ELSEVIER

Transportation Research Part B 38 (2004) 869–887

TRANSPORTATION RESEARCH PART B

www.elsevier.com/locate/trb

Optimal freeway ramp control without origin–destination information

Lei Zhang *, David Levinson

Department of Civil Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Received 1 August 2002; received in revised form 14 November 2003; accepted 24 November 2003

Abstract

This paper develops an analytical framework for ramp metering, under which various ramp control strategies can be viewed as ramifications of the same most-efficient control logic with different threshold values, control methods, and equity considerations. The most-efficient control logic only meters the entrance ramps nearest critical freeway mainline sections so as to eliminate freeway internal queues, which is derived from a new formulation of the optimal ramp control problem. Instead of assuming the availability of real-time origin–destination information, the new formulation takes advantages of the stability and predictability of off-ramp exit percentages. Those properties of the off-ramp exit percentages are supported by empirical data, and allow us to formulate the optimal ramp control problem as a linear program whose input variables are all directly measurable by detectors in real-time. The solution is also tested on a real-world freeway section in a microscopic traffic simulator for demonstration. Time-dependent origin–destination tables and off-ramp exit percentages are compared as two alternative ways to represent the true real-time demand patterns that are important to freeway ramp metering.

Keywords: Ramp metering; Freeway operations; OD table; Off-ramp exit percentage

1. Introduction

The first attempt to solve the ramp metering control problem via optimization at the freeway system level goes back to Wattleworth in two papers (Wattleworth, 1963, 1967). The linear program proposed in these papers and its followers (see Lovell, 1997 for a review) are essentially time-invariant optimization minimizing the total travel time in the freeway system. Some distinct

* Corresponding author. Tel.: +1-612-626-0024; fax: +1-612-626-7750.
E-mail address: zhan0294@umn.edu (L. Zhang).

0191-2615/$ - see front matter
doi:10.1016/j.trb.2003.11.003

attributes of these models include: (a) They all incorporate a constraint equation which ensures that freeways operate under free-flowing conditions. Hence, the difficulties of dealing with freeway mainline dynamics are avoided. (b) Time-dependent origin–destination (OD) demand information is assumed to be available. (c) They assume that there are no diversions from freeways to surface arterial streets. Recently, Lovell and Daganzo (2000) extended Wattleworth's steady-state model to include time-dependency, and developed a computationally-efficient greedy heuristic solution. However, the heuristic is only appropriate for small-scale networks, and OD information is still a required input.

There is also a body of literature combining optimal control theory and macroscopic traffic flow models (Chang and Li, 2002; Kotsialos et al., 2002; Papageorgiou, 1995; Zhang and Recker, 1999). The freeway mainline dynamics therein are described by a set of time-discrete equations based on finite difference approximations of specific macroscopic traffic flow models. Control strategies developed along this line also confront the difficulty of getting accurate OD information in real-time. Unclear reliability of the estimated OD information and computational complexity are two drawbacks preventing these strategies from being implemented widely. The models themselves are usually complicated, which makes it hard to solve to global optimality.

On the other hand, numerous operational ramp metering algorithms have been developed in practice. Over the years, ramp metering systems have extended to many urban areas around the globe [1] after its debut in Chicago, IL, in the early 1960s. Surprisingly, every city has its own strategy, some of which are summarized by Bogenberger and May (1999). Simulation evaluation studies on various ramp control strategies become more and more popular. Although many of these practical algorithms are limited in many aspects and based on extensive engineering judgment instead of optimization models, they are successful in reducing total delay as demonstrated by some field experiments (Levinson et al., 2002) and many simulation studies (Hourdakis and Michalopoulos, 2003; Kwon, 2000; Lomax and Schrank, 2000). Local traffic responsive metering algorithms, which base control decisions on real-time traffic data collected in the vicinity of individual on-ramps, have also been very successful without even touching the issue of system-level optimization, no matter what type of local controller is used (Linear: Papageorgiou et al., 1991; Artificial Neural Network: Zhang, 1997; or Fuzzy-logic: Taylor et al., 1998).

This brief retrospective examination of both the research and practice sides of ramp metering brings our attention to several interesting phenomena and questions. First, there is a gap between the state-of-the-art and the state-of-the-practice, and it is not clear which is behind. Researchers have been working on formal optimization problems with assumptions about data availability and complex mathematical models. In contrast, practitioners have developed various *ad hoc* but operational strategies. In our conversation with several engineers in the Minnesota Department of Transportation who oversee the Twin Cities freeway management system, we were told that researchers in the field of ramp metering are far behind practitioners. It is probably true that practitioners developed more real-world congestion-mitigating ramp control strategies on their

[1] US cities with ramp metering systems: Phoenix, AZ; Fresno, CA; Sacramento, CA; San Francisco, CA; San Diego, CA; Denver, CO; Atlanta, GA; Twin Cities, MN; Las Vegas, NV; Long Island, NY; New York, NY; Cleveland, OH; Lehigh Valley area, PA; Philadelphia, PA; Houston, TX; Arlington, VA; Milwaukee, WI; and Seattle, WA; Non-US cities: Sydney, Australia; Toronto, Canada; Paris, France; and Birmingham and Southampton, UK; Kobe, Japan.

own. But what is it that makes those operational strategies successful while few formal theories or models are adopted by practitioners? People working with real-world metering systems long ago drew the conclusion empirically that the most benefits of ramp metering were from the decision to have a metering system, but not from the sophistication of the algorithm (Newman et al., 1969). Is that true and why?

The second observation is that in theoretical development of optimal ramp control solutions, researchers tend to use time-dependent OD tables to represent true freeway demand patterns, which help formulate the problem mathematically. Many researchers have considered the optimal ramp control problem as a two-stage problem implicitly or explicitly: an accurate and efficient real-time OD estimation procedure should be developed, and then that information can be used as input to the following optimization stage. However, in practice, few operational control strategies use OD tables. There must be some way other than OD tables that the practitioners take care of the true demand patterns. Their success implies that the alternative method is somewhat reasonable. What is it?

Finally, we have seen the following trend in ramp metering studies. The mathematical models become more and more complicated as the scope is expanded from local to coordinated and integrated control. However, the only validation step taken seems to be simulating the final product of all research efforts—the resulting ramp control strategy. If the simulation shows positive results, the theoretical model or procedure is considered as acceptable. However, this reasoning process could be dangerous because sometimes very crude metering algorithms (e.g. a pre-timed) can also significantly reduce total delay. Simulation studies can show whether one strategy outperforms another, but do not shed light on how and why that is the case. If satisfactory efficiency performance is obtained in one control strategy, is it because the traffic flow model successfully predicts real traffic conditions, or because an OD estimation procedure is incorporated, or because equity is put on a low priority, or something else? If we do not pursue answers to the how and why questions, successful simulation, even field evaluation results, do not necessarily imply that the underlying theory is superior. Therefore, an analytical framework under which those questions can be explored is clearly in order.

This paper, as a small step to address the questions and research needs identified in the above discussion, develops an analytical framework for ramp metering studies, with the hope of leading towards a more unified and generic ramp control theory. Under this framework, various individual elements that constitute a complete ramp control strategy can be easily decomposed and studied separately. We formulate the optimal ramp control problem without using the time-dependent OD tables to represent true demand patterns. Instead, the stability of off-ramp exit percentages is studied and used in the analysis. The solution to the new formulation reveals that the most-efficient ramp control logic is actually a very simple one, which to some extent explains the success of many operational ramp control strategies. In this regard, we hope that the paper can also help bridge some of the gaps between research and practice in both directions. A simulation experiment is executed only to demonstrate that the core ramp control logic in the analytical framework can be implemented in real-time. The findings also reveal that the most efficient control strategy is also the least equitable one. Considering the enormous political and public interests on balancing efficiency and equity of ramp meters, this topic is also briefly discussed.

The remainder of the paper is organized as follows. The next section (Section 2) proposes an analytical framework for ramp metering studies. The following section (Section 3) is devoted

exclusively to two alternative ways of considering real-time freeway demand pattern in ramp metering—OD tables vs. off-ramp exit percentages. In order to complete the construction of the analytical framework, a ramp metering logic is required. Therefore, Section 4 details our formulation of the optimal ramp control problem, the solution of which can serve as the control logic. Thanks to the stability and predictability properties of off-ramp exit percentages, it is able to formulate the problem as a linear program. The simulation experiment on the solution to the linear program is described in Section 5, followed by a discussion on equity considerations in Section 6. Conclusions and suggestions for future studies are delivered at the end of the paper.

2. An analytical framework for ramp metering

Many existing or proposed ramp control strategies avoid internal queues on the freeway (i.e. freeway mainline sections operate at free-flow conditions). It has been shown that, in the time-independent ramp control problem, preventing the formation of internal queues is a necessary condition that the optimal solution must satisfy (Wattleworth, 1967). However, in the time-dependent case the benefits of allowing internal queues are not clear. The analytical framework that will be developed herein is suitable for ramp metering controls not allowing freeway internal queues. When implementing a ramp metering logic, one has to specify the following elements to form a complete operational algorithm:

2.1. Threshold values—the "capacity" of each freeway section

The "capacity" of all critical sections must be specified. Since freeway breakdown is essentially a probabilistic phenomenon (Persaud et al., 1998), one can be risk averse and set critical values in the lower tail of the breakdown probability distribution to minimize the probability of failing to ensure that freeway mainline section flows must be strictly lower than capacities (see Fig. 1).

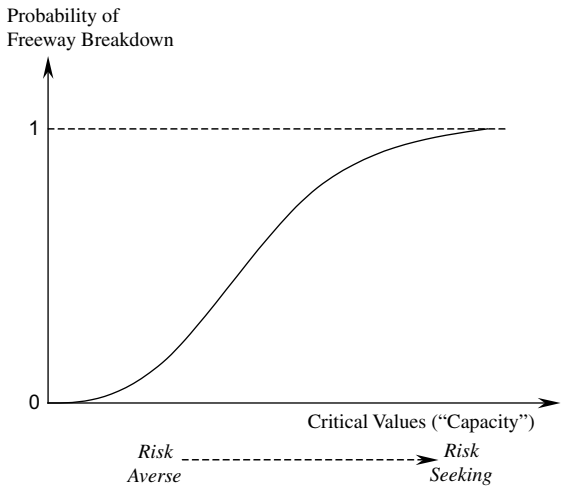

Fig. 1. Two types of critical values.

However, these somewhat smaller critical values may lead to an efficiency loss. On the other hand, a set of more aggressive critical values can be used with a higher risk of freeway breakdown. This risk-seeking strategy may in the long run improve the overall efficiency of the controlled freeway system.

The "capacity" here can be either flow thresholds or density thresholds depending on the control method one chooses (see Section 2.2). In the case of flow-capacity, most of the existing ramp control strategies tend to adopt long-run freeway queue discharging flow rates. The threshold set in this way should be considered as risk-averse decisions, not necessarily optimal values. However, it will remain impossible to take advantage of control methods under uncertainty in the case of ramp metering until the probabilistic nature of freeway breakdown is more thoroughly understood. It should also be noted that ramp metering tends to increase the capacity of freeway bottlenecks as a recent empirical study suggests (Zhang and Levinson, 2003a). Therefore, the threshold values can be better determined by the traffic data when ramp metering is in operation, or a markup over the non-metering thresholds should be used.

2.2. Control methods—how one keeps the flow of a section strictly below capacity

Ramp control strategies keep the flow/density of critical sections below critical values. This is a standard control problem and thereby a control method must be selected. In the case of ramp control, one must specify several control details:

(a) Flow control vs. density control

One can either control flow or control density to achieve the control objective. Earlier ramp control strategies controlled flow, largely because the final control variable, the vector of on-ramp metering rates, was actually a vector of flows, not density. But recently there is an increasing trend of controlling densities. In density control, some rules must be specified to transform densities to final control variables (not necessarily numerically, e.g. fuzzy rules). In a recent study, Zhang and Beegala (2003) show that occupancy thresholds are better indicators of potential breakdown than flow thresholds at one bottleneck, which is evaluated by detection rates and false alarm rates using traffic data collected in more than thirty peak periods. Further studies on other freeway sections can provide additional information. Real-world examples: (Flow control) Twin Cities, Seattle Bottleneck algorithm; (Density control) SWARM, Denver, ALINEA.

(b) Feedback control vs. feed-forward control

If there is a discrepancy between the actual flow/density and the desired threshold values, the controller must take action to eliminate/minimize this difference. If the controller employs a predictor to estimate potential discrepancies and take action before the discrepancies actually occur, this type of control is feed-forward. On the other hand, feedback controllers only adjust control parameters (metering rates) based on the detected differences (already occurred) of the desired threshold values and the observed. A controller using a Kalman filter is a mix of both feed-forward and feedback concepts since it contains both an equation predicting future system states and a feedback equation reducing prediction errors based on measured data. In a sense, feedback controllers are more desirable because they guarantee convergence and do not include any predictive elements. However, to ascertain which type of control is more suitable for ramp

metering, future comparison studies are required. Real-world examples: (Feedback control) ALINEA, ANN, Denver; (Feed-forward control) Twin Cities, SWARM.

(c) Linear control vs. non-linear control

Once a difference term between the desired critical values and the observed/predicted values are identified, will this difference be transformed to control parameters linearly or non-linearly by the controller? For example, in flow control, the simplest controller is the one that directly uses the difference as the metering flow which is a linear controller (*control parameter = difference*). Zhang (1997) found a non-linear controller based on Artificial Neural Network is superior to a standard linear controller in a simulation test. But linear controllers are usually cheaper to implement. Real-world examples: (Linear control) ALINEA, Twin Cities, Denver; (Non-linear control) ANN controller, Fuzzy Rules.

2.3. Purpose of the analytical framework

The analytical framework allows us to view many existing/proposed ramp metering algorithms as ramifications of the same ramp control logic with different threshold values and/or different control methods and/or different equity considerations (equity considerations will be discussed in Section 6). Under this framework, various elements that constitute a ramp control strategy can be decomposed and studied individually. Previous studies have evaluated and compared various ramp metering algorithms in traffic simulators. One algorithm may outperform another according to the simulation results. However, those studies have very limited theoretical implications because with many components in each algorithm, they are unable to answer why the more efficient algorithm is more efficient. Every component, such as the threshold value or the type of the controller, can make a difference on the performance of the entire algorithm. The analytical framework developed facilitates the decomposition of these individual factors, and comparisons can then focus on just one of the many factors keeping all others equal. A research topic under this framework could be, for instance, "all others being equal, is a non-linear controller better than a linear one" or "provided the same threshold values and control methods, how will the efficiency of the control strategy trade off with equity". In the long run, such studies should provide more valuable results than those directly comparing two existing or proposed ramp control strategies.

The last but most important brick of the analytical framework that is missing is a control logic from which various algorithms evolve. Developing the most-efficient control logic, or the solution to the optimal ramp control problem, is the major purpose of Section 4. Before that, some remarks on time-dependent OD tables are discussed, and properties of off-ramp exit percentages studied. They provide necessary background information for the following formulation of the optimal ramp control problem.

3. The trouble with origin–destination tables

For decades, origin–destination tables (OD) have been used as standard tools to represent travel demand patterns in transportation studies ranging from urban travel demand forecasting to

real-time facility management. Researchers have developed numerous methods to estimate such OD information, which is later used to predict or manage traffic conditions on a freeway segment, a corridor, or a transportation network. However, there are at least three pitfalls when the notion of OD information is adopted for real-time, adaptive freeway ramp metering strategies, which we will critique in the following paragraphs. The purpose of the discussion is not necessarily to depreciate the value of the concept of time-dependent OD tables, but to pave a road for alternative ways of formulating and solving the optimal ramp control problem.

3.1. The three pitfalls of using OD tables in freeway operations

The first critique (actually a clarification) we want to make is that there does NOT exist a true time-independent OD table. When a freeway section is of interest, a potential user of that freeway section has a specific destination off-ramp in mind, and chooses a time window to start the trip from an entrance ramp. But travelers do not necessarily stick to their original choices. A driver may change destination after entering the freeway mainline due to perceived congestion, temporal change of activity location, or other reasons. The true demand pattern is related to all these individual travel decisions, and an OD table is simply one way to approximately aggregate these micro-level demand decisions. The extensive application of OD tables in travel demand analysis has elevated its status to be almost equal to the true demand pattern, and all we are doing seems to be estimating the "true" OD table. However, because travelers may change destinations after departure, a true OD table does not exist. There are other ways to represent travel demand. For instance, in agent-based micro-simulation models, individual travel decisions are explicitly considered. Other methods for demand estimation do not have to be interpreted in terms of time-dependent OD tables.

The second critique, recognized by many others, is that it is extremely hard to estimate a real-time OD table for freeway operations. As an aggregate approximation of the true demand pattern, the value of a time-dependent OD table becomes less and less apparent as the time intervals get shorter and shorter. For many freeway ramp control strategies, the control interval is usually shorter than a minute. Is it possible to predict an accurate-enough OD table for the next 30 s? Many studies explore the potentials of using real-time detector counts to estimate real-time OD matrices. However, according to the authors' own experience, the problem is not well formulated and in order to obtain a solution, some kinds of aggregation procedures are unavoidable. The accuracy of the resulting estimates is highly questionable.

Finally, if the OD table is just one way to represent demand patterns and a reliable estimation procedure for it is still elusive, is it because the assumption of the availability of OD information helps solve the optimal ramp control problem mathematically that many introduce OD variables into the formulation of the problem? Unfortunately, the answer is still negative. Although some past studies, ignoring the first-in-first-out conditions at on-ramps, mistakenly formulate the time-dependent freeway ramp control problem with OD information as a linear programming problem, the work by Lovell and Daganzo (2000) has clearly demonstrated that the time-dependent ramp control problem with OD information is highly non-linear and only heuristic solutions are currently available for general freeway systems. A following study (Erera et al., 2002) shows that the discrete-time version of the problem is NP-complete. Therefore, using an OD table does not bring us mathematical advantages.

3.2. An alternative way to estimate real-time demand patterns

All these stimulate one to ask—is there a better way to represent the true freeway demand patterns, which is also predictable from available data with reasonable accuracy, for the purpose of ramp metering. We say yes to the question—there is an alternative.

If the goal of ramp metering is to minimize total travel time, under fixed demand the optimal strategy should maximize total output of the system at any control interval (a formal mathematical proof of this is given in Section 4.2). Then it is the causal effects between on-ramp metering rates (controlled variables) and off-ramp exit percentages that must be known to optimally control freeways. These causal effects can be obtained from an accurate time-dependent OD table if there is one. However, if those causal effects are directly estimable from available data, there is no need to derive them from a real-time OD table, which is estimated from the same data set with a questionable procedure. The key observation of the proposed approach is that at control interval t, a lot of information is already known, such as metering rates (control variables) and off-ramp exit percentages (measurable variables) in all previous control intervals, which can help one estimate off-ramp exit percentages at $t+1$. We understand that the exit percentage at an off-ramp at $t+1$ depends on metering rates of all upstream on-ramps at $t+1$ (let the time be synchronized so there is no need to denote time lags) in a complex way. But we believe that the exit percentage in the next control interval can be reasonably estimated with only the information available at the current control interval for three reasons: (a) Usually, freeway demand patterns change slowly. (b) Keeping metering rates smooth is a standard constraint in practice. (c) The exit percentage depends on metering rates at many upstream on-ramps and mainline demand characteristics at the furthest upstream section, and their combined effects may be averaged out, which can result in some kind of stability in exit percentages. Therefore, it is reasonable to assume that off-ramp exit percentages also change slowly in real-time. Let α denote exit percentages at an off-ramp. This implies that α_{t+1} may be well approximated by $\alpha_t, \alpha_{t-1}, \ldots, \alpha_{t-n}$ where n is a small natural number (e.g. The average α in the previous n control intervals can be an estimate of α_{t+1}). If this time-series estimation method for α_{t+1} based on detector data is reasonable, then there is really no need to use time-dependent OD tables to represent real-time demand patterns, as we shall later see in Section 4.

3.3. Some empirical evidence

Traffic data in the afternoon peak period (14:30–19:30) of a randomly-selected day (November 02, 1999) on a freeway section with 25 off-ramps on Trunk Highway 169 northbound (TH169) in the Twin Cities metropolitan area is examined to evaluate the stability of off-ramp exit-percentages with ramp metering. This freeway section is also used for a simulation test later in the paper and to avoid repetition, its geographical characteristics will only be detailed in Section 5. Loop detectors provide 30-s flow at both off-ramps and freeway mainline sections. Malfunctioning detectors are removed from the analysis. Fig. 2a plots observed exit percentages at a representative off-ramp during a peak period, and demonstrates that exit percentages do not have much variation. Eleven different estimates for the exit percentages in the next 30-s control interval (α_{t+1}) are compared: the average exit percentages in the previous n 30-s control intervals ($n = 1, \ldots, 10$), and the average exit percentage of the whole peak period. Two performance measures are defined

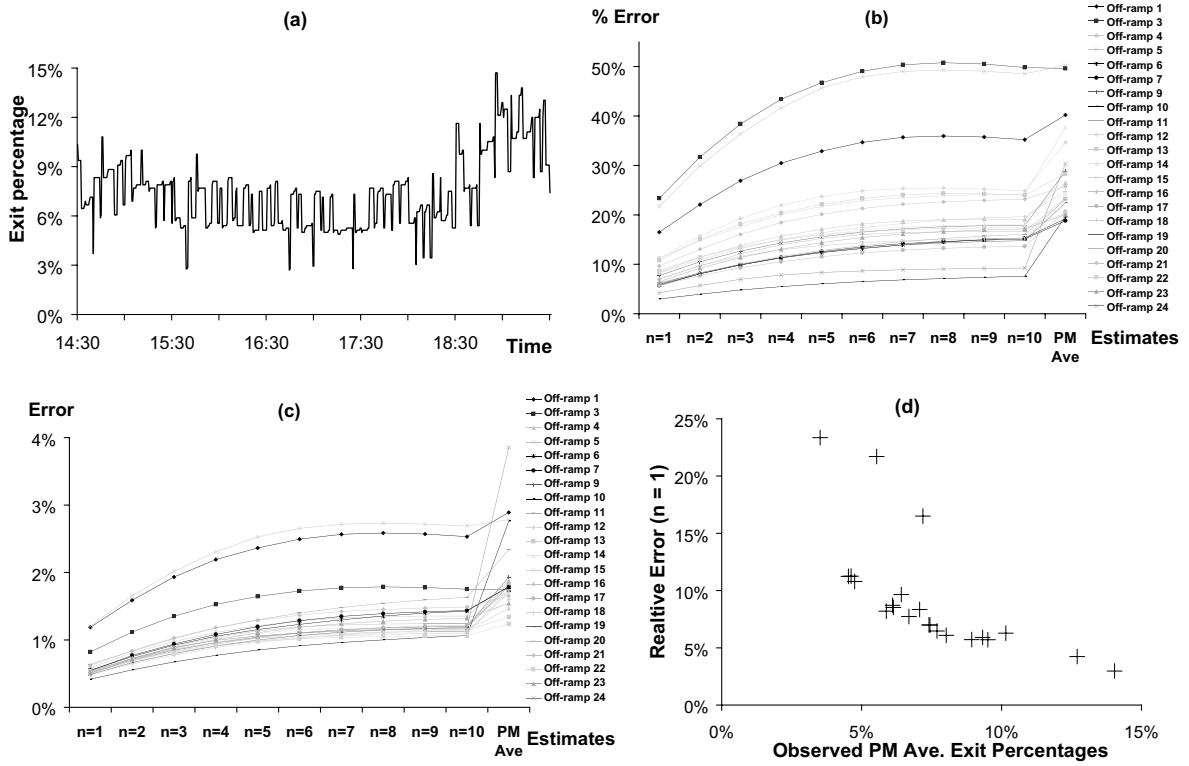

Fig. 2. Estimation methods for exit percentages in the next control interval. (a) Observed exit percentages at an off-ramp in a typical peak period, (b) relative error, (c) obsolute error and (d) relative error vs. exit percentages. *Note*: Off-ramps 2, 8, and 25 were not included due to detector errors.

to evaluate these estimates, absolute and relative errors. For each control interval t (abs(.) calculates the absolute value of a variable):

$$\text{Absolute error} = \text{abs}(\text{estimated } \alpha_{t+1} - \text{observed } \alpha_{t+1})$$

$$\text{Relative error} = \text{abs}(\text{estimated } \alpha_{t+1} - \text{observed } \alpha_{t+1})/\text{observed } \alpha_{t+1}$$

The average relative and absolute errors for each off-ramp during the studied peak period are computed and presented in Fig. 2b and c respectively. It is clear that $n = 1$ is the best estimate at all off-ramps examined. The best estimate for α_{t+1} is simply α_t. With this estimate, the relative errors at most off-ramps are within 10% of the actual exit percentages. In absolute term, the average prediction errors never exceed ±1.2 percentage points at all off-ramps. As the average exit percentages in longer intervals are used, the prediction errors also increase at diminishing rates. However, even the crudest assumption that the exit percentages are constant throughout the peak period only produces relative errors around 20% at most off-ramps. Fig. 2d plots relative errors against average peak period exit percentages. The proposed estimation method predicts exit percentages better at off-ramps with higher flow than those with lower flow.

This empirical evidence supports the presumption that exit percentages in the next control intervals can be reasonably estimated without detailed OD information. We doubt if any existing real-time OD estimation procedure can predict 30-s exit percentages with comparable accuracy. The next section further demonstrates that using exit percentages instead of OD information to formulate the optimal ramp control problem also tremendously reduces the complexity of the mathematical program and the solution algorithm. Some previous studies assume that exit percentages are independent of metering rates. Although the results here somehow support that assumption, we think it is necessary to restate it as follows: the complex dependencies of off-ramp exit percentages on metering rates at upstream on-ramps, together with slow-changing demand patterns and metering rates, allow one to reasonably estimate exit percentages in the next control interval only with known information in the current interval.

4. A theory of optimal ramp control

The analytical framework views various ramp metering algorithms as ramifications of the same ramp control logic. This section formulates the optimal ramp control problem, and provides a solution that can serve as that logic.

4.1. Notation

We will follow these notation conventions hereafter in the paper.

A the furthest upstream flow of the whole freeway system;
B flow on a freeway section measured at the furthest downstream point;
C capacity of a freeway section;
D arrival flow at an entrance ramp;
i index of entrance ramps;
I number of entrance ramps;
j index of exiting ramps;
J number of exiting ramps;
k index of freeway sections;
K number of freeway sections;
M departure flow at an entrance ramp (also the metered flow if the ramp is metered);
q arrival rate for the whole freeway system;
Q departure rate of the whole freeway system;
S standing queue at an entrance ramp, see Eqs. (9) and (10);
t index of time intervals;
t_0 starting time of the ramp metering control period;
T end time of the control period;
X flow at exiting ramps;
α exit percentage at an off-ramp = exiting ramp flow/upstream mainline flow, $\alpha \in [0, 1]$;
γ a $I \times K$ indication matrix, $\gamma_{ik} = 1$ if entrance ramp i is on section k, $\gamma_{jk} = 0$ otherwise;
δ a $J \times K$ indication matrix, $\delta_{jk} = 1$ if exiting ramp j is on section k, $\delta_{jk} = 0$ otherwise;
Δ a $K \times K$ synchronization matrix, $\Delta_{k1,k2}$ = free-flow travel time from section k_1 to k_2;

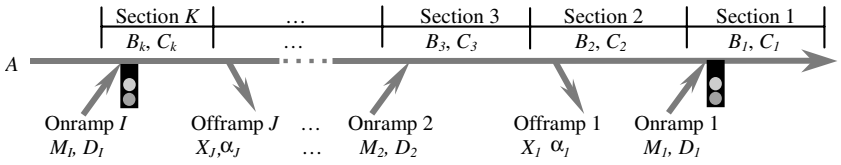

Fig. 3. Coding a typical freeway system.

Fig. 3 illustrates the freeway coding method. A freeway system can be divided into a series of freeway sections (sections 1–K), each of which contains either an on-ramp or an off-ramp. On-ramps could be metered by a particular control algorithm or not metered. A section starts from the location immediately upstream of the on/off-ramp it contains and ends at the starting point of the immediately downstream section. All sections can be categorized into three types: metered-section (e.g. sections 1 and K), unmetered-section (e.g. section 3) and off-ramp-section (e.g. section 2).

4.2. Formulating the optimal ramp control problem

The whole freeway system is considered as a queuing system in which the nth vehicle entering the system may not be the nth vehicle leaving the system. Therefore flow data collected at boundaries of the system do not allow one to track the travel times of individual vehicles since the first-in-first-out (FIFO) conditions are violated. Note that on-ramp queues are still subject to FIFO conditions. The arrival rate of this queuing system q_t at any time interval t is:

$$q_t = A_t + \sum_{i=1}^{I} D_{i,t} \tag{1}$$

The departure rate of this queuing system Q_t is:

$$Q_t = B_{1,t} + \sum_{j=1}^{J} X_{j,t} \tag{2}$$

Total travel time is the area bounded by the arrival curve and the departure curve in the queuing diagram. The objective for optimal ramp metering control is to minimize the total travel time. By assuming a fixed arrival curve, i.e. no route choices before departure, minimizing total travel time is equal to maximizing the area under the departure curve which is the integral of departure rates over the whole control period:

$$\text{Max} \int_{t=t_0}^{T} Q_t \, \mathrm{d}t \tag{3}$$

Papageorgiou (1983) shows that the time-discrete version of this integral is:

$$\text{Max} \sum_{t=t_0}^{T} [(T - t) \cdot Q_t] \tag{4}$$

Substitute Eq. (2) into (4), the objective function becomes:

$$\underset{M}{\text{Max}} \sum_{t=t_0}^{T} \left[(T-t) \left(B_{1,t} + \sum_{j=1}^{J} X_{j,t} \right) \right] \tag{5}$$

This objective is subject to a set of control or physical constraints:

$$B_{k,t} < C_{k,t} \quad \forall k \text{ and } t \tag{6}$$

$$0 \leqslant M_{i,t} \leqslant S_{i,t} \quad \forall i \text{ and } t \tag{7}$$

We also have the following relationships:

$$X_{j,t} = \sum_{k=1}^{K} \left(\delta_{j,k} B_{k,t} \frac{\alpha_{j,t}}{1-\alpha_{j,t}} \right) \tag{8}$$

$$S_{i,t} = S_{i,t-1} + D_{i,t} - M_{i,t-1} \tag{9}$$

$$S_{i,t_0} = 0 \tag{10}$$

$$B_{k,t} = B_{k-1,t-\Delta_{k,k-1}} - \sum_{j=1}^{J} \delta_{j,k} X_{j,t-\Delta_{k,k-1}} + \sum_{i=1}^{I} \gamma_{i,k} M_{i,t-\Delta_{k,k-1}} \tag{11}$$

(5)–(11) complete the formulation of a linear programming form of the optimal ramp control problem (LP). Inequality (6) states that the flows of all freeway sections are not allowed to exceed capacity at any time. Inequality (7) is a physical restriction which states that metered flow rates must be positive and not larger than the current on-ramp demand. Eq. (8) describes the relationship among off-ramp flows, mainline flows and off-ramp exit percentages. Eq. (9) updates the on-ramp standing queues and (10) is the initialization equation. Eq. (11) is a spatially iterative process through which metered flow rates M, the control variables finally come into the objective function after substitution. Although Eqs. (11) and (8) look like two simultaneous equations, they are actually not, as long as the spatially iterative process starts from the furthest upstream section K.

If all input variables are known, the problem can be solved by any standard LP algorithms. However, unless one assumes constant exit percentages at off-ramps (which may not be as crude an assumption as it seems), exit percentages should be updated in every control interval in real-time using the procedure described in the previous section. The benefit of the global version is that it provides globally optimal solution with the assumption that every input parameter is known *a priori*. The real-time version is inevitably myopic, but takes advantage of information that becomes available in the control process. Future information can be predicted, however, it is impossible to accurately predict all input variables for the entire control period in a system as dynamic as freeways. Therefore, we will focus on the real-time version of the new formulation of the optimal ramp control problem.

4.3. A real-time version of the LP

A real-time ramp control strategy at control interval t aims to optimize the system in the next control interval $t+1$ based on all information available at t. A rolling–synchronized–horizon technique is adopted and the global optimization objective function, Eq. (5), becomes:

$$\text{Max}\left(B_{t^s+1} + \sum_{j=1}^{J} X_{j,t^s+1}\right) \quad \text{at } t^s \tag{12}$$

We use superscript s to denote the synchronized time. If the synchronized time t^s at section 1 is the absolute time t, then the synchronized time t^s at section k is absolute time $t - \Delta_{k,1}$. The notion of synchronized time is convenient because it eliminates the complex notation associated with free-flow travel times. Constraints (6) and (7) still hold in synchronized time:

$$B_{k,t^s} < C_{k,t^s} \tag{13}$$

$$0 \leqslant M_{i,t^s} \leqslant S_{i,t^s} \tag{14}$$

In transforming Eq. (5) to (12), dependencies between time slices are thrown out. This ignorance of time dependence allows one to develop a solution procedure without any predictive elements. If there are any benefits to restricting entering vehicles beyond what would be considered optimal during a single control interval, to allow for more efficient usage of the freeway in later control intervals, these benefits will be lost in the above transformation. However, it is reasonable to believe that this potential loss of benefits should not be significant for two reasons: (a) Constraint (6) assures that there are no queues on freeway mainline sections (Rigorously there could be some transient congestion on freeway mainline sections). Therefore, any queues that remain in the system at the end of a control interval and will be dealt with during later control intervals are on-ramp queues. (b) The structure of Eq. (5) reveals that allowing more vehicles to leave the system sooner is preferred to restricting them, since the weights on earlier departures are higher than later ones (the weight is $(T - t)$).

Again, the real-time optimization problem is an LP problem with many fewer control variables compared to the global optimization problem. Standard solution algorithms (Simplex, Interior point, etc.) apply. All input variables, including exit percentages, are directly measurable from currently available traffic detection hardware, and they should be updated in every control interval. However, we further pursue a heuristic solution in the next subsection, obviously not because an optimal solution algorithm is unavailable or inefficient, but because the heuristic solution has very important qualitative meaning.

4.4. A heuristic solution

An intuitive solution to the real-time optimal ramp control problem is developed in this subsection. This simple solution method, with physical meaning easy to understand, very likely provides the maximum system departure rate of all possible solutions.

To facilitate the presentation of the heuristic, we will use a numerical example shown in Fig. 4. Inequality (13) will be referred as constraint 1 and inequality (14) constraint 2 in the following discussion, as well as in Fig. 4. The heuristic consists of a forward process and a backward process (by "forward", we mean the same direction that traffic flows) executed iteratively. In general, the forward process reflects the nature of the optimization objective function, while the backward process is enforced by the two constraints. In the graph, regular fonts stand for either real-time data collected by detector or threshold capacity values preset by the controller. Bold and italic fonts denote values determined by the forward process and the backward process respectively. Again, all

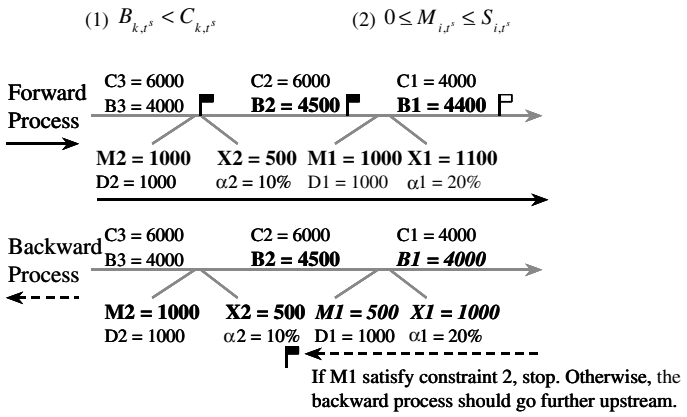

Fig. 4. A numerical example illustrating the heuristic solution procedure.

values in the graph are synchronized-time values. We start the "forward process" (the solid arrow) from the furthest upstream section 3 and let all vehicles waiting at the on-ramps (1 and 2) enter the freeway ($M_1 = S_1, M_2 = S_2$, the right half of constraint 2, $M_i \leqslant S_i$, is guaranteed to be satisfied since all the following steps can only decrease the metering rates). Then given these metering rates, we calculate the flow of the immediately downstream mainline section (e.g. $B_2 = 4500$ based on B_3, M_2 and α_2). Whenever this forward process proceeds to a new downstream section, we check whether constraint 1 is satisfied (e.g. when the forward process goes to section 2, we check if $B_2 < C_2$. A black-flag in the figure means it is satisfied, a white flag otherwise). When in a section, constraint 1 is not satisfied, it becomes a critical section and a "backward process" will be applied starting from this section and going upstream (the dashed arrow). The backward process first adjusts the on-ramp metering rate at the nearest on-ramp (on-ramp 1) to satisfy constraint 1 at the critical section ($M_1 = 500$). The resulting M_1 at this time could be negative (A negative M_1 means that more than one on-ramp needs to be restricted). Then we check whether the left half of constraint 2, $0 \leqslant M_i$, is satisfied at the adjusted on-ramp (on-ramp 1). If it is satisfied, we stop the backward process and start a new forward process from the critical section. Otherwise (negative M_1), the backward process needs to go further upstream (restrict more on-ramps to satisfy both constraints).

The above heuristic solves the real-time version of the LP since all constraints are satisfied. We will then show that this heuristic also gives the optimal solution in the numerical example. The solution we have now is $M(M_1 = 500, M_2 = 1000)$. We denote another solution as $M'(M'_1, M'_2)$. Since M' is a solution, it must satisfy both constraints that apply (this means $M'_2 \leqslant M_2$ since $M_2 = S_2$). Now there are only three possible types of M':

(a) $M'(M'_1 \leqslant M_1$ and $M'_2 M_2$, but $\sim (M'_1 = M_1$ and $M'_2 = M_2))$;
(b) $M'(M'_1 > M_1$ and $M'_2 < M_2)$;
(c) $M'(M'_1 > M_1$ and $M'_2 = M_2)$.

It is obvious that (c) does not satisfy constraint 1 in freeway mainline section 1, so it is not a feasible solution. Also, under (a) the resulting total departure rate of the system is less than that under M and hence (a) is an inferior solution to M.

Whether M is the optimal solution now becomes a comparison between $M'(M_1' > M_1$ and $M_2' < M_2)$ with M. Let $M_1' = M_1 + n$, where n is a positive number. We have $M_2' \leqslant M_2 - n/\alpha_2$ in order to satisfy constraint 1 in section 1. Then the total system departure rate under M and M' can be compared:

$$Q = B_1 + X_1 + X_2$$

$$\text{Under } M : Q_M = C_1 + C_1[\alpha_1/(1-\alpha_1)] + B_2\alpha_2/(1-\alpha_2)$$

$$\text{Under } M' : Q_{M'} \leqslant C_1 + C_1[\alpha_1/(1-\alpha_1)] + B_2\alpha_2/(1-\alpha_2) - n\alpha_2/(1-\alpha_2)$$

$$\text{Therefore} : Q_{M'} < Q_M$$

$M'(M_1' > M_1$ and $M_2' < M_2)$ is also an inferior solution to M. Therefore M is the optimal metering rate for this numerical example. Since our derivation does not depend on the concrete values in the numerical example, this heuristic solution should also be the optimal solution to this sample network with any different parameters. The essential philosophy of the heuristic is that, by metering the nearest on-ramp(s) to the critical freeway section, one minimizes the efficiency loss at off-ramps upstream of the critical freeway section. For a larger network, to show M is still the optimal solution, the logical reasoning we just presented in the numerical example would take more texts to explain because the number of possible solutions increases exponentially. However, the philosophy underlying the heuristic solution procedure should also apply in larger networks.

The heuristic solution has very important qualitative meaning. It states that the most-efficient ramp metering control logic is the one metering the nearest upstream entrance ramp(s) to any critical freeway section so as to keep the flow of this section strictly below capacity. This explains why some local metering algorithms, and coordinated algorithms that specifically target bottlenecks are successful—they are really close to the most-efficient metering logic. The logic also provides some theoretical support to actions such as temporary ramp closure. It seems that practitioners have, probably implicitly, taken advantage of the stability of off-ramp exit percentages to deal with real-time dynamic demand patterns in developing those control strategies. It should be noted that the heuristic itself is apparently not a local metering logic because the backward process requires coordination among several on-ramps.

The heuristic is also very desirable from a computational feasibility perspective. First, it only uses information that has been accumulated to the decision point and no prediction is required. Second, the computation work involved in the heuristic is a straightforward iterative process with simple mathematical operations (only plus and minus) in each iteration step. However, since the original real-time optimal ramp control problem is an LP, the additional computational advantages offered by the heuristic may not be very valuable. Finally, with the least number of on-ramps being controlled to provide free-flow conditions on freeway mainline for all commuters, the most-efficient control logic is also expected to be the least equitable one.

5. A simulation experiment

A ramp control strategy that directly implements the heuristic (most-efficient ramp control logic) is developed using risk-averse critical values and linear-feedback-flow control. This new strategy is coded in C++ and tested in a microscopic traffic simulator, AIMSUN2, to demonstrate

Table 1
Summary of simulation result on TH169 northbound test site

Unit: veh h	Total travel time	Total ramp delay
Heuristic	6240	401
Minnesota	6412	661
No Control	6929	0

that the control logic can be directly implemented in real-time. An introduction to this simulator is available in Barceló et al. (1994). A 20-km section of Trunk Highway 169 (TH169) northbound from I-494 to I-94 in the Twin Cities, MN, is selected as the test site. Most of the test section consists of two lanes with ten weaving sections. It has 24 entrance ramps, of which one is unmetered. The metered ramps include 4 HOV bypasses and two freeway-to-freeway ramps from TH62 and I-394. The test site contains 25 exit ramps. The temporal (14:00–19:30 PM) and spatial boundaries of the simulation experiment are free of congestion and the traffic demand data were collected on March 21, 2000. The simulated freeway network of this site has been calibrated in the AIMSUN2 simulator in an earlier study (Hourdakis and Michalopoulos, 2003).

An existing ramp control strategy—the Minnesota Zonal Algorithm (coded for the AIMSUN2 by Hourdakis and Michalopoulos (2003)), along with the no-control scenario are also simulated on the same test site for comparison. The details of the Minnesota Zonal Algorithm (referred to as the Minnesota algorithm hereafter) are discussed in Minnesota Department of Transportation (1998). The simulated total travel times and total ramp delays under the three control scenarios are derived from simulation outputs of five replications with the same randomly-generated random seeds, and summarized in Table 1. The total travel time is 10% less under the heuristic control compared to the no-control scenario. The Minnesota algorithm only shortens the total travel time by 7%. It takes only two minutes of CPU time on a Pentium 1.7 GHz PC with 256 mb memory to run the heuristic on the 20-km test freeway section for the whole peak period. This simulation is essentially a controlled experiment because all critical values and control methods of the strategy based on the most-efficient metering logic are set to be the same as the Minnesota algorithm. Under the proposed analytical framework, it is able to explain the two reasons why the Minnesota algorithm gives inferior measures of effectiveness: (a) it controls more on-ramps to relieve critical sections—some equity consideration; (b) it only concerns traffic conditions at several fixed historical bottlenecks while the most-efficient logic considers every freeway section as a potential bottleneck.

6. Equity consideration

The trade-off between efficiency and equity in freeway ramp metering has been pointed out in several previous studies (Kotsialos and Papageorgiou, 2001; Levinson et al., 2002). The heuristic developed in section 4 also suggests that the most efficient ramp control strategy is the least equitable one. Coordinating on-ramp meters is often a necessary step to eliminate freeway mainline queues. However, it can also be viewed as an equity consideration. A theoretical way to consider equity in ramp control has not been previously studied, but some practical equity

considerations have evolved implicitly over time in real-world ramp control strategies. The first control constraint that improves ramp control equity is probably the maximum queue length restriction, although the original motivation of this constraint is to prevent ramp queue spillover to local streets. In many practical ramp control strategies, there is a minimum/maximum metering rate constraint which is also beneficial from an equity point of view. More specifically, the Denver strategy has a so-called "helper algorithm" among on-ramps in which if one ramp is operated at its most restrictive rates, the ramp that immediately upstream of it will be operated more restrictively in the next control interval to release the downstream one to some extent. The Minnesota Zonal algorithm controls all on-ramps in a control zone to assure the flow at the zone bottleneck is below capacity. The new Minnesota Stratified Zonal algorithm has a maximum ramp delay restriction which ensures that the maximum ramp delay will not exceed four minutes for each individual driver. A more systematic way to consider equity in ramp metering probably requires a change in the objective function itself. Stated preference surveys and laboratory driving simulation experiments disclose that drivers value ramp delays and free-flow travel time differently. Freeway operators may need to consider minimizing total perceived travel time instead of absolute travel time. A more detailed account on that topic is provided in Zhang and Levinson (2003b).

7. Conclusions

The analytical framework developed in this research should assist both researchers and practitioners. It is shown analytically that the most efficient ramp control logic is the one that meters the on-ramps closest to any critical freeway sections such that there is no internal queue on freeway mainline. With different types of threshold values, control methods, and equity considerations, the metering logic can evolve into various practical ramp control strategies. The developed framework also provides a platform on which elements in a ramp control strategy can be decomposed and compared separately. In the long run, these comparisons would allow us to answer why one strategy outperforms another.

The stability of off-ramp exit percentages is essential in deriving the most efficient ramp control logic, which is supported by some empirical data on one freeway in the Twin Cities. The predictability of the exit percentage in the next control horizon using known information accumulated up to the decision point is a very desirable property of freeway traffic. With this predictability, the global optimal ramp control problem can be formulated as a linear program. All input variables for the optimization program become directly measurable by loop detectors. Future studies may examine the stability of off-ramp exit percentages on other freeways. Data required for such studies should be widely available. Formal analysis tools for time-series data may also be used to study properties of off-ramp exit percentages. The scope of this study is limited to the optimal control of a single freeway, and how the stability of off-ramp exit percentage can help formulate and solve integrated corridor (a freeway and parallel arterial streets) control problems has not been explored.

It is interesting, though not very surprising, that the findings suggest the most efficient ramp control logic is also the least equitable one. To achieve efficiency goals, we must meter the least number of on-ramps in order to provide free-flow conditions for all commuters on freeway

mainline, a majority of whom access the freeway through other on-ramps with less restricted metering rates. With efficiency as the sole criterion, we may have done an engineering job very well. However, such a strategy is not politically palatable and may lack public acceptance. Minimized travel time for the system as a whole is a good thing. However, if that is achieved by helping some drivers at the expense of others, there is also a serious equity issue that should be considered. Future studies should pursue a mechanism balancing the efficiency and equity of ramp meters.

Acknowledgements

This research was part of the project *Measuring the Equity and Efficiency of Ramp Meters* funded by the Minnesota Department of Transportation. The authors would like to thank the Center for Transportation Studies at the University of Minnesota and International Road Federation for providing additional support. The authors want to thank James Aswegan, John Bieniek, John Hourdakis, Rich Lau, and Frank Lilja for their assistance. The opinions and errors remain those of the authors.

References

Barceló, J., Ferrer, J.L., Grau, R., 1994. AIMSUN2 and the GETRAM simulation environment. Internal report. Departamento de Estadística e Investigación Operativa, Facultad de Informática, Universitat Politècnica de Catalunya.

Bogenberger, K., May, A.D., 1999. Advanced coordinated traffic responsive ramp metering strategies. California PATH Paper UCB-ITS-PWP-99-19.

Chang, T.H., Li, Z.Y., 2002. Optimization of mainline traffic via an adaptive co-ordinated ramp metering control model with dynamic OD estimation. Transportation Research 10C, 99–120.

Erera, A., Daganzo, C.F., Lovell, D., 2002. The access control problem on capacitated FIFO networks with unique O–D paths is hard. Operations Research 50 (4), 736–743.

Hourdakis, J., Michalopoulos, P.G., 2003. Evaluation of ramp control effectiveness in two twin cities freeways. Transportation Research Record 1811, 21–30.

Kotsialos, A., Papageorgiou, M., 2001. Efficiency versus fairness in network-wide ramp metering. Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference, Oakland, USA.

Kotsialos, A., Papageorgiou, M., Mangeas, M., Haj-salem, H., 2002. Coordinated and integrated control of motorway networks via non-linear optimal control. Transportation Research 10C, 65–84.

Kwon, E., 2000. Comparative analysis of operational algorithms for coordinated ramp metering. Center for Transportation Studies, University of Minnesota.

Levinson, D., Zhang, L., Das, S., Sheikh, A., 2002. Evaluating ramp meters: evidence from the Twin Cities ramp meter shut-off. Presented at the 81st TRB Annual meeting, Wahsington, DC. Under Review Transportation Research Part C.

Lomax, T.J., Schrank, D.L., 2000. A comparison of ramp metering in Minneapolis–Saint Paul and peer US cities. Minnesota Department of Transportation.

Lovell, D.J., 1997 Traffic control on metered networks without route choice. Ph.D. Dissertation, University of California at Berkeley.

Lovell, D., Daganzo, C.F., 2000. Access control on networks with unique origin–destination paths. Transportation Research 34B, 185–202.

Minnesota Department of Transportation Traffic Management Center, 1998. Ramp metering by zone—the Minnesota algorithm report. Minneapolis, MN.

Newman, L., Dunnet, A., Meis, G.J., 1969. Freeway ramp control—what it can and can not do. Traffic Engineering 39, 14–21.

Papageorgiou, M., 1983. Application of Automatic Control Concepts to Traffic Flow Modeling and Control. Springer-Verlag, New York.

Papageorgiou, M., Hadj-Salem, H., Blosseville, J., 1991. ALINEA: A local feedback control law for on-ramp metering. Transportation Research Record 1320.

Papageorgiou, M., 1995. An integrated control approach for traffic corridors. Transportation Research 3C, 19–30.

Persaud, B., Yagar, S., Brownlee, R., 1998. Exploration of the breakdown phenomenon in freeway traffic. Transportation Research Record 1567.

Taylor, C., Meldrum, D., Jacobson, L., 1998. Fuzzy ramp metering—design overview and simulation results. Transportation Research Record 1634.

Wattleworth, J.A., 1963. Peak-period control of a freeway system—some theoretical considerations. Ph.D. Dissertation, Northwestern University.

Wattleworth, J.A., 1967. Peak-period analysis and control of a freeway system. Highway Research Record 157.

Zhang, H.M., 1997. Freeway ramp metering using artificial neural network. Transportation Research 5C, 273–286.

Zhang, L., Beegala, A., 2003. Prediction of freeway breakdown using Artificial Neural Network. Accepted for presentation at the 14th ITS America Annual Meeting, St. Antonio, TX.

Zhang, L., Levinson, D.M., 2003a. Ramp metering and the capacity of active freeway bottlenecks, Presented at the 83rd TRB Annual Meeting, Washington DC. Under Review Transportation Research Part A.

Zhang, L., Levinson, D.M., 2003b. Balancing efficiency and equity of ramp meters. Presented at the 82nd TRB Annual Meeting, Washington DC. Under Review ASCE Journal of Transportation Engineering.

Zhang, H.M., Recker, W.W., 1999. On optimal freeway ramp control policies for congested traffic corridors. Transportation Research 33B, 417–436.