a similar example for the growing congestion case which we now consider.

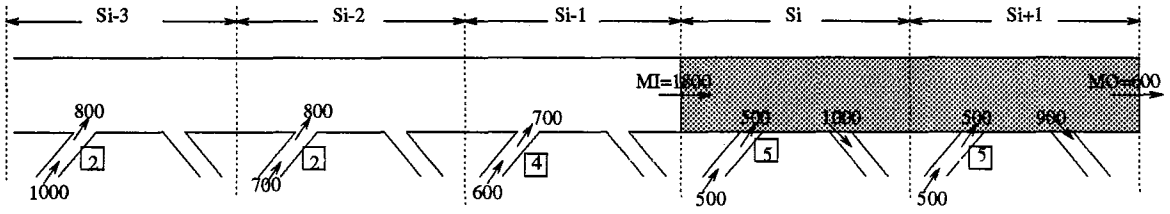


Figure 36: An example of the growing congestion

We calculate the ramp metering rates for the next control time interval as follows:

1. $F_I = M_I + r_i' + r_{i+1}' = 2800$, $F_O = M_O + e_i + e_{i+1} = 2500$.

2. To find the optimal value of R for the utility function, we first calculate
   $\alpha_1 = -1$, $\beta_1 = 3300$, $\alpha_2 \approx 0.0166$, and $\beta_2 \approx -55$.

3. The total optimal metering rate of the utility function is $R^* = -\frac{\alpha_1\beta_2 + \beta_1\alpha_2}{2\alpha_1\alpha_2} \approx 3300$.

4. $a_{i+1} = a_i = 800 - 60C$, $a_{i-1} = 840 - 60C$, $a_{i-2} = 820 - 60C$, $a_{i-3} = 1120 - 60C$.
   If $C = 6$, we get $\Sigma_{j=i-3}^{i+1} a_j = 2580$.

5. The ramp metering rates are :
   $r_{i+1}^* = 584$, $r_i^* = 584$, $r_{i-1}^* = 624$, $r_{i-2}^* = 604$, $r_{i-3}^* = 904$.

We show the above calculation for growing congestion for only one iteration. We should execute similar calculation steps for different numbers of sections to get an optimal control area.

We simulated and compared our congestion resolution algorithm with two cases when all input flows to the entrance ramps were free to enter the freeway (No Blocking), and

when all the ramps were completely closed (Total Blocking). The performance indices were congestion duration and the average queue length. Simulation was done for different ranges of density values per congested section assuming that we have four consecutive congested sections initially.

We performed the simulation with the following assumptions and parameters. The density ranges from 70 to 100 in each congested section. We assume that a section is congested if its density is greater than 40. Hence, the measured congestion duration is the elapsed time for a congested section to have a density lower than 40. At the beginning of the simulation, sections S_i to S_{i+3} were congested. The mainline flow rate was uniformly distributed in the range of [E-100, E+100], where E=1100 for S_i, E=600 for S_{i+3} in the first ten minutes, and E=1000 for other sections. For subsequent control intervals, mainline flow rates were calculated while the mainline flow rate coming into the system boundary still follows our assumption of uniform distribution. Congestion was growing for the first 10 minutes and then began to shrink with the out-bound flow rate of S_{i+3} being 1800 vehicles per hour (saturation flow). Input flow rates of the entrance ramps and exit flow rates are uniformly distributed in the range of [800, 1000]. We assumed initially that the ramp metering rate was equal to 900, and the ramp queue length was equal to two, except that $Q_{i-4} = Q_{i-3} = 3$, $Q_{i-2} = 4$, $Q_{i-1} = 5$, and $Q_i = Q_{i+1} = 7$. We plot the simulation results in Figure 37.

As seen in Figure 37, congestion duration increases with the density. As expected, the Total-Blocking strategy has the shortest congestion duration, and the No-Blocking strategy has the longest congestion duration. Note that in the Total-Blocking approach, even though the algorithm resolves congestion in the shortest time, it creates serious congestion on the surface streets. The congestion effect gets worse with increasing traffic density. On the other hand, the congestion resolution time of the No-Blocking strategy is nearly 10 times longer than that of the Total-Blocking strategy. We simulate our algorithm assuming the queue length of entrance ramps is a critical performance factor. In this case, we can reduce the congestion resolution time by 10 minutes, with a minimal impact on the service quality of surface streets. By a simple adjustment to the weighting factor of the utility function, we can speed up the congestion resolution time at the cost of increased average queue length.

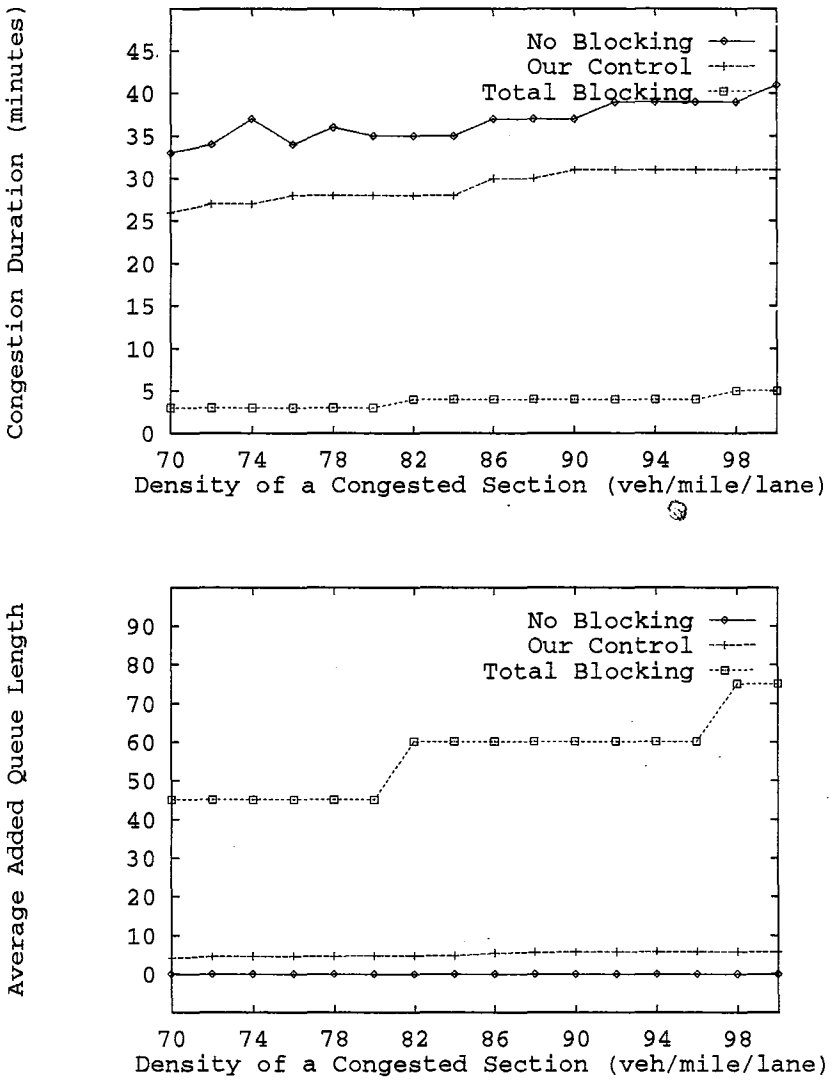


Figure 37: The congestion duration and average queue length of three different schemes when four sections are congested

5 CONCLUSION

In this report we have presented an *Advanced Real-time Ramp Metering System (ARMS).* ARMS consists of three levels of control algorithms integrated for control of freeway ramp metering systems. For each level of traffic control, we proposed flexible control algorithms which can be easily optimized for different local environments.

The first level of control algorithm treats traffic as free flows; the algorithm makes metering decisions of a ramp using the system-wide traffic flow information. Metering decisions of a target area are optimized based on a system-wide objective function which takes into account the correlation between adjacent sections. Our free-flow control scheme differs from existing algorithms by taking into account the risk of congestion when the ramp metering rate (or freeway throughput, equivalently) increases. As a result, we can effectively reduce the possibility of recurrent congestion caused by peak demand, and, thus, we can achieve a better level of freeway service. The algorithm has a modular structure, and it can scale for incremental implementation. Our algorithm for prediction of the O-D traffic distribution is quite accurate and can adapt to traffic pattern changes quickly.

The second level control algorithm deals mainly with congestion prediction. We present a complete solution to the optimal design of a congestion predictor with self-learning capability. This algorithm is able to predict short term traffic flow breakdown in a freeway section so we can prevent congestion before it occurs. We use a stochastic model based pattern recognition technique to capture sequences of traffic patterns for flow-breakdown prediction. Our breakdown prediction algorithm is fundamentally different from the existing solutions such as the one proposed by Nihan and Berg [19]. Our solution improves the bottleneck prediction algorithm by making prediction based on sequences of traffic data instead of snapshots. We can also use our prediction scheme to predict interdependent events which is impossible with existing schemes. Our algorithm begins with an arbitrary set of traffic patterns for congestion prediction, and it improves the accuracy of the prediction with results obtained from on-line operation.

The third level control scheme is a congestion resolution scheme. This algorithm balances the congestion resolution time and the surface street service by proper selection of the control area and metering rates. The dynamic congestion resolution algorithm overcomes drawbacks

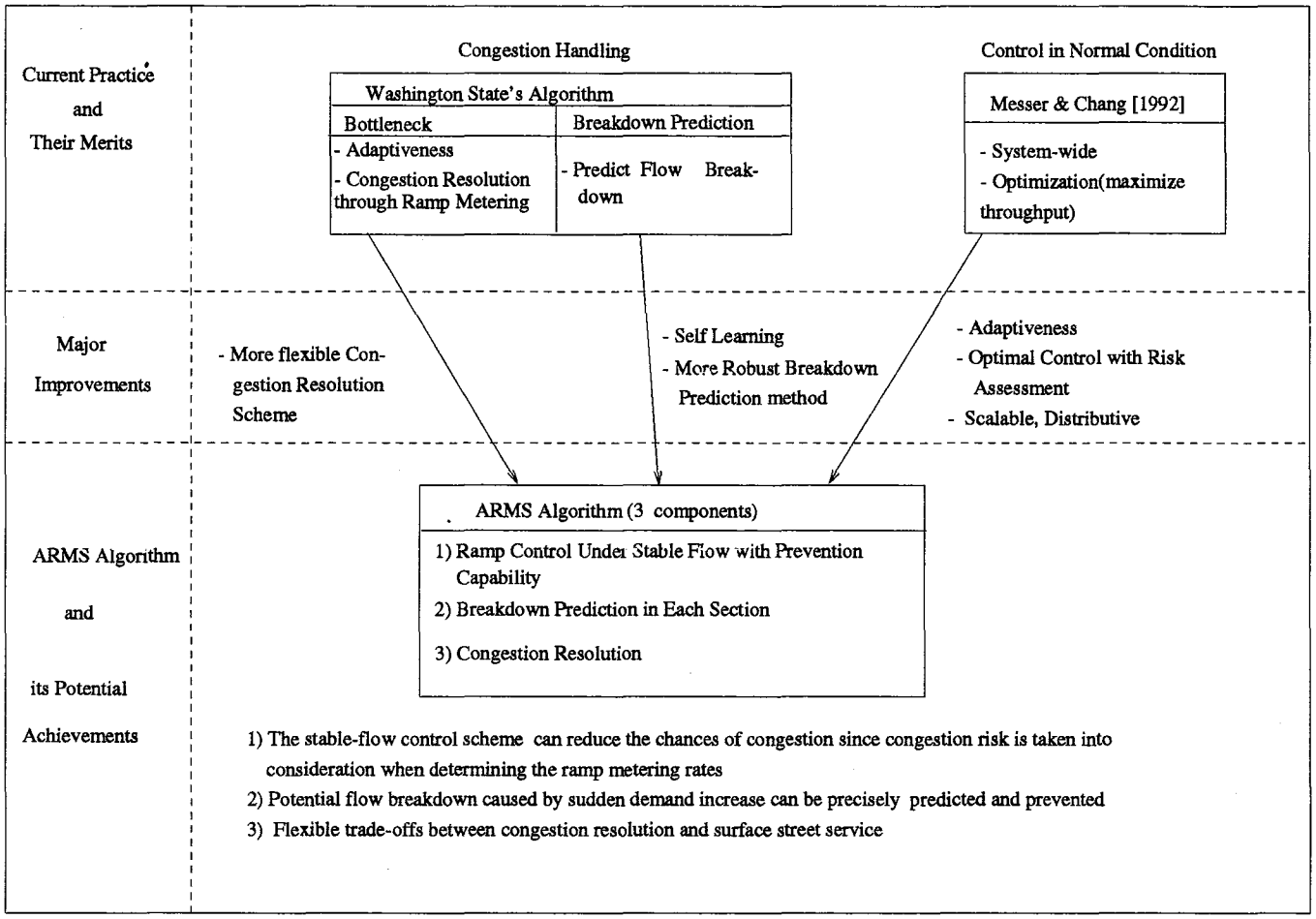


Figure 38: Summary of ARMS Algorithm

of existing approaches by taking into account traffic conditions at both the freeway and surface streets. In this way, the algorithm can resolve congestion based on the local traffic conditions, and possibly by dictations from traffic operators.

Figure 38 illustrates a summary of the technical contributions of the ARMS algorithm. We believe the ARMS provides a comprehensive solution that cohesively integrates different system elements including traffic control algorithms, resilient system architecture, and scalable deployment strategies for implementation of the next generation of freeway ramp metering systems.

REFERENCES

[1] A. J. Ballard, D. Morris, and S. M. Smith. Loop 410 (I-410) Origin-Destination Study, San Antonio, Texas. Technical report, Texas Transportation Institute, February 1985.

[2] D. Chazan and W. Miranker. Chaotic Relaxation. *Linear Algebra and its Applications*, 2, 1969.

[3] L. L. Chen and A. D May. Freeway Ramp Control Using Fuzzy Set Theory for Inexact Reasoning. Research report, University of California, Berkeley, 1989. UCB-ITS-RR-88-10.

[4] A. R. Cook and D. E. Cleveland. Detection of Freeway Capacity-Reducing Incidents by Traffic-Stream Measurements. *Transportation Research Record*, (495):1–11, 1974.

[5] D. R. Drew. *A Study of Freeway Traffic Congestion.* PhD thesis, Texas A&M University, 1964.

[6] Federal Highway Administration. *Integrated Motorist Information System, Phase III: Development of Detailed Design and PS & E*, 1982.

[7] N. H Gartner and R. A. Reiss. Congestion Control in Freeway Corridors: the IMIS System. In *Flow Control of Congested Networks.* Springer Verlag, 1987.

[8] M. E. Goolsby and W. R. McCasland. Freeway Operations on the Gulf Freeway Ramp Control System. Research report, Texas Transportation Institute, August 1969. TTI24-25.

[9] Institute of Traffic Engineers. *Transportation and Traffic Engineering Handbook*, 1976.

[10] L. N. Jacobson, K. C. Henry, and O. Mehyar. A Real-Time Metering Algorithm for Centralized Control. Technical report, Washington State Department of Transportation, November 1988.

[11] M. James. *Pattern Recognition.* New York: Wiley, 1988.

[12] L. Kleinrock. *Queueing Systems: Computer Applications.* New York: Wiley, 1976.

[13] P. H. Masters, J. K. Lam, and K. Wong. Incident Detection Algorithms for COMPASS - An Advanced Traffic Management System. In Vehicle Navigation & Information Systems Conference Proceedings, pages 295–310, Dearborn, Michigan, October 1991. Part1.

[14] A. D. May. Gap Availability Studies. *Highway Research Record*, 72, 1965.

[15] Adolf D. May. *Traffic Flow Fundamentals.* Prentice-Hall, 1990.

[16] W. R. McCasland. Study of Local Ramp Control at Culebra Entrance Ramp on the Southbound IH 10 Freeway in San Antonio. Research report, Texas Transportation Institute, May 1974. TTI165-15.

[17] C. J. Messer and E. C.-P. Chang. Design of Improved Freeway System Ramp Metering Strategies for Texas. Research report (draft), Texas Transportation Institute, September 1992. TTI1232-5-1.

[18] N. L. Nihan. Procedure for Estimating Freeway Trip Tables. *Transportation Research Record*, 895, 1982.

[19] N. L. Nihan and D. B. Berg. Predictive Algorithm Improvements for a Real-Time Ramp Control System. Technical report, September 1991. WA-RD213.1.

[20] D. Owens and M. J. Schofield. Motorway Access Control: Implementation and Assessment. TRRL Research Report 252, June 1990.

[21] M. Papageorgiou, H. Hadj-Salem, and Blosseville J.-M. ALINEA: A Local Feedback Control Law for On-Ramp Metering. *Transportation Research Record*, 1320:58–64.

[22] H. J. Payne, E. D. Helfenbein, and H. C. Knobel. Development and Testing of Incident Detection Algorithms: Vol. 2 - Research Methodology and Detailed Results. Technical report, Federal Highway Administration, Washington, D. C., April 1976. Report No. FHWA-RD-76-20.

[23] H. J. Payne and S. C. Tignor. Freeway Incident - Detection Algorithms Based on Decision Trees with States. *Transportation Research Record*, 682:30–37, 1978. TRB.

[24] B. T. Poljak and Y. T. Tsypkin. Pseudogradient Adaption and Training Algorithms. *Automatic Remote Control*, 3, 1973.

[25] G. P. Ritch. Digital Computer Programs for the Ramp Metering Systems on the Gulf Freeway. Research report, Texas Transportation Institute, August 1971. TTI139-12.

[26] S. G. Ritchie and N. A. Prosser. Real-time Expert System Approach to Freeway Incident Management. *Transportation Research Record*, (1320):7–16, 1991.

[27] B. A. Sanders. An Asynchronous Distributed Flow Control Algorithm for Rate Allocation in Computer Networks. *Transactions on Computers*, 37(7), 1988.

[28] S. I. Schwartz. Traffic Metering of High Density Sectors. In *Strategies To Alleviate Traffic Congestion, Proceedings of ITE's 1987.* Institute of Transportation Engineers, 1987. National Conference.

[29] G. R. Walsh. *An Introduction to Linear Programming.* New York: Wiley, 1985.

[30] J. A. Wattleworth. Peak Period Control of a Freeway System–Some Theoretical Considerations. Chicago area expressway surveillance project, 1963. Report 9.

[31] J. A. Wattleworth. Peak-Period Analysis and Control of a Freeway System. Research report no. 24-15, Texas Transportation Institute, Texas A&M Univ., October 1965.

[32] A. E. Willis and A.D. May. Deriving Origin-Destination Information from Routinely Collected Traffic Counts. Research report, Institutes of Transportation Studies, University of California, 1981. UCB-ITS-RR-81-8.