Effect of Response Limitations on Traffic-Responsive Ramp Metering

JAMES H. BANKS

Simulations of ramp meter responses were used to study the feasibility of replacing locally based moving-average estimates of mainline flow currently used in San Diego with estimates based on upstream data. The choice of estimation methodology made little practical difference in the performance of the system; instead, the major problem in providing precise control of meter outputs was the limited ability of the system to respond. The most important limitation is that the difference between the maximum and minimum metering rates tends to be small relative to normal random variation in mainline input flows at the minimum counting interval of 30 sec. Consequently, meters respond with their maximum or minimum rates most of the time, which leads to biases in average responses. Other response limitations include a comparatively large number of ramps at which demands are less than minimum metering rates and insufficient total metering capacity. For multiramp systems, the most promising way to prevent biases caused by the meters' limited response ranges is to set flow targets for upstream meters to cause the average response of the bottleneck meter to be about halfway between its maximum and minimum rates. This strategy may be employed only where there is sufficient total metering capacity and may conflict with other strategies for setting flow targets in multiramp systems, such as the so-called Wattleworth strategy.

Traffic-responsive ramp metering is an important technique of freeway traffic control. Its most obvious advantage is its potential to hold flows through bottlenecks close to capacity, despite fluctuations in traffic volumes arriving from upstream. In order to do this, it is necessary to calculate meter responses based on information about the state of the traffic stream.

Most past discussions of traffic-responsive metering strategies were focused on the best way to calculate the meter's response, given certain types of information about the system, or on determining the best indicators of the flow state. Little explicit attention is paid to the system's ability to respond and how this might affect the control strategies; at most, certain limitations may be included implicitly as constraints in mathematical models. Important response limitations include maximum and minimum limits to metering rates, the possibility that ramp demand may fall below the minimum metering rate, and constraints on ramp queue lengths.

The work reported here began as a feasibility study of a minor proposed modification to an existing traffic-responsive metering system. In the evaluation of that proposal, it became evident that the most important issue was the way in which the system's output was affected by its response limitations, particularly the maximum and minimum metering rates, and not the way in which the nominal response was calculated. A scheme for analyzing response errors in traffic-responsive metering systems is presented in this paper, and the scheme is applied to the simulated results of proposed modifications to the San Diego system. Some of the ways in which response limitations affect traffic-responsive strategies are discussed.

Civil Engineering Department, San Diego State University, San Diego, Calif. 92182-0189.

TRAFFIC-RESPONSIVE METERING STRATEGIES

Traffic responsive metering strategies may be divided into two types. Local strategies involve control of a single ramp to produce capacity flow through a bottleneck; in addition, considerable attention has been given to so-called "gap-acceptance" strategies (*1*). These strategies, which are not covered here, are intended to smooth flow without necessarily providing capacity operation. Global strategies manage a number of ramps to control flows throughout a more extended section of freeway. Because it is often impossible to hold flows approaching bottlenecks to capacity without metering several ramps, control of flow at the bottleneck is often the ultimate goal of multiramp systems; however, global strategies are usually expressed in terms of target flow states (usually volumes, traffic densities, or lane occupancies) distributed throughout the system.

Several global traffic-responsive metering strategies based on automatic control theory have been proposed. The most common variety is based on minimization of a quadratic performance functional that penalizes deviations from nominal values of flow and traffic density throughout the system and employs some variation of the overlapping decentralization scheme of Isaksen and Payne (*2–4*). Another proposed global strategy is the hierarchical system of Papageorgiou (*5,6*). This consists of three functional layers: (a) an optimization layer based on the steady-state linear programming formulation proposed originally by Wattleworth and Berry (*7*) and later extended by Wattleworth and others (*8–11*); (b) an adaptation layer that reacts to congestion and significant deviations from assumed origin-destination trajectories; and (c) a direct-control layer that implements local feedback controls. Papageorgiou's scheme allows the overall metering problem to be decomposed into separate global and local strategies, with the results of the global strategy entering the local strategy as output flow targets (or average metering rates) at the various meters. Such a decomposition simplifies the design of global traffic-responsive systems but also raises the issue of the compatibility of particular global and local strategies.

Local traffic-responsive strategies include demand-capacity strategies, traditional occupancy-based strategies, and feedback strategies (*12*). Demand-capacity and traditional occupancy-based strategies are similar. In both, metering rates are calculated as the difference between a target output volume and the traffic volume measured (or estimated) just upstream of the ramp. The only difference is that traditional occupancy-based schemes estimate upstream volumes from occupancies instead of measuring them directly; the advantage of this is that it does not require detectors for all lanes. In both cases, the normal response calculation is overridden whenever high occupancies indicate congestion; in this case, minimum metering rates are employed. Feedback strategies base meter responses on traffic conditions downstream of the ramp. Recent European work has included the development and testing of the occupancy-based feedback rule ALINEA (*12–14*). Real-life tests on European freeways have indicated that ALINEA is superior to demand-capacity and traditional occupancy-based strategies.

The superiority of feedback strategies appears to lie in their quicker response to the transition from uncongested to congested flow. Under uncongested conditions, all local strategies are similar and can be thought of as variations of the following rule:

$$M(t) = Q - \hat{q}(t) + \Delta(t) \qquad \text{for } M_{min} \le M(t) \le M_{max} \tag{1}$$

otherwise $M(t) = M_{max}$ or M_{min}, as appropriate, where

$M(t)$ = metering rate for time period t,
M_{max} = maximum metering rate,
M_{min} = minimum metering rate,
Q = output flow target,
$\hat{q}(t)$ = estimated mainline flow arriving from upstream during time period t,
$\Delta(t)$ = estimated difference between the actual ramp output and the metering rate; $\Delta(t) = M(t) - p(t)$, where $p(t)$ is the ramp passage count.

Alternatively, the relationship between the passage count and the metering rate may be expressed as the ratio $r(t) = M(t)/p(t)$, and

$$M(t) = \frac{Q - \hat{q}(t)}{r(t)} \tag{2}$$

as long as $M_{min} \le M(t) \le M_{max}$.

In either version, this rule states that the metering rate for time period t is a function of the difference between the target output flow and an estimate of the mainline input flow for the same time interval. The various strategies differ only in the way in which the current upstream flow is estimated. Demand-capacity strategies estimate it on the basis of past upstream volume counts; traditional occupancy-based strategies estimate it on the basis of past upstream occupancy measurements; and feedback strategies estimate it on the basis of past downstream volumes or occupancies. In addition, estimates may involve various types and degrees of data smoothing.

EVALUATION SCHEME

The ultimate goal of local traffic-responsive metering strategies is to reduce delay by maximizing flow through the bottleneck. All are based to some extent on the so-called two-capacity phenomenon, in which maximum uncongested flows have been found to exceed queue discharge rates. Consequently, they are designed to hold uncongested flow as close as possible to its maximum value, and to act quickly to restore uncongested flow whenever flow breakdown does occur. They can be successful only if they can hold uncongested flows above the queue discharge rate. Since recent research indicates that maximum uncongested flow rates exceed queue discharge rates by no more than 5 to 6 percent (*15–19*), metering is required to be quite precise.

The most direct means of evaluating such control strategies is to implement them to determine whether they can produce bottleneck flows that exceed the queue discharge rate over extended periods of time. Where this is not possible, they may be simulated and analyzed by their ability to produce a predetermined target output flow.

In carrying out this evaluation, it is important to consider the time frame. Most traffic-responsive systems use short count intervals, often 1 min or less. The meter's responses are updated with similar frequency, so that it is possible to evaluate the system's ability to produce specific 30-sec or 1-min output volumes. It is not clear, however, what deviations from the target flow are significant at this level of disaggregation. Short-term volume counts normally display a great deal of random variation, and this is true of both queue discharge and uncongested flow. The *Highway Capacity Manual* (*20*) uses 15-min flows to define capacity; other research related to capacities, such as the recent work related to the two-capacity phenomenon, has used intervals of at least 5 or 6 min. Meanwhile, average queue discharge rates, which are the critical point of comparison in evaluating the performance of the system, may often be measured over even longer time intervals.

It seems reasonable to assume that the value of Q used in Equation 1 or 2 will represent flow over a period of at least 5 or 6 min, although the meter's response is updated much more frequently. Consequently, the primary criterion for evaluating the accuracy of a particular strategy is the difference between the output and the target flow, averaged over a period of 5 min or more; a secondary criterion is the variance or standard deviation of the differences measured at the minimum response interval, since this reflects the accuracy of the response at the minimum interval.

Let $e(t)$ represent the response error for time period t; $e(t)$ is defined as the actual output flow minus the flow target, that is

$$e(t) \equiv q_o(t) - Q \tag{3}$$

where q_o represents the actual output flow. Three sources of response error are to be expected. These are (a) predictive error e_q, which results from the difference between q and $\hat{q}$, (b) response error e_r, which results from nominal metering rates falling outside the limits M_{max} and M_{min}; and (c) response error e_p, which results from the differences between $M + \Delta$ or rM and the actual ramp flow. If Equation 2 is used to determine the metering rate, these are

$$e_q(t) = q(t) - \hat{q}(t) \tag{4}$$

$$e_r(t) = \begin{cases} r(t)M_{min} - Q(t) + \hat{q}(t) & \text{if } [Q(t) - \hat{q}(t)]/r(t) < M_{min} \\ r(t)M_{max} - Q(t) + \hat{q}(t) & \text{if } [Q(t) - \hat{q}(t)]/r(t) > M_{max} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and

$$e_p = p(t) - r(t)M(t) \qquad (6)$$

The overall error in the output is the combination of these errors.

Now consider the ways in which these various sources of error might combine to produce the distribution of flow downstream of the ramp. In cases in which all predicted flows fall within the response limits, flow downstream of the meter is

$$q_o = Q + e_q + e_p$$

If the parameters of the predictive model are correctly estimated, both e_q and e_p should have zero mean. Also, there is every reason to believe that they are statistically independent, so their variances should sum. Thus, the distribution of q_o should have mean Q (the target flow) and variance $VAR(e_q) + VAR(e_p)$.

If some predicted flows fall outside the response limits, matters become more complicated. In this case

$$q_o = Q + e_q + e_p + e_r$$

but e_r does not have zero mean. Instead, it has negative mean when the calculated value of M exceeds M_{max} and positive mean when the calculated value of M is less than M_{min}. Also, e_r is not independent of e_q; instead, they should be negatively correlated. For instance, if $\hat{q}$ is too high, there will be a negative error e_q; meanwhile, there will be increased probability that the calculated value of M will be less than M_{min}, which results in a positive error e_r. Since e_r exists only when the meter's response is M_{max} or M_{min}, the distribution of e_r is a combination of the conditional distribution of q, given that $(Q - \hat{q})/r > M_{max}$ and the conditional distribution of q given that $(Q - \hat{q})/r < M_{min}$. As such, it depends on the dispersion and shape of the distribution of the actual upstream count q and on the relative probabilities that the calculated meter responses fall outside the limits M_{max} and M_{min}.

METER RESPONSE SIMULATION

The main objective of this study was to evaluate a minor proposed modification to the San Diego ramp metering system. The system in question is a multiramp traffic-responsive system, in which all field controllers communicate with a central computer. At the time of the study, the system consisted of approximately 100 controllers; most were grouped into six major subsystems of 7 to 22 controllers each. In addition, there were a number of isolated controllers or small groups. Most controllers are used to operate meters, although a few only collect data.

At present, coordination between the various metering locations consists of a set of flow targets that are intended to hold flows at the bottlenecks to their capacities. These are preset, having been determined over time by trial and error. In a sense, they correspond to the optimization layer in Papageorgiou's proposed system (*5,6*), although no actual optimization is involved. In addition, there is a limited adaptive capability: under certain circumstances, high occupancies will cause restrictive metering rates to be passed upstream.

Local traffic-responsive control is of the demand-capacity variety. Detectors, as is common in demand-capacity systems, are located just upstream of the on-ramps. At present, the upstream flow is calculated as a moving average of several past 30-sec counts, with the number of counts varying by location. Meter cycles (and hence metering rates) vary in discrete steps, with the number of steps depending on the location. Each ramp has maximum and minimum rates, and if the calculated metering rate falls outside these limits, it is set at the appropriate limit. In addition, when high occupancies indicate that a mainline queue is present, this calculation is overridden and the metering rate is set at its minimum value. In contrast to most systems, San Diego does not have maximum ramp queue length constraints; instead, minimum metering rates are set to control the growth of ramp queues.

One alternative to the existing strategy is to base the estimate of $\hat{q}$ in Equation 1 or 2 on data from upstream meters. Under uncongested conditions, average speeds in San Diego have been found to vary only slightly with flow, and there seems to be an emerging consensus that this is true of most North American data (*21,22*). If this is the case, it should be relatively easy to project variations in uncongested flow downstream. It was proposed to do this by using a moving average of flow at an upstream meter, offset from the response interval by the travel time between the two meters. The estimated flow was further modified by a factor intended to account for count biases and expected flows at intervening off-ramps (off-ramp flows are generally not counted in San Diego) and by adding actual or expected flows at intervening on-ramps. Details of the proposed evaluation scheme and the calibration of the parameters of the flow model may be found elsewhere (*23*).

The proposed control system modification was evaluated by simulating the performance of alternative estimates of $\hat{q}$ based on data from three freeway segments in San Diego. Figures 1 and 2 are schematic diagrams showing lane configurations, distances between detectors, and approximate maximum flow rates for selected locations in these sections. The parameters of the models were calibrated for these sections, and actual 30-sec counts from the peak periods of three different days were used to calculate values of $\hat{q}$ at the downstream end of each section. These were compared with actual 30-sec volume counts. The response of a meter at or near the downstream end of each section was then simulated on the basis values of Q set for these locations. In the case of Sections 1 and 3, no bottlenecks were present, and values of Q were set arbitrarily. In the case of Section 2, the Q-value was set to approximate the capacity of the College Avenue bottleneck; it should be noted, however, that what was simulated was the output of the College Avenue meter, although past studies indicate that the bottleneck is probably just upstream of the on-ramp (*15*). The output of the meter was estimated to be

$$\hat{q}_o(t) = q(t) + r(t)M(t)$$

where $q(t)$ represents the actual 30-sec count for time period t; $e(t)$ was then estimated to be $\hat{q}(t) - Q$. Metering rates between the maximum and minimum limits were assumed to vary continuously, thus ignoring the discrete cycle lengths used in practice.

Since the simulated metering rates were not necessarily the same as the actual rates in the data, it was not possible to

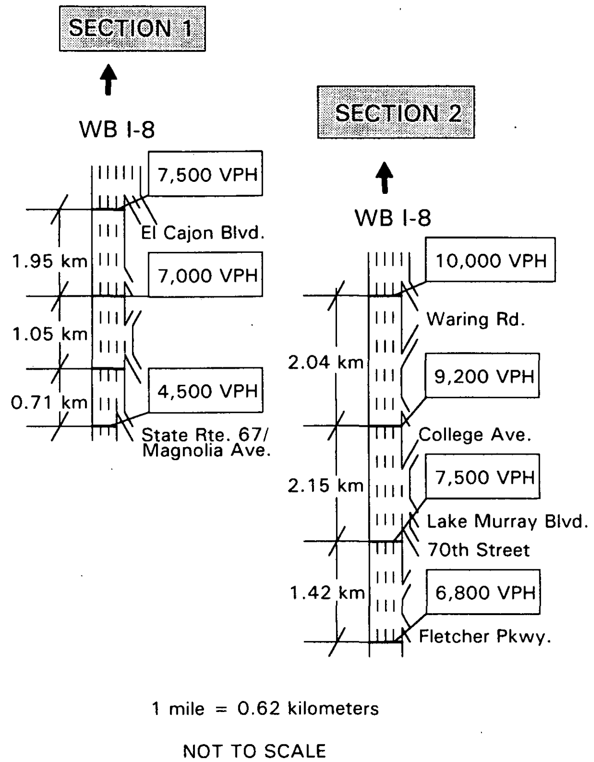

FIGURE 1 Schematic diagrams of Interstate 8 study sections.

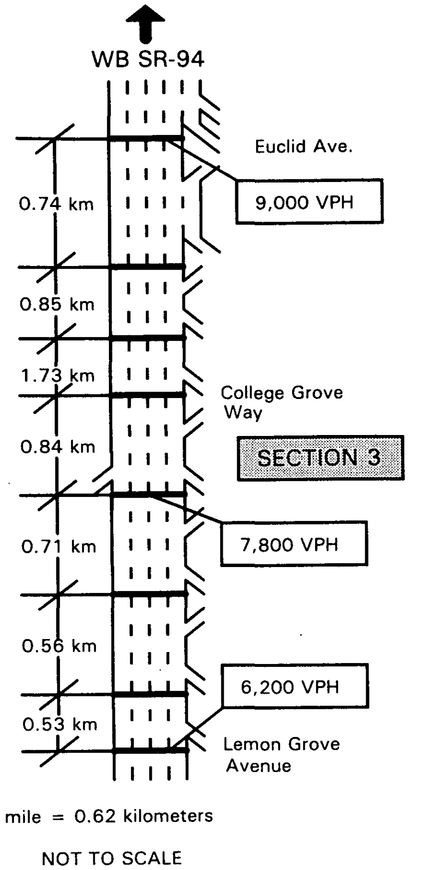

FIGURE 2 Schematic diagram of State Route 94 study section.

estimate e_p. Also, given the complexities of the error distribution in the case in which some values of $\hat{q}$ fall outside the response limits, no attempt was made to analyze the distribution of e_r in detail. Instead, alternative ways of estimating $\hat{q}$ were compared on the basis of the distributions of e_q and of the combined errors $e_q + e_r$. Means and standard deviations were calculated for e_q and the sum $(e_q + e_r)$ for 6-min intervals, and the mean of e_r was calculated by subtracting the mean of e_q from the sum. Also, these quantities were calculated for 30-min intervals during which mainline flows produced by the existing system are nearly constant and at their maximum values.

Estimation schemes based on upstream data were compared with those based on local data. Also, within each category, different averaging intervals were used to determine the effect of data smoothing. Finally, upstream-based schemes using data from different locations were compared.

RESULTS

All models were reasonably good at estimating $\hat{q}$ over periods of 5 to 6 min. Mean values of e_q, when averaged over 6 min, ranged from near 0 to about 1.7 percent of the target flow, and in most cases were less than 1 percent. Given the corresponding variances of e_q, these means do not appear to be significantly different from zero.

Ability to predict individual 30-sec counts, on the other hand, was low in all cases. This is indicated by relatively high standard deviations of the distributions of e_q. When calculated over 30-min intervals, the standard deviation of e_q ranged from 7 percent up to about 20 percent of the mean flow, with most cases in the range of 10 to 15 percent.

Mean values of e_r were highly variable and quite sensitive to the relationship between the average input flow and the target flow, the response range of the meter, and daily variations in the time series of the input flow. On the whole, they tended to be similar for both upstream-based and locally based models. For both types of model, their absolute values tended to increase as averaging intervals decreased.

Absolute values of the mean of e_r, calculated over 30 min, ranged from 0 to around 7 percent of the target flow. They tended to be least when the mean of the input flow was "centered" in the response range of the meter—that is, when the mean input flow was approximately equal to $Q - r(M_{max} + M_{min})/2$—and to increase as the mean input flow approached $Q - rM_{max}$ or $Q - rM_{min}$. The response error e_r tended to dominate the total error in cases in which the mean input flow was not centered in the response range and led to average total errors of roughly 2 to 5 percent of the mean flow over periods as long as 30 min.

EFFECTS OF RESPONSE LIMITS

To understand why the output is so sensitive to response errors, it is necessary to realize that the normal random variation in 30-sec counts almost always exceeds the difference between the maximum and minimum metering rates. Figure 3 is a typical example of a time series of 30-sec counts from the San Diego system. The section in question has four lanes in one direction, and a maximum average flow of around 57 vehicles per 30 sec or 6,800 vehicles per hour. (The comparatively low flow is because the location in question is some distance upstream of the bottleneck.) Note that the individual counts fall in a band that is about 20 to 30 vehicles wide. This corresponds to a variation in flow rate of around 2,400 to 3,600 vehicles per hour. Meanwhile, differences between maximum and minimum metering rates for meters in the sections studied ranged from 1.2 to 8.5 vehicles per 30 sec or 144 to 1,020 vehicles per hour, with a range of around 5 vehicles per 30 sec or 600 vehicles per hour being most common. As a result, meters are responding with maximum or minimum rates most of the time; in the case of the simulations described here, 45 to 96 percent of the responses were either M_{max} or M_{min}, depending on the flow model, the mean input flow, and the response range of the meter.

When the input flow is centered in the meter's response range, M_{max} and M_{min} occur about equally, and the response errors tend to cancel out. When mean input flow approaches $Q - rM_{min}$, M_{min} predominates and the output is biased high; the opposite happens when q approaches $Q - rM_{max}$. Consequently, the main difficulty in providing precise control of bottleneck flows is the limited response of the meter in the face of highly variable mainline flows and not the accuracy of the estimate $\hat{q}$ on a count-by-count basis.

In fact, overly accurate prediction of 30-sec counts is counterproductive, as is evidenced by the fact that the short-averaging-interval upstream-based models, which produced the best estimates of individual 30-sec counts, resulted in the largest response errors. Consider the case of a perfectly accurate estimator of 30-sec flow attempting to respond to a mean input flow of $Q - rM_{min}$. When the 30-sec count exceeds the average value, the meter responds with M_{min}, and the output exceeds Q. When the 30-sec count is less than average, it responds (correctly) with something more than M_{min}, and the output is Q. Consequently, these will be a substantial positive bias in the output. An estimator that was so heavily smoothed that it always estimated the next 30-sec count to be equal to the long-term average of q, on the other hand, would result in roughly half the output counts being less than Q and half being greater, so that the average output would be approximately Q.

In addition to the phenomenon just described, other limitations to the system's ability to respond were observed. For instance, it is normally assumed that actual flow from a metered ramp will be approximately equal to the metering rate or that the two may be related by a fairly stable factor, such as r in Equation 2. In calculating r values for the San Diego system, it was discovered that they not only varied by time of day, but that, at some locations, they also were often significantly less than 1.0 throughout the peak. Of the 16 ramps included in the sections studied here, 6 had r values of less than 0.95 throughout the peak or exceeded 0.95 for no more than 5 or 6 min; several other meters in Section 3 exceeded 0.95 for a total of 15 to 30 min during a 4-hr potential metering period.

This appears to indicate that ramp demand is less than the minimum metering rate at many locations. This was particularly true where ramp spacings were fairly close (as would be expected) and where there were high-occupancy vehicle (HOV) lane bypasses around ramp queues. The San Diego system, in contrast to some other systems, does meter HOV lanes, but it appears that demand for HOV lanes is rarely equal to the minimum metering rate.

In Figures 4 through 6, ramp passage counts are compared with metering rates for a variety of situations. Figure 4 shows the beginning of metering on a fairly high-volume ramp. It

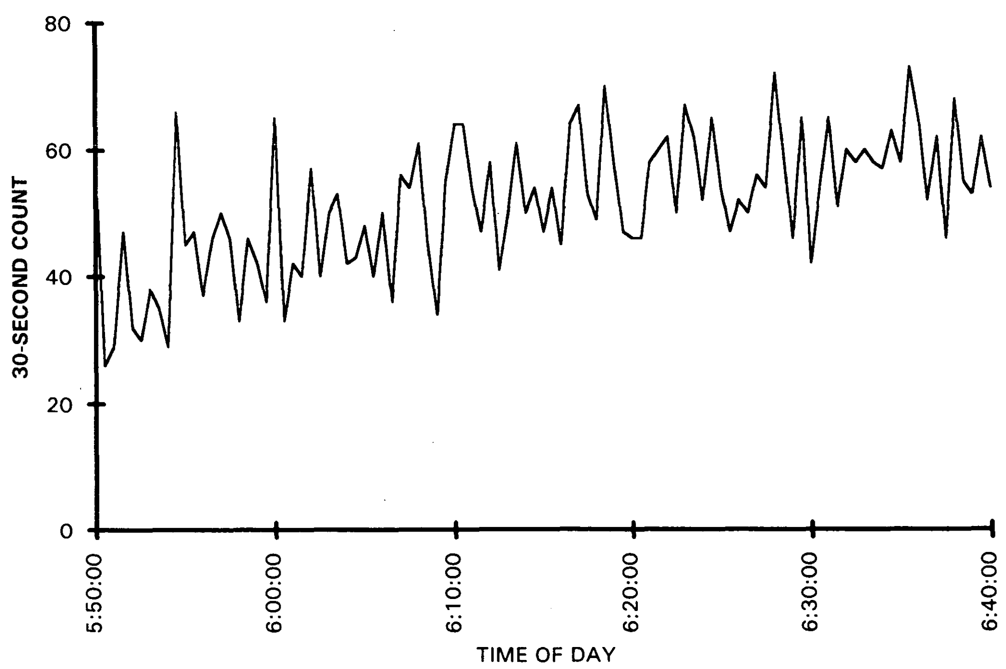

FIGURE 3 30-sec count versus time, westbound Interstate 8 at Fletcher Parkway, June 11, 1991.

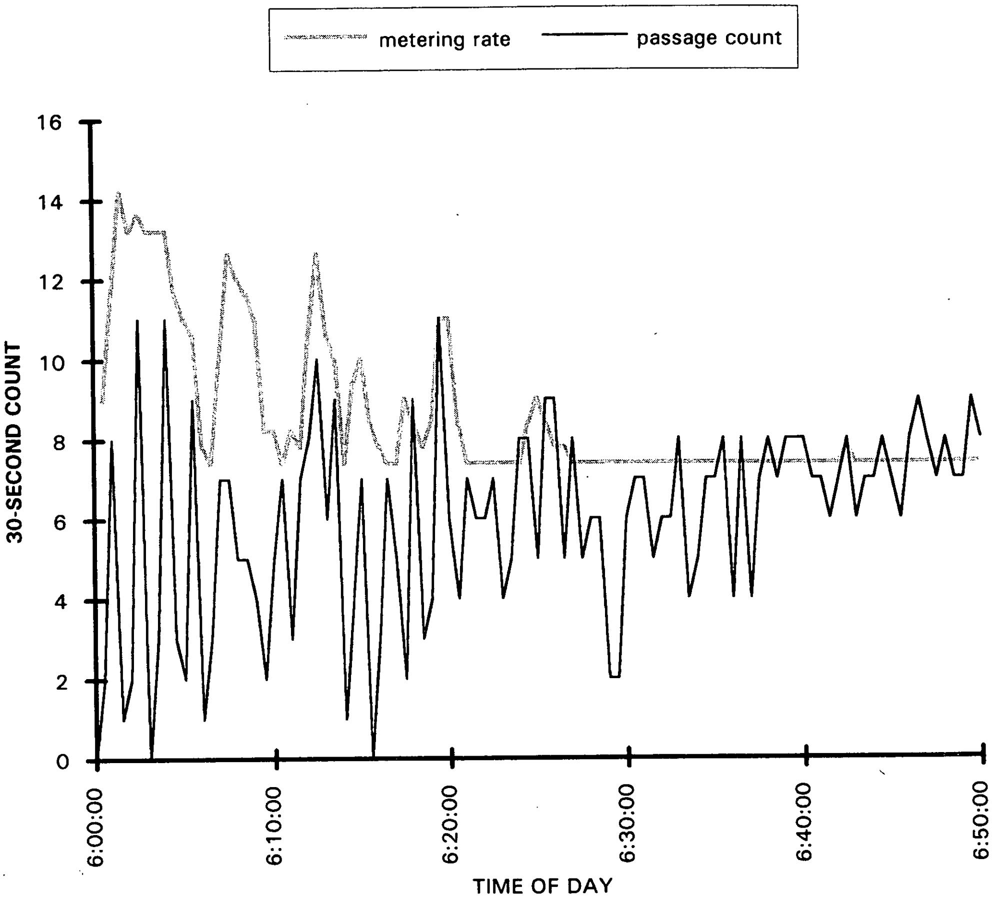

FIGURE 4 Theoretical metering rate versus passage count, Fletcher Parkway on-ramp to Interstate 8, August 5, 1991.

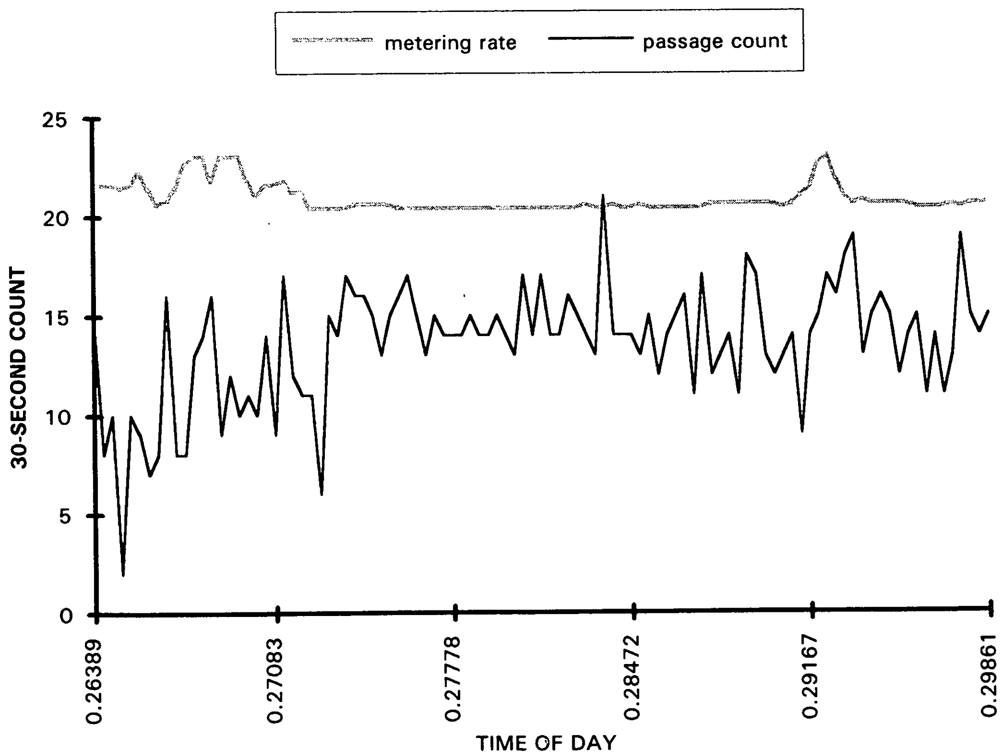

FIGURE 5 Theoretical metering rate versus passage count, 70th Street/Lake Murray Boulevard on-ramp to Interstate 8, August 5, 1991.

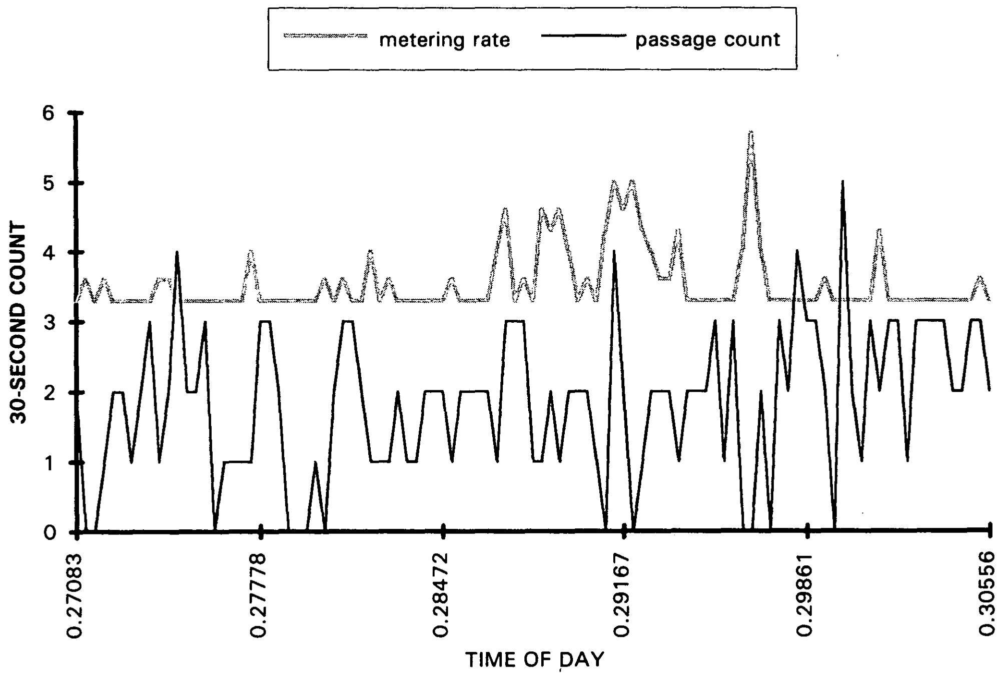

FIGURE 6 Theoretical metering rate versus passage count, College Grove Way on-ramp to State Route 94, August 5, 1991.

takes about 20 min for the metering rate to settle down to its minimum value and an additional 15 min or so for the passage count to settle down to the metering rate; however, it eventually does so. Figure 5 represents a high-volume ramp with an HOV lane, and Figure 6 represents a low-volume ramp with an HOV lane. In both of these cases, the average passage count is substantially less than the metering rate and there is substantial variation on a count-by-count basis.

Finally, a major response limitation in San Diego appears to be a lack of adequate total metering capacity, despite an extensive system. The portions of the system studied involved inbound flow during the morning peak and metering, which extends to the outer edge of the metropolitan area. It was observed that all meters throughout this portion of the system were responding with their minimum rates during the most intense portions of the peak, but that this was not always sufficient to prevent flow breakdown. Obviously, if all meters must be at their minimum rates for extended periods of time to prevent flow breakdown at the bottlenecks, there can be no real traffic-responsiveness during those periods.

IMPLICATIONS FOR METERING STRATEGIES

The preceding section identified three ways in which the San Diego ramp metering system is limited in its ability to respond to variations in traffic flow. These limitations, which may well apply to other traffic-responsive multiramp systems, include insufficient total metering capacity, a significant number of ramps at which demand is less than the minimum metering rate, and meter response ranges that are considerably less than the normal variation in mainline volumes at the minimum count interval. The following discussion is focused on the last of these, since not much can be done about the other two unless minimum metering rates can be reduced.

A traffic-responsive metering strategy may be adjusted in three ways to reduce the output biases that result from frequent use of maximum or minimum metering rates. They are as follows: (a) set flow targets at the upstream meters so that the input flow at the bottleneck will be centered in the meter's response range; (b) smooth the flow estimate $\hat{q}$ to reduce the response to short-term input flows that do fall within the response range; or (c) use an offset value for the flow target (i.e., use some value other than the actual output flow target for Q in Equation 1 or 2).

Of these three modifications, only the first shows much promise of success. The simulations showed that where input flows were not well-centered in the response range, smoothing of flow estimates based on local data by increasing the averaging interval from 30 sec to 6.5 min will normally reduce the response error by about 30 to 50 percent. The remaining errors, however, are still on the order of 2 to 5 percent of the target flow, and there will almost always be an increase in the variance of the output.

The possibility of using offset flow targets was not investigated in detail. Offset targets could be determined by simulations in which the target in the response calculation was varied until the desired mean output was achieved. This would be unlikely to produce consistently good results, however, because the simulations also showed considerable variation in the response error from day to day when input flows were not centered in the response range.

Setting upstream flow targets centered in the meter's response range to provide input flows at the bottleneck could be somewhat more effective. Based on the simulations performed here, this method shows promise of being able to hold mean output flows (averaged over 30 min) to within 1.5 percent of their targets. In order to do this, it would probably be necessary to set all upstream targets based on the assumption that the mean metering rate at every ramp downstream was approximately halfway between M_{max} and M_{min}. By doing this, flows could be held close to their targets throughout the system.

Unfortunately, this could create a another problem if there were an attempt to implement the optimization layer of Papageorgiou's hierarchical control system (*5,6*). In this scheme, flow targets are set based on the steady-state linear programming formulation of Wattleworth (*7,8*), which is unlikely to result in metering rates halfway between M_{max} and M_{min}. Obviously, there may be a trade-off between the objective of precise control of flow at the bottleneck and that of minimization of delay for the system as a whole, which is the basis of the Wattleworth strategy.

CONCLUSION

Described in this paper is a study in which simulations of ramp meter responses were used to explore the feasibility of replacing the locally based moving-average estimates of mainline flow used by the San Diego ramp metering system with estimates based on upstream data. These simulations showed that choice of estimation methodology made little practical difference in the performance of the system and that the major problem in using ramp metering to provide precise control of bottleneck flows is the limited ability of the system to respond and not the accuracy of the flow estimate. The most important limitation appears to be the limited range of response of individual meters, which is typically much less than the normal variation in flow at the minimum count interval of 30 sec used in San Diego. This leads to a situation in which meters respond with their maximum or minimum rates much of the time and in which mean output flows will be biased unless mean input flows are centered in the response range. Other response limitations include a comparatively large number of ramps at which demands are less than minimum metering rates and insufficient total metering capacity.

The most promising way of preventing biases in the mean output of meters is to set flow targets for upstream meters to provide mean mainline input flow at the bottleneck approximately equal to the output target minus $(M_{max} + M_{min})/2$. This strategy will normally require that mean metering rates at all meters be approximately halfway between M_{max} and M_{min}. It would be ineffective for systems such as that in San Diego, however, because at present all meters are required to operate at their minimum rates for much of the peak in order to hold flows at bottlenecks close to their capacities. It is also likely that this strategy conflicts with any strategy that sets metering rates based on some other criterion, such as minimization of delay.

These findings raise several issues requiring further research. One of these is how best to resolve conflicting objectives in hierarchical ramp metering control systems. A second issue is that of the relationships between measurable (and potentially controllable) characteristics of the traffic stream and flow breakdown. For instance, how important to the process of flow breakdown is the variation in volume counts over intervals such as 30 sec as opposed to the mean flow over periods of 5 or 6 min? The goal of this research should be clarification of the goals of the control system in terms of the type of flow it should be attempting to produce. A third issue is how metering strategies other than those simulated here are affected by response limitations. For instance, is the reported superiority of the occupancy-based feedback law ALINEA due to reduced vulnerability to the effects of response limits (which seems unlikely) or to some other cause, such as faster response to the onset of congestion?

ACKNOWLEDGMENTS

This research was funded by the State of California Department of Transportation, District 11. The contents of this paper reflect the views of the author, who is responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation. Special thanks are due to research assistant Sadam Panduranga.

REFERENCES

1. D. R. Drew. Gap Acceptance Characteristics for Freeway-Ramp—Surveillance and Control. In *Highway Research Record 157,* HRB, National Research Council, Washington, D.C., 1967, pp. 108–135.
2. L. Isaksen and H. J. Payne. Suboptimal Control of Linear Systems by Augmentation with Application to Freeway Traffic Regulation. *IEEE Transactions on Automatic Control,* Vol. AC-18, 1973, pp. 210–219.
3. N. B. Goldstein and K. S. P. Kumar. A Decentralized Control Strategy for Freeway Regulation. *Transportation Research,* Vol. 16B, 1982, pp. 279–290.
4. D. P. Looze, P. K. Houpt, N. R. Sandell, and M. Athans. On Decentralized Estimation and Control with Applications to Freeway Ramp Metering. *IEEE Transactions on Automatic Control,* Vol. AC-23, 1978, pp. 268–275.
5. M. Papageorgiou. A Hierarchical Control System for Freeway Traffic. *Transportation Research,* Vol. 17B, 1983, pp. 251–261.
6. M. Papageorgiou. Multilayer Control System Design Applied to Freeway Traffic. *IEEE Transactions on Automatic Control,* Vol. AC-29, 1984, pp. 482–490.
7. J. A. Wattleworth and D. S. Berry. Peak Period Control of a Freeway System—Some Theoretical Investigations. In *Highway Research Record 89,* HRB, National Research Council, Washington, D.C., 1965, pp. 1–25.
8. J. A. Wattleworth. Peak-Period Analysis and Control of a Freeway System. In *Highway Research Record 157,* HRB, National Research Council, Washington, D.C., 1967, pp. 10–21.
9. A. D. May. A Proposed Dynamic Freeway Control System Hierarchy. *Proc., 3rd IFAC/IFIP/IFORS International Symposium on Control in Transportation Systems,* Columbus, Ohio, 1976, pp. 1–12.
10. M. Papageorgiou. A New Approach to Time-of-Day Control Based on a Dynamic Freeway Control Model. *Transportation Research,* Vol. 14B, 1980, pp. 349–360.
11. R. L. Pretty. Control of a Freeway System by Means of Ramp Metering and Information Signs. In *Highway Research Record 388,* HRB, National Research Council, Washington, D.C., 1972, pp. 62–69.
12. H. Hadj-Salem, J.-M. Blosseville, and M. Papageorgiou. ALINEA: a Local Feedback Control Law for On-Ramp Metering—A Real Life Study. *Proc., 3rd IEE International Conference on Road Traffic Control,* London, England, 1990, pp. 194–198.

13. S. Smulders and F. Middelham. *Isolated Ramp-Metering: Real Life Study in the Netherlands.* DRIVE-Office, Brussels, Belgium, 1991.
14. M. Papageorgiou, H. Hadj-Salem, J.-M. Blosseville. ALINEA: A Local Feedback Control Law for On-Ramp Metering. In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 58–64.
15. J. H. Banks. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287,* TRB, National Research Council, Washington, D.C., 1990, pp. 20–28.
16. J. H. Banks. The "Two-Capacity" Phenomenon at Freeway Bottlenecks—A Basis for Ramp Metering? In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 83–90.
17. J. H. Banks. The "Two-Capacity" Phenomenon: Some Theoretical Issues. In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 234–241.
18. F. L. Hall and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 91–98.
19. T. Urbanik II, W. Hinshaw, and K. Barnes. Evaluation of High-Volume Urban Texas Freeways. In *Transportation Research Record 1320,* TRB, National Research Council, Washington, D.C., 1991, pp. 110–118.
20. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
21. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225,* TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.
22. F. L. Hall, V. R. Hurdle, and J. H. Banks. A Synthesis of Recent Work on the Nature of Speed-Flow and Speed-Occupancy (or Density) Relationships on Freeways. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
23. J. H. Banks. *Predictive Flow Model for Ramp Metering.* Final Report, Civil Engineering Report Series 9210. San Diego State University, San Diego, Calif., 1992.

Publication of this paper sponsored by Committee on Freeway Operations.