An Analytic Stochastic Model for the Transit Vehicle Holding Problem

Mark D. Hickman
Department of Civil Engineering and Engineering Mechanics,
The University of Arizona, Tucson, Arizona 85721-0072

This paper describes an analytic model that determines the optimal vehicle holding time at a control stop along a transit route. This model is based on a stochastic transit service model presented by Andersson and Scalia-Tomba (1981) and enhanced by Marguier (1985). The use of a stochastic service model allows greater realism in the analytic modeling. Making use of these results, the paper presents an analytic model that may be used to determine the optimal holding time for a vehicle at a control stop. As it is formulated, the single vehicle holding problem is a convex quadratic program in a single variable, and is easily solved using gradient or line search techniques. The analytic holding model overcomes two noted problems in the literature: it includes stochastic service attributes of vehicle running times and passenger boarding and alighting processes, and the model may be used for real-time control purposes. The use and potential benefits of the model are illustrated in a simple example. This model may be useful in developing a computerized decision support system to enhance the effectiveness of transit operational decision-making.

One of the continual challenges faced by transit managers is how to provide better service in a dynamic operating environment. In many areas, services are constantly subject to delays and disruptions due to traffic congestion, weather, vehicle breakdowns, and other events. These delays and disruptions are a considerable annoyance to passengers, many of whom are dependent on transit for their basic mobility. It is well known that transit passengers consistently rank on-time performance and schedule reliability as one of the most critical factors affecting their use of transit. While good quantitative evidence is lacking, the inevitable service delays and disruptions do adversely affect transit ridership and its overall level of service to travelers (see, for example, Abkowitz et al. 1978 and Abkowitz 1980). Moreover, from the point of view of the transit operator, service delays and disruptions have a real monetary cost, in terms of lower utilization of vehicles and operators. Conservative estimates are that service delays account for 3–5% of both operating and capital (vehicle) costs. Some agencies have demonstrated simple techniques to improve vehicle routing, scheduling, and service monitoring that have shown associated cost reductions of over 10% (Goeddel 1996). It appears that such improvements can result in significant cost savings to the transit agency and in improved service to transit passengers.

Transit routes and schedules are often subject to unanticipated delays, disruptions, and uncertainties in the anticipated demand. In response, drivers, field supervisors, and dispatchers can identify control actions to respond to these dynamic changes. One of the most commonly employed actions is holding a vehicle at a stop (Levinson 1991). A vehicle may be held for some additional time at a stop in order to improve the "regularity" or reduce the variability of headways downstream of this control stop. By adjusting vehicle headways in this manner, the level of service as measured by passenger waiting times may be improved.

0041-1655/01/3503/0215$05.00
1526-5447 electronic ISSN

The particular holding strategy (or strategies) employed depends on a route's operating conditions, including the volume of passengers along the route, the frequency of service, the variability of route running times and passenger loads, and other factors. Nonetheless, a hold is typically applied using the driver's or supervisor's best judgment. Other common holding techniques include threshold-based and schedule-based holding. The threshold technique involves holding a vehicle only if the preceding headway is below a certain amount of time (e.g., the desired headway); in this case, the vehicle is held only until the threshold time and then dispatched. On the other hand, schedule-based holding involves holding a vehicle only until its scheduled departure time and dispatching it immediately if it arrives later than this scheduled time. Typically, the objective of threshold- or schedule-based holding strategies is to minimize either the downstream passenger waiting time or some combination of downstream waiting time and the delay to passengers already on board.

The primary contribution of this paper is to present the optimal holding problem *in the context of a detailed stochastic vehicle operations model*. Relative to existing analytic methods, what is needed is an analytic model for transit vehicle holding that accounts for: (1) observed route-level operational dynamics, and (2) stochastic travel times and boarding and alighting times for transit vehicles. Also, relative to existing simulation models, an analytic model allows a more general modeling construct for measuring the effects of holding policies. Finally, in contrast to threshold holding policies, the analytic model may also take advantage of real-time knowledge of current route conditions (e.g., from an automated vehicle monitoring system). This paper provides a first step in this direction by presenting an analytic holding model that explicitly accounts for route-level operational dynamics, for stochastic travel times and passenger boarding and alighting times, and for real-time knowledge of vehicle status.

To develop such a holding model, this paper is based on a "microscopic" model of stochastic and dynamic transit operations developed by Andersson and Scalia-Tomba (1981) and refined by Marguier (1985). This route-level operations model allows certain simplifications in the modeling of the holding problem, while maintaining microscopic modeling features. It is important to note that this model can accommodate holding actions under both minor and major disruptions and deviations from service. However, the microscopic operations model assumes that when a hold is applied, the subsequent evolution of the system is consistent with more typical operating conditions (i.e., only minor disruptions and deviations). In this context, the proposed method assumes that system recovery behaves according to historical operating characteristics in both expectation and variance.

A review of the literature in the area of transit vehicle holding is presented in §1. Because Marguier's work has not previously been formally published, §2 presents several important results from his stochastic model of a transit route. Using this framework, §3 presents an analytic and stochastic formulation of the transit vehicle holding problem. The formulation and a corresponding solution algorithm are described. The holding model is illustrated using a simple route example in §4. Finally, §5 offers some conclusions on this model and suggests several areas for future research.

1. Literature Review

Research into vehicle holding strategies originated with the analytic model of Osuna and Newell (1972), with further refinements and extensions by Newell (1974), Barnett (1974), Barnett and Kleitman (1978), and Barnett (1978). These models assume fairly simple transit networks, typically some kind of shuttle service or a simple service loop, with a limited number of vehicles (one or two). The models also include some probability distribution of vehicle running times for the service. Using this framework, these models derive an optimal value of a threshold, corresponding to a threshold-based holding policy. Because of the limitations on the route structures, these models are generally not directly applicable to a more typical fixed-route transit service.

The analytic work of Turnquist and Blume (1980) examines conditions under which a threshold model is likely to produce benefits by reducing passenger

waiting and travel time. Using an analytic model, the research found that the benefits of holding are highest when the headway coefficient of variation (COV) is high and/or when the ratio of on-board passengers to expected downstream passengers is small. They note that this is somewhat conflicting: headway COV tends to increase along the route, while the ratio of on-board to downstream waiting passengers tends to decrease. Hence, the choice of a holding station is not entirely obvious. The research concludes that control should be enacted at a stop where there is already substantial variation in headways, the vehicle load is light, and the number of downstream passengers is significant. These results on the selection of control sites were echoed in subsequent studies by Turnquist (1981), Abkowitz and Engelstein (1984), Abkowitz and Tozzi (1986), Abkowitz et al. (1986), Vandebona and Richardson (1986), and Senevirante (1990), among others.

A much larger literature on transit vehicle holding has developed using heuristics and Monte Carlo simulation. This approach is appealing because vehicle operations on a transit route are inherently stochastic and hence difficult to describe analytically. Examples of route-level simulations that examine threshold- and schedule-based holding strategies include: Bly and Jackson (1974), Koffman (1978), Andersson et al. (1979), Andersson and Scalia-Tomba (1981), Abkowitz and Engelstein (1984), Abkowitz et al. (1986), Abkowitz and Tozzi (1986), Vandebona and Richardson (1986), Senevirante (1990), Lin et al. (1995), and Adamski and Turnau (1998). These models typically assume a random pattern of passenger arrivals at each stop (i.e., Poisson arrivals) and a binomial distribution for the number of alighting passengers at each stop. Generally, these studies have concluded that threshold-based holding is more effective than schedule-based holding, while both methods offer substantial improvements over no holding at all. In summary, the extensive simulation of holding strategies has resulted in certain threshold values that yield improved levels of service under specific operating conditions.

Recent studies by Eberlein (1995), Eberlein et al. (2001), Furth (1995), O'Dell and Wilson (1999), and Shen and Wilson (2000) present more general analytic models for the transit vehicle holding problem, using more detailed transit route models. These models each determine the optimal holding times for a set of vehicles at a dispatch point. More specifically, each model formulates the vehicle holding problem as a quadratic programming problem to solve for the optimal set of vehicle headways at a dispatch point. Eberlein provides an iterative heuristic to determine the optimal set of headways, while the models of Furth and O'Dell and Wilson solve these problems using nonlinear optimization methods. Shen and Wilson's model includes integer variables to accommodate short-turning and expressing of vehicles; their solution method uses a piecewise linear objective and linear constraints within a branch-and-bound technique. In each of these formulations, the transit service model includes more realistic characteristics of fixed-route service, such as minimum acceptable vehicle spacing and the dynamics of vehicle running and dwell times along a transit route. The models of O'Dell and Wilson and Shen and Wilson also include vehicle capacity constraints, although this requires integer variables in the formulation. The drawback to these analytic models, however, is that they use deterministic running and dwell times—there are no stochastic service attributes in these models.

A somewhat separate literature has developed that formulates the transit vehicle holding and dispatching problem as an optimal control problem. In this case, the decision to hold a vehicle is essentially one of determining the appropriate control, given a deviation from either the scheduled headway (in a headway-based control model) or the scheduled arrival and departure time (in a schedule-based control model). There has been considerable research and application of these models, including several works by Adamski (1979, 1983, 1993, 1996) and Adamski and Turnau (1998). In particular, these papers have offered analytic solutions to the threshold-based holding problem using optimal control techniques such as half-wave and shifted linear rectifiers, polynomial rectifiers, and step function controls. While optimal control provides a different modeling construct for the optimal transit vehicle holding problem, the resulting solution methods and solution properties are similar to the optimization-based methods cited above.

Also, similar to the cited work on threshold policies, the optimal control models show promise in reducing headway variability and improving on-time performance.

2. A Result due to Marguier

The literature indicates that stochastic service conditions have not been adequately addressed in the analytic models for the transit vehicle holding problem. This is somewhat surprising, in that there is a considerable literature on the stochastic aspects of transit operations. Stochastic arrival and boarding processes of passengers are described in the models of Jolliffe and Hutchinson (1975), Bowman (1979), Turnquist (1978), Bowman and Turnquist (1981), Marguier and Ceder (1984), Ceder and Marguier (1985), Guenthner and Sinha (1983), Adamski (1992), and Saidi et al. (1994). The modeling of transit running times and overall stochastic route performance measurement have been addressed in the analytic models of Andersson and Scalia-Tomba (1981), Powell and Sheffi (1983), Marguier (1985), and Adebisi (1986). These recent models build on a vast literature of route performance of a more deterministic nature, such as Newell and Potts (1964), Potts and Tamlin (1964), Heap and Thomas (1976), Newell (1976), and Chapman and Michel (1978).

Surprisingly, none of these operations models have explicitly considered how an operations control action (such as a vehicle hold) might affect route performance. Nonetheless, the more recent stochastic models of Andersson and Scalia-Tomba (1981) and Marguier (1985) calculate the first and second moments of vehicle travel times and of passenger boarding and alighting processes. In turn, these approximations can be used to estimate the impacts of the various holding actions on vehicle loads and headways and on passenger waiting times. Because these models explicitly incorporate stochastic elements in transit operations, they provide the foundation for the proposed holding model.

2.1. Assumptions and Notation

The following section describes the stochastic model of transit operations on a single route given by Marguier. Marguier's work describes the trajectories of transit vehicles on a single fixed route. The following assumptions are made.

1. Vehicles arrive frequently (i.e., headways of no more than 10–12 minutes) at each stop, making it appropriate to assume random passenger arrivals at each stop.

2. Passengers arrive randomly, as a homogeneous Poisson process, at each stop. In this case, we assume that the arrival rate does not change over the time period of interest.

3. Passenger waiting time is an important measure of performance. Because passenger arrivals are random, the waiting time is a function of the vehicle headway.

4. When a vehicle arrives at the stop, a subset of passengers will alight. Subsequently, all waiting passengers (for which this vehicle serves their destination stop) will board. That is, these processes occur in *series*. As noted by Boyd (1983), many previous studies simply assume that boarding and alighting can occur simultaneously and that boarding will dictate the total dwell time.

5. The number of passengers alighting at a stop is based on the current passenger load on the vehicle when it enters a stop; specifically, there is a binomial probability distribution of the number of passengers alighting at each stop. Moreover, the probability of a passenger alighting at a given stop does not change over the time period of interest. The binomial distribution was suggested by Andersson and Scalia-Tomba, under the assumption that passengers are independent (i.e., do not travel in groups). Other models use similar distributions (e.g., Powell and Sheffi 1983, Andersson et al. 1979, Adamski 1992). Marguier (1985) argues that the binomial distribution is appropriate if, "given the total load, the number of people in the origin-destination subgroups of this load are jointly multinomially distributed" (p. 76)—which is not unreasonable for smaller time periods (e.g., a morning peak period). The use of the binomial distribution is reasonable for modeling vehicle dynamics, but may not be entirely suitable for modeling passenger movements.

6. The incremental boarding and alighting times for each passenger are constant.

7. The travel time of a vehicle between stops has first and second moments, i.e., a mean and variance of between-stop travel times. The moments of this travel time also do not change over the time period of interest. Running times are assumed to be independent, reflecting both variations in traffic conditions and variability of driver behavior.

8. The boarding time of passengers that arrive at the stop while passengers are alighting and boarding is negligibly small.

With these assumptions, Marguier outlined a set of relationships governing vehicle and passenger movements on a single route. The following notation will be used in presenting his results.

- k is a subscript denoting the stop (or station); $k = 1, \ldots, N$.
- i is a subscript denoting the vehicle run number; $i = 1, \ldots, I$.
- L_{ik} is a random variable denoting the load on a vehicle i as it leaves stop k. In this manner, the load on vehicle i as it enters stop k is $L_{i,k-1}$.
- R_k is a random variable denoting the running time between stop $k-1$ and stop k.
- H_{ik} is a random variable denoting the headway between the preceding vehicle $i-1$ and vehicle i, as they depart stop k. That is, it is the time between the departure of vehicle $i-1$ at stop k and the departure of vehicle i from stop k.
- λ_k denotes the passenger arrival rate at stop k (passengers per unit time).
- p_k denotes the probability that a passenger who is on board the vehicle entering stop k will alight at stop k.
- a denotes the lost time due to accelerating and decelerating the vehicle at each stop.
- b_A denotes the incremental time for a single passenger to alight from the vehicle.
- b_B denotes the incremental time for a single passenger to board the vehicle.
- $E[\cdot]$ denotes the expected value of a random variable.
- Var$[\cdot]$ denotes the variance of a random variable.
- Cov$[\cdot, \cdot]$ denotes the covariance of two random variables.

2.2. Marguier's Model

The transit vehicle spends time both traveling between stops and dwelling at stops to allow passengers to board and alight. The time spent traveling between stops is a random variable (R_k), with expectation $E[R_k]$ and variance Var$[R_k]$. The time spent dwelling at each stop is given as the sum of the time spent accelerating and decelerating (a) and the time passengers spend alighting and boarding. Mathematically,

$$D_{ik} = a + b_A \cdot A_{ik} + b_B \cdot B_{ik}, \tag{1}$$

where D_{ik} is the dwell time for vehicle i at stop k, A_{ik} is a random variable of the number of people alighting from vehicle i at stop k, and B_{ik} is a random variable of the number of people boarding vehicle i at stop k. Using the properties of a binomial random variable, one arrives at the following mean and variance of the number of people alighting, given the expectation of the incoming vehicle load $L_{i,k-1}$:

$$E[A_{ik}] = p_k \cdot E[L_{i,k-1}], \tag{2}$$

$$\text{Var}[A_{ik}] = p_k \cdot (1 - p_k) \cdot E[L_{i,k-1}]. \tag{3}$$

Also, using the properties of a Poisson random variable, the mean and variance of the number of people boarding, given the expected headway when the vehicle enters the stop of $H_{i,k-1}$, is given as

$$E[B_{ik}] = \lambda_k \cdot E[H_{i,k-1}], \tag{4}$$

$$\text{Var}[B_{ik}] = \lambda_k \cdot E[H_{i,k-1}]. \tag{5}$$

In total, then, the dwell time at each stop has the following mean and variance:

$$E[D_{ik}] = a + b_A \cdot p_k \cdot E[L_{i,k-1}] + b_B \cdot \lambda_k \cdot E[H_{i,k-1}], \tag{6}$$

$$\text{Var}[D_{ik}] = b_A^2 \cdot p_k \cdot (1 - p_k) \cdot E[L_{i,k-1}] + b_B^2 \cdot \lambda_k \cdot E[H_{i,k-1}]. \tag{7}$$

The equations assume that a vehicle's headway and load are given. However, the headway and the load on vehicle i at any stop k depend on the trajectory and loading on that vehicle as it runs along the route. Specifically, Marguier identified the following dynamics. The headway of a vehicle i at a stop k depends

on the observed headway at the previous stop $k-1$, the running time for both vehicle $i-1$ and vehicle i to stop k, and the dwell time for vehicles $i-1$ and i at stop k. Mathematically,

$$H_{ik} = H_{i,k-1} + R_{ik} - R_{i-1,k} + D_{ik} - D_{i-1,k}. \quad (8)$$

With the assumptions above, the following expression gives the expected headway as it evolves with a particular vehicle i across the set of stops:

$$\begin{aligned} E[H_{ik}] &= E[H_{i,k-1}] + E[D_{ik}] - E[D_{i-1,k}] \\ &= E[H_{i,k-1}] + b_A p_k \cdot (E[L_{i,k-1}] - E[L_{i-1,k-1}]) \\ &\quad + b_B \lambda_k \cdot (E[H_{i,k-1}] - E[H_{i-1,k-1}]). \end{aligned} \quad (9)$$

Note that this requires knowing not only the expected headway but also the expected load on the vehicle.

In a similar manner as the headway, the load on vehicle i leaving stop k is given as a function of the incoming load as follows:

$$L_{ik} = L_{i,k-1} - A_{ik} + B_{ik}. \quad (10)$$

This gives the following expression for the expected load, as it evolves with a particular vehicle i across the set of stops:

$$\begin{aligned} E[L_{ik}] &= E[L_{i,k-1}] - E[A_{ik}] + E[B_{ik}] \\ &= (1-p_k) \cdot E[L_{i,k-1}] + \lambda_k \cdot E[H_{i,k-1}]. \end{aligned} \quad (11)$$

Again, the expected load depends on both the expected load and headway from the previous stop.

A more succinct matrix notation of the system evolution is as follows:

$$M_{ik} = F_k M_{i,k-1} + G_k M_{i-1,k-1}, \quad (12)$$

where

$$M_{ik} = \begin{bmatrix} E[H_{ik}] \\ E[L_{ik}] \end{bmatrix},$$

$$F_k = \begin{bmatrix} 1 + b_B \cdot \lambda_k & b_A \cdot p_k \\ \lambda_k & 1 - p_k \end{bmatrix},$$

$$G_k = \begin{bmatrix} -b_B \cdot \lambda_k & -b_A \cdot p_k \\ 0 & 0 \end{bmatrix}.$$

It is important to note that, mathematically, this model does allow vehicle overtaking. Under more extreme values of headways and loads, the expected headway could be negative; i.e., one vehicle overtakes another. This occurs when the expected dwell time of the previous vehicle $i-1$ exceeds the dwell time of the current vehicle i by at least the expected headway at the previous stop $k-1$. This occurs under some operating conditions where there is considerable variation in vehicle running times and passenger boarding and alighting times. Practically, however, whether or not "overtaking" at a stop k occurs is a matter that depends on local decision rules. In this model, we allow overtaking, and use a change of the vehicle index when overtaking occurs. However, it is important to note that this feature of Marguier's model makes it much more relevant to bus operations than to rail operations, which typically have physical and operational restrictions on vehicle overtaking.

Marguier also derived the variance of the headway and load, using similar system dynamics. In order to simplify his derivation, he made the following additional assumption.

9. Either (1) the time spent by a vehicle decelerating to and accelerating from a stop, a, is negligible, or (2) a vehicle dwells for a nonzero time at *all* stops, resulting in a being incorporated in each stop's R_k.

Together with the assumptions given above, this led Marguier to the following remarkable result. The result is given directly; readers seeking a detailed derivation of this result should consult the source (Marguier 1985). [There are several mathematical discrepancies between the equations in (13) and (14) and the matrix notation that appears in Marguier's dissertation. This paper corrects what appear to be typographic and transcription errors.] The derivation follows the same line of thought as the derivation of the expectations above but is much more mathematically tedious.

$$\begin{aligned} V_{ik} &= 2F_k S_k F_k^T + 2G_k S_k G_k^T - F_k S_k G_k^T - (F_k S_k G_k^T)^T \\ &\quad + F_k V_{i,k-1} F_k^T + G_k V_{i-1,k-1} G_k^T + F_k Q_{i,k-1} G_k^T \\ &\quad + (F_k Q_{i,k-1} G_k^T)^T + \overline{F}_k \overline{M}_{i,k-1} F_0^T + \overline{G}_k \overline{M}_{i-1,k-1} G_0^T, \end{aligned} \quad (13)$$

where F_k and G_k are the matrices defined previously, and

$$V_{ik} = \begin{bmatrix} \text{Var}[H_{ik}] & \text{Cov}[H_{ik}, L_{ik}] \\ \text{Cov}[H_{ik}, L_{ik}] & \text{Var}[L_{ik}] \end{bmatrix},$$

$$S_k = \begin{bmatrix} \text{Var}[R_k] & 0 \\ 0 & 0 \end{bmatrix},$$

$$Q_{ik} = \begin{bmatrix} \text{Cov}[H_{ik}, H_{i-1,k}] & \text{Cov}[H_{ik}, L_{i-1,k}] \\ \text{Cov}[H_{i-1,k}, L_{ik}] & \text{Cov}[L_{ik}, L_{i-1,k}] \end{bmatrix},$$

$$\overline{F}_k = \begin{bmatrix} b_B \cdot \lambda_k & -b_A \cdot p_k \cdot (1-p_k) \\ \lambda_k & p_k \cdot (1-p_k) \end{bmatrix},$$

$$\overline{G}_k = \begin{bmatrix} b_B \cdot \lambda_k & -b_A \cdot p_k \cdot (1-p_k) \\ 0 & 0 \end{bmatrix},$$

$$F_0 = \begin{bmatrix} b_B & -b_A \\ 1 & 1 \end{bmatrix},$$

$$G_0 = \begin{bmatrix} b_B & -b_A \\ 0 & 0 \end{bmatrix},$$

$$\overline{M}_{ik} = \begin{bmatrix} E[H_{ik}] & 0 \\ 0 & E[L_{ik}] \end{bmatrix}.$$

In (13), the first four terms (the first line in the equation) represent the effect on the variance of the headway and load that is caused by the variability in running times. Higher variability in running times would be expected to lead to higher variability in headways and loads at each stop. The fifth and sixth terms (in the variance-covariance matrix V) represent the propagation of stop-to-stop irregularity—one might expect the variance to propagate along a route. With increasing variance of headway and load at one stop, the subsequent stop will also have higher variance of the headway and load. First-order trip-to-trip instability is described in the seventh and eighth terms. These terms represent the effect of what has typically been observed as negative correlation between headways and loads of adjacent vehicles that may result in vehicle "bunching." That is, as one headway becomes longer, the subsequent headway becomes shorter, with the result that two vehicles move closer together in time and space as they move down the route. Finally, the last two terms in (13) are terms in expected values, since the variance also depends on the expected headway and load for that vehicle.

The other quantity yet to be defined is the matrix of lagged covariances Q_{ik}. Marguier also derives this matrix to be

$$\begin{aligned} Q_{ik} = {} & F_k Q_{i,k-1} F_k^T + G_k V_{i-1,k-1} F_k^T \\ & + G_k Q_{i-1,k-1} G_k^T + F_k S_k G_k^T + (F_k S_k G_k^T)^T \\ & - F_k S_k F_k^T - \overline{G}_k \overline{M}_{i-1,k-1} \overline{F}_0^T \end{aligned} \tag{14}$$

with

$$\overline{F}_0 = \begin{bmatrix} b_B & 0 \\ 1 & 1 \end{bmatrix}.$$

The first three terms in (14) represent propagation of lagged covariances as a vehicle moves down a route. The fourth through sixth terms represent the effect of running time variance on the correlation of headways and loads between vehicles. Finally, the last term includes the effect of the current headways and loads for vehicles i and $i-1$ on the covariance between the two vehicles.

Hence, first and second moments of the vehicle load and the vehicle headway, across vehicles and across stops on a transit route, can be defined through iterative use of (12), (13), and (14), with appropriate boundary conditions for the vehicles as they are dispatched from stop 1. This was one of the major results of the work of Marguier.

The reason why Marguier's result is so remarkable is that the first and second moments of vehicle headways and loads, given in the system dynamics of (12), (13), and (14), are linear. This fact can be exploited in understanding the effects of changes in system behavior under various control actions, as will be shown later.

One additional result due to the system dynamics, not noted by Marguier, is formalized in the following theorem.

THEOREM 1. *Under the assumptions of* (12), (13), *and* (14), $\text{Cov}[H_{ik}, L_{ik}] \geq 0\ \forall i, k$ *and all four covariance terms in* Q_{ik} *are* $\leq 0\ \forall i, k$. *As a result, the headway variance* $\text{Var}[H_{ik}]$ *is a strictly monotonic (increasing) function of the stop* k.

PROOF. The proof is straightforward from multiplication of terms in (13) and (14). □

The conditions of Theorem 1 are fairly intuitive. The condition that $\text{Cov}[H_{ik}, L_{ik}] \geq 0 \ \forall i, k$ states that the headway and load should not be negatively correlated. Rather, one might expect a longer headway to result in higher vehicle loads, and smaller headways to produce smaller vehicle loads. The second condition, that the covariance terms in Q_{ik} should not exhibit positive correlation, implies that one might expect a longer headway to be followed by a shorter headway (and vice versa), and a high vehicle load to be followed by a smaller vehicle load (and vice versa).

2.3. Extension to Real-Time Data

Marguier's modeling construct is static in the sense that it does not reflect how the evolution of the vehicle trajectories may be updated with real-time data. For the purposes of this analysis, with existing technology, one may assume the availability of both vehicle location data and passenger load information at a stop (Casey 1999, Casey et al. 2000). To illustrate this point, by 1999, 61 transit agencies had automatic vehicle location (AVL) systems in place in the United States, and 100 more agencies were installing or planning to install AVL systems in the near future (Casey 1999). The same report cites 24 agencies with operational Automatic Passenger Counter (APC) systems in the United States, and 30 more agencies that were installing or planning to install APC systems. While the integration of AVL and APC systems is not without its technical challenges (Schiavone 1999), several agencies in the United States now have integrated systems that are communicating passenger boardings and alightings in real time (Casey et al. 2000).

With these capabilities, Marguier's model may be extended to include the following basic information that might be known to a driver or to field supervisory personnel when a vehicle arrives at a stop:

- The observed load (number of persons) on board the vehicle i entering the stop k, $L_{i,k-1}$.
- The observed number of passengers waiting at the stop, denoted $\tilde{N}_k$.
- The time since the previous vehicle $i-1$ left the current stop k, denoted $\tilde{H}_{ik}$.
- The latest recorded headway and load of the immediately preceding vehicle $i-1$.
- The latest recorded headways and loads of all vehicles *upstream* of the current vehicle i and stop k. That is, the model assumes to know H_{mn} and L_{mn} for all vehicles $m > i$ and at each vehicle's latest stop, $n \leq k$.

Under these conditions, the expected headway and load, and the associated variances and covariances of vehicle i as it enters stop k, are calculated using the following set of equations. These are easily shown to be straightforward extensions of the previous modeling results that account for the realizations of random variables that are known to a driver or supervisor once a vehicle arrives at the stop.

$$E[H_{ik}] = \tilde{H}_{ik} + b_A p_k \cdot L_{i,k-1} + b_B \cdot \tilde{N}_k, \quad (15)$$

$$E[L_{ik}] = (1-p_k) \cdot L_{i,k-1} + \tilde{N}_k, \quad (16)$$

$$\text{Var}[H_{ik}] = b_A^2 p_k (1-p_k) \cdot L_{i,k-1}, \quad (17)$$

$$\text{Var}[L_{ik}] = p_k (1-p_k) \cdot L_{i,k-1}, \quad (18)$$

$$\text{Cov}[H_{ik}, L_{ik}] = b_A p_k (1-p_k) \cdot L_{i,k-1}, \quad (19)$$

$$Q_{ik} = 0 \quad \text{(all lagged covariance terms are 0)}. \quad (20)$$

Using realizations of random variables rather than expectations, the lagged covariance terms associated with the previous vehicle $i-1$ (i.e., the terms in Q_{ik}) are zero. A similar extension to the expectations, variances, and covariances for subsequent vehicles (i.e., for vehicles $i+j$, $j \geq 1$) at stop k are determined through a similar argument as

$$E[H_{i+j,k}] = (1 + b_B \lambda_k) \cdot (E[H_{i+j,k-1}] + \Delta R) + b_A p_k \cdot L_{i+j,k-1} - E[D_{i+j-1,k}], \quad (21)$$

$$E[L_{i+j,k}] = (1-p_k) \cdot L_{i+j,k-1} + \lambda_k \cdot (E[H_{i+j,k-1}] + \Delta R), \quad (22)$$

where ΔR is the difference in running times for vehicles $i+j$ and $i+j-1$, with $\Delta R = E[R_k] - R_{ik}$ for $j = 1$ (since R_{ik} is known with certainty) and $\Delta R = 0$ otherwise ($j \geq 2$). Similarly, with R_{ik} known, the variances and covariances $V_{i+1,k}$ and $Q_{i+1,k}$ are recalculated as

$$V_{i+1,k} = 2F_k S_k F_k^T + F_k V_{i+1,k-1} F_k^T + F_k Q_{i+1,k-1} G_k^T + (F_k Q_{i+1,k-1} G_k^T)^T + \overline{F}_k \overline{M}_{i+1,k-1} F_0^T, \quad (23)$$

$$Q_{i+1,k} = F_k Q_{i+1,k-1} F_k^T. \quad (24)$$

The variance and covariance matrices for $j \geq 2$ are determined as before, using (13) and (14).

Vehicle expectations, variances, and covariances are then determined for stops $k+1, \ldots, N$ using these initial conditions (15–24) and the evolution from (12), (13), and (14).

2.4. Example

In order to clarify the model, a small example from Marguier is used. A single bus route has 10 stops, with buses being dispatched from stop 1 and reaching the terminus at stop 10. For illustrative purposes 10 buses are dispatched from stop 1 at 6-minute headways. This headway is sufficiently short that, during the time period of interest, an assumption of randomly arriving passengers at all stops is justified. The expected running time from one stop to the next is five minutes in all cases, resulting in a total expected one-way trip running time of 45 minutes from stop 1 to stop 10. Expected dwell times add approximately 5 minutes to this time, for a total of 50 minutes in one direction. Table 1 gives the parameters for the Poisson passenger arrival rates at each stop λ_k (in passengers per minute); the probability an arriving passenger alights at each stop p_k; and the expected running time and the variance of the running time to each stop $E[R_k]$ and $\mathrm{Var}[R_k]$ (in minutes and squared minutes, respectively). One may note that the coefficient of variation for the running times is at most 0.2, reflecting considerable variation in traffic congestion and in driver behavior. In addition, the incremental boarding time b_B per passenger is 0.05 of a minute (3.0 seconds), and the incremental alighting time per passenger b_A is 0.03 of a minute (1.8 seconds).

The initial conditions are that each bus is dispatched from stop 1 precisely at its given headway, with $\mathrm{Var}[L_1] = E[L_1]$ and all other variance and covariance terms equal to 0. The expectations and variances of bus headways and loads on the route, using (12), (13), and (14), are shown in Table 2 and illustrated in Figure 1. Note that all bus headways and loads will be identical, in expectation; hence, only one bus headway and load are shown in the table. As will be illustrated later, the expected total waiting time for randomly arriving passengers, as calculated using (28), is 2185.2 minutes. Without the variance terms, the expected waiting time is 1755 minutes; this implies that headway variability adds 430.2 minutes (about 24.5%) to the expected passenger waiting time.

From Table 2, the variance of the headway increases over the course of the route, while the variance of the load shows higher variance toward the center of the route, where the expected load is highest. The coefficient of variation of the headway and load over the route is shown in Figure 2. This figure more clearly illustrates the increase in the variability of the headway as the bus moves down the route. Also, Figure 2 shows that while the variance of the load is highest in the middle of the route, the coefficient of variation of the load also tends to increase along the route.

A Monte Carlo simulation was developed to generate specific instances of route operating conditions. Random passenger arrivals were generated using a

Table 1 Example Bus Route Parameters

Stop	λ_k	p_k	$E[R_k]$	$\mathrm{Var}[R_k]$
1	0.75	0.0	—	—
2	1.5	0.0	5.0	0.8
3	0.75	0.1	5.0	0.2
4	3.0	0.25	5.0	1.0
5	1.5	0.25	5.0	0.4
6	1.0	0.5	5.0	0.4
7	0.75	0.5	5.0	0.4
8	0.5	0.1	5.0	0.1
9	0.0	0.75	5.0	0.6
10	0.0	1.0	5.0	0.6

Table 2 Example Bus Trajectories

Stop	$E[H]$	$E[L]$	$\mathrm{Var}[H]$	$\mathrm{Var}[L]$
1	6.00	4.50	0.00	4.50
2	6.00	13.50	2.03	17.10
3	6.00	16.65	2.77	25.15
4	6.00	30.49	7.49	101.29
5	6.00	31.87	11.03	142.88
6	6.00	21.93	15.70	96.25
7	6.00	15.47	20.39	68.65
8	6.00	16.92	22.63	94.50
9	6.00	4.23	27.06	9.08
10	6.00	0.00	29.40	0.00

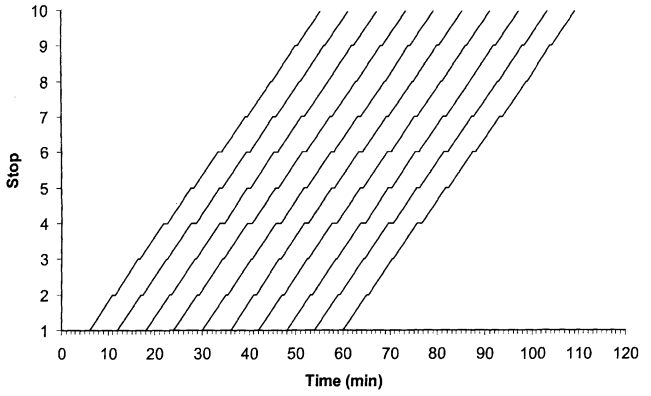

Figure 1 Expected Bus Trajectories

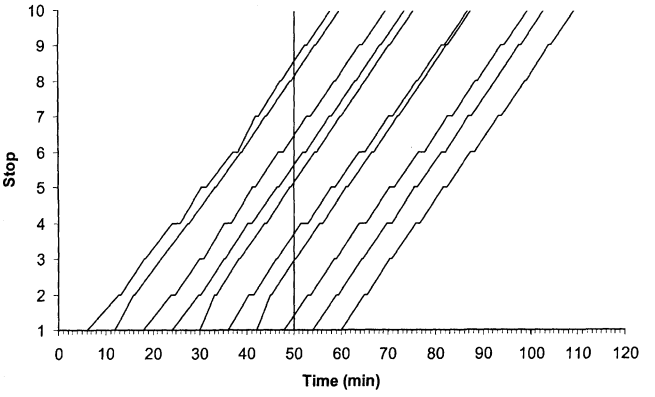

Figure 3 Example Bus Trajectories at $t = 50$ minutes

negative exponential distribution with parameter λ_k at the 10 stops. When a bus arrives at a stop, the number of alighting passengers is generated randomly by sampling from a binomial distribution. Finally, running times are generated randomly using a lognormal distribution with mean and variance parameters as in Table 1.

One illustrative set of bus trajectories from the simulation is shown in Figure 3, taken at a time 50 minutes into the simulation. The last noted stops for the earlier buses are given in Table 3. The values in this table are generated using data available from a real-time monitoring system: the time of arrival of a bus at each stop, the number of waiting passengers at the stop when it arrives, and the load on the bus when it enters the stop. The expected headway and load from the table are calculated using (15) and (16). Other initial conditions (not shown in the table) are similarly calculated using (17–24).

The expected headways and loads, and associated variances and covariances, are then calculated iteratively using (12), (13), and (14). As an example, the mean and variance of the headway and load for bus 7 as it continues along the route from stop 2 is given in Table 4. Note that there is still considerable variability in the headway and load for bus 7. Also, the evolution results in bus 7 getting closer and closer to bus 6. Similar bunching of trajectories is shown in Figure 3 for buses 4 and 5, while the headways for bus 8 are already anticipated to be very large.

At this state of the system, the expected remaining waiting time for the 10 buses as they ply the remainder of the route is found using Equation (28) and is 1318.2 minutes. This excludes the waiting times observed in the first 50 minutes; if this is included, the total wait sums to 2281.9 minutes.

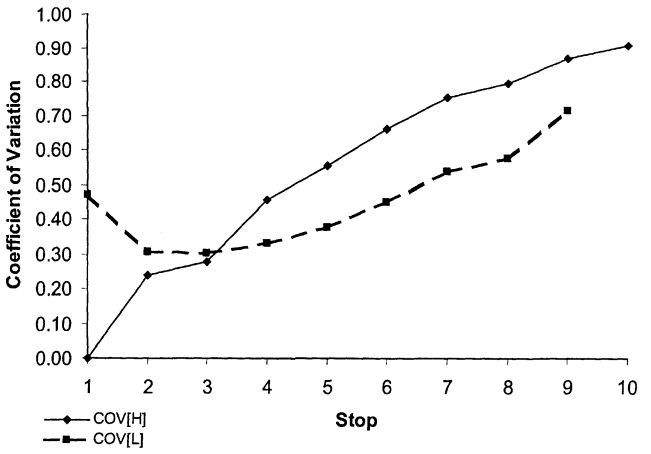

Figure 2 Coefficient of Variation of Bus Headway and Load

Table 3 Example Bus Route Conditions After 50 Minutes

Bus	Last Stop	Time Arriving Last Stop	Expected Headway	Expected Load
1	8	47.11	N/A	21.80
2	8	49.25	2.14	7.50
3	6	46.72	8.78	31.50
4	5	46.49	5.08	24.25
5	5	49.14	2.47	18.60
6	3	46.25	8.07	26.60
7	2	44.99	3.89	11.00
8	1	48.00	6.00	6.00

Table 4 **Evolution of Bus 7 in Example From $t = 50$ Minutes**

Stop	$E[H]$	$E[L]$	Var$[H]$	Var$[L]$
2	3.89	11.00	0.00	0.00
3	3.91	12.93	0.44	4.14
4	3.21	21.42	3.72	40.35
5	2.63	20.88	5.85	63.70
6	1.91	13.07	8.49	46.57
7	1.31	7.97	11.21	34.09
8	1.02	7.82	12.46	47.00
9	0.59	1.96	15.23	4.40
10	0.45	0.00	16.99	0.00

3. The Vehicle Holding Problem

Consider the following operational strategy to improve transit service. Under some conditions, it may be advantageous to hold a vehicle at a stop, so as to provide some regularity of headways on the route. By holding a vehicle at a stop, the total waiting time for all passengers may be reduced, even though there may be some slight inconvenience to those who are on board the vehicle as it holds. One challenge in determining the holding time is to minimize passenger waiting time without adding too much additional idle time for those on board.

As explained by Welding (1957) and Osuna and Newell (1972), the expected waiting time for a randomly arriving passenger is given as a function of both the mean and variance of the headway:

$$E[WT] = \frac{E[H]}{2} \cdot \left(1 + \frac{\text{Var}[H]}{E[H]^2}\right), \tag{25}$$

where $E[WT]$ is the expected waiting time per person, $E[H]$ is the expected headway, and Var$[H]$ is the variance of the headway. The total expected waiting time of all passengers at a particular stop, assuming random arrivals according to a Poisson process of rate λ_k, is given as

$$E[TW_k] = \frac{\lambda_k}{2} \cdot \sum_{i=1}^{I} E[H_{ik}^2], \tag{26}$$

where $E[TW_k]$ is the expected total waiting time at stop k, H_{ik} is the headway of vehicle i at the stop, and I is the total number of vehicles arriving at the stop. The total expected waiting time spent at all stops, $E[TW]$, is given by

$$E[TW] = \sum_{k=1}^{N} \frac{\lambda_k}{2} \cdot \sum_{i=1}^{I} E[H_{ik}^2] \tag{27}$$

$$= \sum_{k=1}^{N} \frac{\lambda_k}{2} \cdot \sum_{i=1}^{I} (\text{Var}[H_{ik}] + E[H_{ik}]^2). \tag{28}$$

In addition, the expected delay to passengers on board a vehicle at stop k as it is held for a time t is given by $E[L_{ik}] \cdot t$.

Consider now the problem that the transit operator (driver) or supervisor faces once a given vehicle i enters a stop k. Once the vehicle enters the stop, it is up to the operator or supervisor to determine whether or not the vehicle should be held at the stop and, if so, for how long. Note in this case that the decision is defined as determining the desired hold for the current vehicle, *independent* from any future holds or other control actions that might be taken on subsequent vehicles. In this sense the resulting holding decision is determined in real time, for a single vehicle, using the most recent information available. However, this also assumes that multiple holds will not be considered simultaneously (i.e., the decision is strictly local).

The information available for this decision is given in §2.4: the observed load entering the stop $L_{i,k-1}$, the observed number of passengers waiting $\tilde{N}_k$, the time since the previous vehicle left the current stop $\tilde{H}_{ik}$, the latest recorded headway and load of the immediately preceding vehicle $i-1$, and the last stop and the latest expected headways and loads of all vehicles *upstream* of the current vehicle. The on-vehicle and in-stop number of passengers may be observed either by the operator or by the transit supervisor. One might assume that current schedule adherence and estimated loads are available from a central or distributed database and communicated over a wireless or wireline network.

3.1. Model Development and Formulation

The objective of the decision to hold, and how long to hold, is to minimize the total passenger waiting time and the delay to on-board passengers. It is important to note that, if a hold is enacted, it will also affect

the trajectories of subsequent vehicles on the route, from the control stop k to all downstream stops. In particular, the effect of the hold includes changes in headways for the current vehicle i as well as all subsequent vehicles $m > i$, and for all stops from k to the end of the line N. The objective function for the holding problem then becomes a variant of (28), adding in the delay to on-board passengers:

$$\min_{t\geq 0} E[TW] = \sum_{m=k}^{N} \frac{\lambda_m}{2} \cdot \sum_{n=i}^{I} \left(\mathrm{Var}[H_{mn}|t] + E[H_{mn}|t]^2\right) + \theta \cdot E[L_{ik}] \cdot t, \quad (29)$$

where t is the length of time the vehicle i is held at stop k, after passengers have alighted and boarded. In this case, the model includes a parameter θ that allows one to weight this on-board delay less than, equal to, or greater than the value of waiting time at a stop.

Also, the objective includes an assumed set of stops ($m = k$ to N) and set of vehicles ($n = i$ to I) that will be affected by the hold. The choice of N and I as the range for the impact analysis is important. The stop N is chosen because it is a downstream stop where additional control may be applied. The work of Eberlein (1995, 2001) and the analysis here suggests that the vehicle-to-vehicle effect of a hold drops off rapidly, so that a small number of vehicles (i.e., three or four vehicles after i) is likely to be sufficient to capture the main effects of the hold. Nonetheless, since the computational burden is small, the full set of vehicles is advocated for this analysis.

The expression for $E[TW]$ in (29) constitutes the objective function for the analytic model. Note that the function is uniquely comprised of terms in the variance and expected value of the headways, only for vehicle movements "downstream" (in both space and time) from the current stop and time.

The derivation of the analytic model then depends on understanding the system dynamics once a hold is (or is not) enacted on the current vehicle i at stop k. Once the vehicle i arrives at the stop k, if *no hold* is enacted, the expected headway and load for the vehicle are determined using (15–24). Note that these equations give a set of "initial" conditions for the system evolution described in (12), (13) and (14).

Alternatively, one might expect that if a hold of duration $t > 0$ is enacted on vehicle i at stop k, then a reformulation of (15–20) with a holding time of t yields:

$$E[H_{ik}|t] = t + \widetilde{H}_{ik} + b_A p_k \cdot L_{i,k-1} + b_B \cdot \widetilde{N}_k, \quad (30)$$

$$E[L_{ik}|t] = (1-p_k) \cdot L_{i,k-1} + \widetilde{N}_k + \lambda_k \cdot t, \quad (31)$$

$$\mathrm{Var}[H_{ik}|t] = b_A^2 p_k (1-p_k) \cdot L_{i,k-1} + b_B^2 \lambda_k \cdot t, \quad (32)$$

$$\mathrm{Var}[L_{ik}|t] = p_k (1-p_k) \cdot L_{i,k-1} + \lambda_k \cdot t, \quad (33)$$

$$\mathrm{Cov}[H_{ik}, L_{ik}|t] = b_B \lambda_k \cdot t - b_A p_k (1-p_k) \cdot L_{i,k-1}, \quad (34)$$

$$Q_{ik} = 0. \quad (35)$$

Within this set of equations, the assumption is made that the holding time may be sufficiently long that passenger arrivals at the holding stop cannot be neglected. Marguier had assumed that arrivals during boarding and alighting are negligible (see assumption 8), but this assumption is likely to be violated if the hold is long. For this reason, (30–35) include terms reflecting arrivals during the holding time t. The expected load increases by $\lambda_k \cdot t$, and the variance and covariance terms are similarly adjusted. Of particular note is the fact that a hold of t increases the variance of the headway.

If a hold is enacted on vehicle i at stop k, the additional time that vehicle i spends at stop k will affect the dwell times of subsequent vehicles $i+j$, $j \geq 1$, at that stop. In particular, the expected headways, loads, and variance-covariance terms no longer evolve strictly according to (12), (13), and (14). Instead, the change in headway for vehicle i at k will lead to a shorter headway for the subsequent vehicle $i+1$ at k, since those passengers arriving during t will board vehicle i instead of $i+1$. Hence, the corresponding initial conditions for the expected headway and load of vehicle $i+j$, $j \geq 1$, at stop k are:

$$E[H_{i+1,k}|t] = E[H_{i+1,k}] - \left(\frac{1}{1-b_B\lambda_k}\right) \cdot t \quad (36)$$

$$\text{and } E[H_{i+j,k}|t] = E[H_{i+j,k}] + \left(-\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j \cdot t \quad \text{for } j \geq 2, \quad (37)$$

$$E[L_{i+1,k}|t] = E[L_{i+1,k}] - \left(\frac{1}{1-b_B\lambda_k}\right)\lambda_k \cdot t \quad (38)$$

$$\text{and } E[L_{i+j,k}|t] = E[L_{i+j,k}] + \left(-\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j \lambda_k \cdot t \quad \text{for } j \geq 2. \quad (39)$$

Note that a hold of t on vehicle i causes a direct reduction of t in the headway of vehicle $i+1$. Also, this headway is also slightly shorter because fewer people will have arrived during that shorter headway. In particular, the factor $(\frac{1}{1-b_B\lambda_k})$ includes the small fraction of time that is eliminated by *not* boarding passengers arriving during t. For subsequent vehicles $i+j$, $j \geq 2$, the only contributing effect on the headway is the time saved (or lost) by not boarding (or boarding) passengers arriving during the slight perturbation to the previous headway. A similar argument motivates (38) and (39) for the expected load.

An important fact in the initial conditions (36–39) is that the effect of a hold of t on subsequent vehicles at stop k drops off quickly, assuming $b_B\lambda_k$ is small. For example, if passengers board 10 times faster than they arrive (e.g., $b_B\lambda_k = 0.1$), then the effect of the holding time t drops off by approximately an order of magnitude for every subsequent vehicle at k after vehicle $i+1$.

A similar derivation of variance and covariance terms, based on a hold of t, yields the following adjustments for initial conditions at stop k for vehicles $i+j$, $j \geq 1$:

$$\text{Var}[H_{i+j,k}|t] = \text{Var}[H_{i+j,k}] + \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j \times \frac{b_B \cdot t}{1-b_B\lambda_k}, \quad (40)$$

$$\text{Var}[L_{i+j,k}|t] = \text{Var}[L_{i+j,k}] + \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j \times \lambda_k \cdot t, \quad (41)$$

$$\text{Cov}[H_{i+j,k}, L_{i+j,k}|t] = \text{Cov}[H_{i+j,k}, L_{i+j,k}] + \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j b_B\lambda_k \cdot t, \quad (42)$$

$$\text{Cov}[H_{i+j+1,k}, H_{i+j,k}|t] = \text{Cov}[H_{i+j+1,k}, H_{i+j,k}] - \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j b_B^2\lambda_k \cdot t, \quad (43)$$

$$\text{Cov}[H_{i+j+1,k}, L_{i+j,k}|t] = \text{Cov}[H_{i+j+1,k}, L_{i+j,k}] - \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j b_B\lambda_k \cdot t, \quad (44)$$

$$\text{Cov}[H_{i+j,k}, L_{i+j+1,k}|t] = \text{Cov}[H_{i+j,k}, L_{i+j+1,k}] - \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j b_B\lambda_k \cdot t, \quad (45)$$

$$\text{Cov}[L_{i+j+1,k}, L_{i+j,k}|t] = \text{Cov}[L_{i+j+1,k}, L_{i+j,k}] - \left(\frac{b_B\lambda_k}{1-b_B\lambda_k}\right)^j \lambda_k \cdot t. \quad (46)$$

These equations (40–46) are derived by direct calculation of the variance and covariance of the perturbations to the headways and loads of vehicles at stop k. Again, note that a holding time of t results in increases to the variance of headways, as included in (40).

Mathematically, the choice of whether or not to hold a vehicle i at stop k, and the length of the holding time t, can be determined through the following univariate quadratic program:

$$\min_{t \geq 0} Z(t) = E[TW|t] = \sum_{m=k}^{N} \frac{\lambda_m}{2} \cdot \sum_{n=i}^{I} (\text{Var}[H_{mn}|t] + E[H_{mn}|t]^2) + \theta \cdot E[L_{ik}] \cdot t \quad (47)$$

subject to:

System evolution constraints: (12), (13), and (14)

$\forall m = k, \ldots, N, \; n = i, \ldots, I$

Nonnegativity constraints: $t \geq 0$

Initial conditions: (15–24), (30–35), and (36–46)

Note that all constraints are linear in the decision variable t, and the objective function is a quadratic function in the decision variable t.

3.2. Solution Technique

An analytic solution for this univariate quadratic program is possible using first-order conditions on the objective function, although this is tedious to derive. A more straightforward approach, at the cost of slightly greater computational time, results from

recognizing that the objective function (47) of the quadratic program is convex in the decision variable t, subject to linear equality constraints. The following theorem makes this explicit.

THEOREM 2. *With the substitution of linear equality constraints from* (12), (13), *and* (14), (15–24), *and* (30–46), *the objective function in* (47) *is strictly convex in* t, *for* $t \geq 0$.

PROOF. First, note that all terms in the initial conditions (30–46) are linear in t. Second, since the system evolution constraints (12), (13), and (14) are also linear, $E[H_{nm}]$ and $\text{Var}[H_{nm}]$, for $n \geq i$ and $m \geq k$, are also linear in t. Then, $\frac{dE[H_{nm}]}{dt}$ is a nonzero constant for $t > 0$. The second derivative of $Z(t)$ is then

$$\frac{d^2Z(t)}{dt^2} = 2 \cdot \left(\frac{dE[H_{nm}]}{dt} \right)^2, \tag{48}$$

which is strictly positive for all values of $t > 0$. Because the second derivative is strictly positive, $Z(t)$ is strictly convex. □

Hence, any local minimum of the objective function (47) is also a global minimum. Rather than solving for first-order conditions, a simple line search technique is adequate for finding the optimal holding time t. The only difficulty of such a line search is the calculation of the objective function values, which (though tedious by hand) can be calculated very quickly using a computer. In the example problem shown later, calculations take only a small fraction of a second using a 166 MHz Pentium computer, and could easily be implemented on less powerful computing platforms with more than adequate speed for on-line decision support.

While the holding problem statement and solution approaches have some similarities, the model here is not entirely comparable to the work of Eberlein (1995), Eberlein et al. (2001), Furth (1995), O'Dell and Wilson (1999), and Shen and Wilson (2000) for the vehicle holding problem. The main differences are as follows.

1. The model here is used for the holding of a single vehicle, while the other models mentioned are applied to a larger set of vehicles. Particularly, the other models are intended to solve for the optimal holding times for *all* vehicles (i through I) at a particular control point (although one might reoptimize upon each vehicle arrival at the control point). The model presented here only solves for a single vehicle i at the control point.

2. The operations model of Marguier allows vehicle overtaking, which is explicitly prohibited in the other models. This is an important limitation, so that the model is not useful for rail systems (in contrast to the models of Eberlein, O'Dell and Wilson, and Shen and Wilson).

3. The model presented here explicitly incorporates stochastic vehicle travel times, and vehicle dwell times, in the model. The other models are strictly deterministic.

The myopia noted in the first point above is a limitation of the model here, and an area of further research. However, according to Turnquist (1989), it is unlikely that two successive vehicles will receive a hold under minor service disruptions. Turnquist observed that unless the observed percentage deviations from the desired headway are large (greater than six times the ratio of on-board passengers to downstream passengers), the optimal control action should be enacted on a single vehicle independently, and generally will not be applied on two or more successive vehicles. This suggests that the benefits of a model that optimizes across all vehicles is likely to be small compared to the proposed model. At the same time, Turnquist's model only considers minor service disruptions and is not applicable to major disruptions.

The third point above is also very important. Because the variance of the headway is a strictly increasing function of the stop (by Theorem 1), small changes in headways at the control point may propagate into significant changes in variability of the headway at downstream stops. As a result, this model is *more sensitive* to the downstream effects of a particular holding decision than a deterministic model. This is as it should be: Because the variability of service increases down the line, the decision to hold a vehicle at a stop should yield substantial effects in the variability of service at downstream nodes, and more so as the vehicle gets further from the control point.

4. Holding Example

The holding model presented above is illustrated using a simple hypothetical bus route. In the first

part of this section, the bus route and its modeling is described, and the holding results for this route are described in the second part. The purpose of this example is to illustrate how the model may be applied in practice, and how its performance may be measured. It is important to keep in mind, however, that these results are illustrative and not intended to be conclusive about the value of holding strategies using this model.

4.1. Study Design

The same bus route described in §2.4 is used in this section. Hence, Table 1 gives the various parameters, while Table 2 and Figure 1 illustrate the expected trajectories for this example. The simulation model was further developed to examine holding actions at a particular control stop on the route. The simulation is an event-based simulation, with the trigger event being the arrival of a bus at a particular stop. When a bus i arrives at a stop k, the expected headways, loads, and associated variance and covariance matrices are updated, according to (15–24) and then by (30–46) if a hold is enacted. These expectations and variance and covariance matrices are then evolved up to stop 10 (the end of the route) using (12), (13), and (14). Finally, changes in waiting time are then computed using the expectation and variance of headways in (29). The simulation then continues on to the next "scheduled" event—another bus arrival at a stop. This process repeats until all buses have reached the end of the route.

For the purposes of illustration, holding was considered at stop 3. With the selected route parameters, this stop is just upstream of the peak loading point (stop 4) and the heaviest load segment (stop 5 to stop 6). This is consistent with the results of most simulation experiments with holding, which suggest that a stop just upstream of the maximum load point is likely to provide the highest reduction in passenger waiting times from a holding action. Using the analysis of Turnquist and Blume (1980) directly, with the parameters in Tables 1 and 2, stops 2 and 3 are acceptable candidates for holding. Finally, for the purposes of weighting the value of passenger delay on board versus passenger waiting time, θ was chosen to be 0.5. This corresponds to a value of waiting time that is twice as large for passengers at a stop than it is for passengers on board.

When a bus arrives at stop 3 (the control stop), the simulation calculates the expected headways and loads, and the variance-covariance matrices, using (15–24). These are evolved to the end of the route (stop 10) using (12), (13), and (14). The objective function in (47) is evaluated, and then a holding decision is made using the following algorithm.

Algorithm for the Holding Problem.

Initialize $n = 0$.

Calculate (15–24) for this bus i at the control stop k.

Evolve the expectations and variance-covariance matrices for i, from $k+1$ to K, using (12), (13), and (14).

Calculate the expected total wait with $t = 0$ in (47): let this be TW_0.

while ($n = 0$ or $TW_n < TW_{n-1}$) **do**

{

Set $n = n+1$.

Set $t = stepsize \cdot n$.

Calculate (30)–(35) for this bus.

Calculate (36)–(46) for subsequent buses at the control stop k.

Evolve the expectations and variance-covariance matrices for this bus i and all subsequent buses (to I), from stop $k+1$ to N, using (12), (13), and (14).

Calculate the objective function TW_n using (47).

}

Set optimal holding time $t = stepsize \cdot (n-1)$.

The critical part of this algorithm is a simple line search on t, using for illustrative purposes a *stepsize* of 0.05 minutes (3 seconds). The effect of the hold is measured in terms of the waiting time and delay to on-board passengers using (47). In this example the control stop is stop 3, and holds are considered for all buses up to bus 10. However, to avoid end-of-horizon effects, the total number of buses in the simulation was increased to 15; i.e., $I = 15$. This means the effect of holding was measured up to stop 10 (the end of the line) and to all buses to $I = 15$. One simulated period includes the movement of all 15 buses from stop 1 to stop 10. However, in reporting the results, only passenger waiting times up to bus 10 were calculated.

The trajectories from a single simulation run, identical to the example in §2.4 and Figure 3 but with holding, are shown in Figure 4. In this case, holds are enacted on buses 2, 5, 7, and 10 for 1.85 minutes, 2.25 minutes, 1.75 minutes, and 0.7 minutes, respectively. Note that even with these more substantial holding times, the effect of the holds is very small on the trajectories of subsequent buses.

For illustrative purposes, a total of 50 simulations were conducted for seven cases:

1. No holding is performed.
2. Holding is performed using the formulation in (47), with the associated constraints (the "full model").
3. Holding is performed using the formulation in (47), with the associated constraints, *but excluding the variance terms* ("no variance"). This emulates a strictly deterministic model.
4. Holding is performed using a threshold strategy, with a threshold of 4.5 minutes.
5. Holding is performed using a threshold strategy, with a threshold of 5.0 minutes.
6. Holding is performed using a threshold strategy, with a threshold of 5.5 minutes.
7. Holding is performed using a threshold strategy, with a threshold of 6.0 minutes.

The threshold models were included to compare the performance of the proposed analytic technique with a traditional threshold technique.

Within each simulation run, the following performance measures were calculated: (1) the total waiting time experienced by passengers; (2) the number of holds enacted; (3) the duration of each hold; and (4) the total delay experienced by passengers on board the bus during each hold.

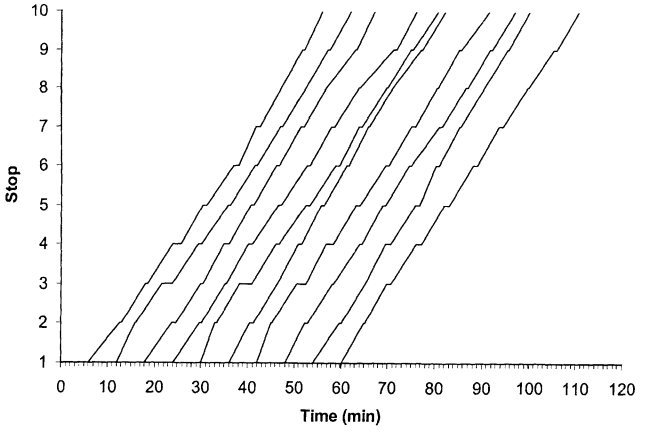

Figure 4 Example Bus Trajectories with Holding

The results of the simulation are discussed below. First, the results for the full model are presented. Second, the results from the full model are compared with those from the "no variance" model and with the threshold models. Finally, a sensitivity analysis of the full model to selected model parameters is also presented.

4.2. Simulation Results

Table 5 gives the *observed* value of the objective (waiting time plus one-half the on-board delay) for all 50 simulation runs. It is notable that the full model yields a total objective that is higher than the no-holding case in 20 runs. This suggests that, while justified in terms of *expected* passenger waiting time and delay, the holding actions may in many cases result in an actual value that is worse than if no holding action were taken (40% in this example). While a full analysis of this phenomenon is beyond the scope of this paper, this suggests that on a day-to-day basis, universally applied holding strategies may not always improve passenger level of service.

At the same time, the distribution of time savings from holding is skewed toward positive values, as illustrated in Figure 5. Specifically, one notes that holding produces the largest reductions in the objective in the simulation runs where the waiting times are highest. In over one-third of the cases (17 of 50, or 34%), the reductions in the objective exceed 100 minutes (5% of the total waiting time, or 27% of the excess waiting time due to headway variability). In nine cases (18%), the reductions in the objective exceed 200 minutes (about 10% of the total waiting time, or 55% of the excess waiting time due to headway variability). This supports the intuition that the value of holding as a control action appears greatest when the variability of headways is highest. Further experimentation is necessary to confirm this intuition with the proposed model.

Summary statistics from the 50 simulations runs, for each of the seven scenarios, are shown in Table 6. Compared with the no-holding case, all holding

Table 5 Objective Values from Simulation (in Passenger Minutes)

Simulation Number	No Hold	Full Model	No Variance	Threshold 4.5 min	5.0 min	5.5 min	6.0 min
1	1901.5	2008.7	2008.7	1942.1	1984.4	2053.8	2194.3
2	1867.0	1928.0	1914.4	1887.0	1930.7	2024.3	2123.6
3	2449.2	2148.2	2151.7	2252.4	2079.8	2038.8	2054.8
4	2130.7	2099.9	2099.9	1991.5	2033.9	2165.3	2231.5
5	2083.0	2078.0	2079.0	2083.0	1973.7	1992.9	1985.0
6	1959.8	2171.1	2131.0	2108.4	2167.2	2221.6	2295.5
7	2121.9	1895.9	1904.0	1897.2	1864.5	1863.8	1873.9
8	2147.5	2235.5	2221.1	2147.5	2135.4	2196.4	2128.9
9	2352.8	2372.3	2376.6	2384.9	2423.1	2424.9	2374.3
10	2258.5	2248.3	2248.3	2259.1	2229.3	2393.8	2554.2
11	1979.1	2034.6	2025.5	1965.6	1966.0	2014.1	2040.5
12	1980.3	1980.3	1980.3	1980.3	1981.6	1855.5	1872.2
13	1974.5	1942.8	1938.1	1942.6	1941.0	1900.3	1949.9
14	2474.7	2308.5	2350.1	2456.4	2455.2	2383.6	2393.9
15	2027.5	2133.9	2140.9	2035.2	2083.1	2129.6	2136.6
16	1777.5	1713.3	1717.9	1795.0	1782.2	1826.6	1852.1
17	1890.0	1902.1	1909.0	1890.0	1908.4	1869.9	1832.9
18	2520.9	2521.8	2491.5	2521.5	2491.4	2480.2	2498.4
19	2354.7	2310.5	2300.9	2340.5	2312.2	2304.0	2266.7
20	1976.2	1932.0	1922.8	1983.5	1998.0	1960.7	2045.2
21	2313.7	2154.0	2154.0	2199.3	2219.4	1992.6	2098.2
22	1997.4	2069.7	2076.0	1997.4	2001.9	2057.7	2149.9
23	2257.2	2411.9	2400.0	2400.3	2383.2	2428.2	2356.8
24	2023.9	2015.9	2019.8	2009.9	2060.0	2110.2	2136.1
25	2483.3	2252.7	2239.7	2357.1	2314.1	2373.0	2235.8
26	2452.9	2237.2	2239.6	2453.4	2264.5	2175.5	2158.5
27	1854.2	1909.2	1909.6	1874.0	1893.9	1944.5	2050.1
28	2232.6	2225.4	2215.2	2230.4	2219.3	2214.4	2264.8
29	2157.1	1953.5	2005.9	2073.2	1996.1	1992.7	1891.8
30	2027.7	1993.8	1993.8	2015.4	2021.4	2089.4	2124.0
31	2015.1	2018.0	2018.0	1940.9	1941.1	2057.4	2135.9
32	1976.4	1989.5	1999.7	1972.7	1977.1	1943.9	1949.6
33	2254.8	2234.3	2248.3	2235.6	2226.0	2240.0	2408.0
34	1866.9	1762.7	1784.3	1831.4	1762.8	1847.3	1866.7
35	2423.4	2194.0	2196.1	2270.5	2312.2	2175.5	2329.6
36	1903.6	1785.7	1785.7	1776.3	1755.9	1840.9	1830.5
37	2043.4	1917.2	1917.2	1896.7	1868.7	1865.2	1834.7
38	2263.7	2012.6	2013.9	2010.9	2001.5	1996.8	2004.0
39	2178.7	2044.4	2044.4	2165.1	2152.3	2194.1	2117.5
40	2160.1	2185.3	2180.2	2210.3	2230.7	2255.8	2354.1
41	1957.6	1927.0	1953.4	1957.6	1950.5	1927.6	1952.4
42	2090.8	2153.8	2105.0	2102.3	2146.5	2179.5	2199.9
43	2014.7	1884.7	1889.0	2007.0	1967.8	1941.9	1980.5
44	2149.1	2190.7	2203.3	2140.8	2186.0	2284.4	2315.5
45	2325.3	2075.5	2074.9	2014.6	2076.8	2222.4	2276.7
46	1690.3	1691.8	1699.8	1690.3	1690.3	1661.8	1663.9
47	1965.8	1973.6	1983.8	1968.3	1990.6	1981.6	2052.5
48	2103.1	2147.6	2128.3	2206.4	2167.8	2176.6	2083.9
49	2246.4	2105.4	2104.3	2227.0	2187.1	2140.9	2172.4
50	2378.9	2100.9	2114.4	2144.9	2063.1	2072.4	2102.7

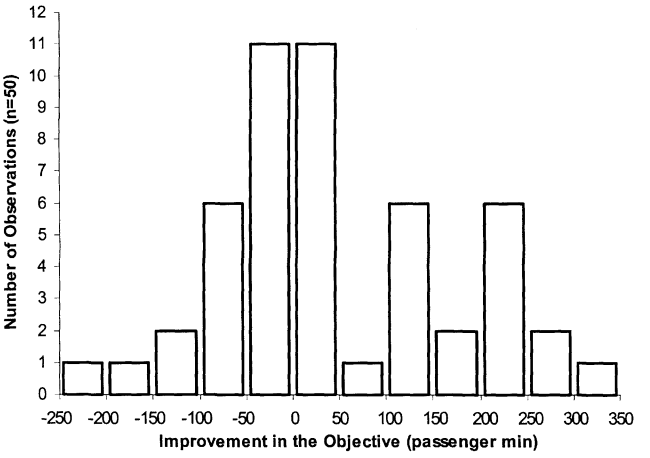

Figure 5 Histogram of Objective Improvement: No Holding—Full Model

strategies reduced both the average objective value and the average total waiting time. The full model gave an average reduction in passenger waiting time of 73.2 minutes, about 3.5% of the total waiting time, per scenario. In a different light, when compared with the total added waiting time due to headway variability (2120.7 minutes − 1755.0 minutes = 365.7 minutes), the holding model reduces this additional waiting time by 20.0%. These potential savings are considerable. Including the delay caused to passengers on-board while holding with a weight of 0.5, the full model still results in savings of 49.0 minutes, about 2.3% of the total waiting time or 13.4% of the incremental waiting time due to headway variability.

4.3. Model Comparison

There are several interesting observations made by comparing the results across the different holding models. First, from Table 6, the performance of the deterministic holding model ("no variance") is only slightly worse than the full model, based on the total waiting time plus on-board delay. This result is not statistically significant and indicates that, in total, the two models may be very similar in the net result. What is also notable in Table 6 is the number and duration of holds for these two holding models. While the objective function values are similar, the full model results in about 25% fewer holds than the "no variance" model. The full model results in holds being applied to 30% of the buses, while over 40% of the buses are held in the "no variance" case.

Table 6 Summary of Holding Time Results

	No Holding	Full Model	No Variance	Threshold 4.5 min	Threshold 5.0 min	Threshold 5.5 min	Threshold 6.0 min
Avg. waiting time per simulation (min)	2120.7	2047.5	2046.2	2075.6	2058.4	2060.6	2067.2
Avg. on-board delay per simulation (min)		48.4	52.0	18.5	33.9	58.2	97.6
Total objective (min)	2120.7	2071.7	2072.2	2084.9	2075.4	2089.7	2116.0
Total holds ($n = 500$)		150	201	85	135	199	288
% of Buses held		30.0%	40.2%	17.0%	27.0%	39.8%	57.6%
Avg. hold time (min)		1.27	0.99	0.93	1.00	1.10	1.21
Avg. on-board delay per hold (pass.-min)		16.4	12.9	10.9	12.6	14.6	16.9
Avg. no. of on-board passengers per hold		12.9	13.0	11.7	12.5	13.3	14.0

Moreover, the average duration of the holds are considerably longer for the full model (1.27 minutes/hold) compared with the "no variance" model (0.99 minutes/ hold).

To explore this phenomenon further, Figure 6 gives a distribution of the holding times for the full model and the "no variance" model, at a 0.25 minute (15 second) level of resolution. One notes in this figure that the full model produces few holds under 0.5 minutes (30 seconds) in duration. In contrast, the "no variance" model produces a much larger number of holds in the 0–0.75 minute range (0–45 seconds). Interestingly, such short holds appear to have only a very modest effect, if any, on the total objective. The most obvious explanation for the difference in short holds is the effect of variance on the total objective. As noted in (32) and (40), the headway variance increases with t. Under very small values of the holding time t, these increases in headway variance more than compensate for small reductions in waiting time due to changes in the expected headways. As a result, short holds may be relatively ineffective in improving passenger service in a stochastic transit operating environment.

It appears that longer holding times (over 0.5 or 0.75 seconds) have greater effect in reducing passenger waiting times. Interestingly, from Figure 6, the number of holds in the full model match those in the "no variance" case fairly closely with holding times of at least 0.75 minutes. In fact, for holds of 0.75 minutes or more, the two models select identical situations to apply holds, and the optimal holding times selected by each of the two models differ by at most 0.05 minutes. As a result, both models seem to perform equally well when longer holds are necessary. This result is due to the dominance of the expected headway terms in the objective (47) as the holding time t increases.

One may also compare the proposed model with the various threshold policies. From Table 6, the objective value with the 5-minute threshold policy also

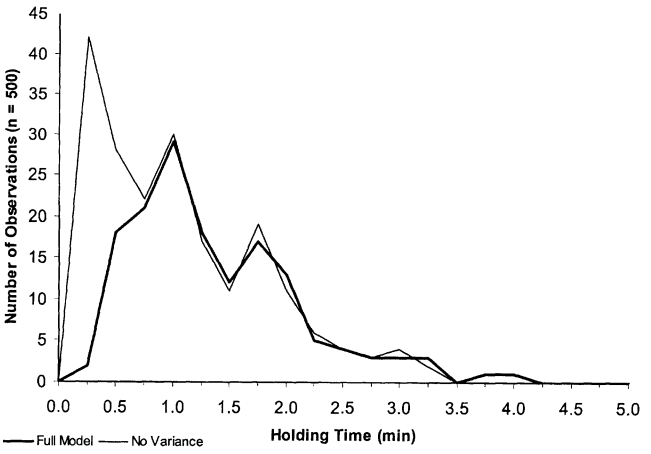

Figure 6 Distribution of Holding Times, Full Model and "No Variance" Model

is close to the minimum objective value with the full model. This indicates that a threshold policy may also perform reasonably well, in aggregate, in this instance. The percentage of buses held in the threshold models ranges from 17% for the 4.5-minute threshold to 57.6% for the 6-minute threshold. As may be expected, the average holding times and average on-board delay also increases as the threshold value increases.

One may compare the circumstances under which the full model and the threshold models might apply a hold. Descriptive statistics are shown in Table 7, giving the total number of holds for each technique, the number of holding opportunities in which both the full model and the threshold model agree to apply a hold, and this number as a percentage of all holds under each threshold policy. From the table, the number of opportunities where both the full model and the threshold model choose to apply a hold increases as a function of the threshold value. However, as a percentage of all holds under each threshold policy, these cases decrease with the threshold value.

These results suggest that threshold policies can be detrimental to service, and that the percentage of holds that are ineffective rises with the value of the threshold. Specifically, one may look at the percentage of cases in which no hold should be applied ($t = 0$ in the full model), but a hold is applied using the threshold model. In these situations, the objective function value will clearly increase with a hold. The percentage of such cases rises from 9% (8 of 85) of the holds for the 4.5-minute threshold, to 19% (25 of 135) for the 5.0-minute threshold, to 33% (66 of 199) for the 5.5-minute threshold, to a full 51% (148 of 288) for the 6.0-minute threshold. Moreover, these percentages are lower bounds, in the sense that even in some cases when both the full model and the threshold model apply a hold, the threshold value may also result in an increase in the objective function.

Table 7 Comparison of Holds, Full Model, and Threshold Models

		Cases Where Full Model and Threshold Model Apply a Hold	
Holding Policy	Total Holds ($n = 500$)	No. of Cases	% All Holds with Threshold Policy
Full model	150		100
4.5-min threshold	85	77	91
5.0-min threshold	135	110	81
5.5-min threshold	199	133	67
6.0-min threshold	288	140	49

Figure 7 gives the distribution of the holding times for the full model and each of the threshold models, again at a 0.25-minute (15-second) level of resolution. One notes immediately that there is little correspondence between the holds applied for the full model and those for the threshold models. Since the threshold values are not directly tied to an objective, there is no reason to expect any correspondence. Also, the threshold models produce a substantial number of short holds that are likely to have only a very modest effect on passenger waiting times. In summary, based on these experimental results, the threshold policies appear to be inferior to the proposed model in determining both whether and how long to hold a vehicle.

4.4. Sensitivity Analysis

To illustrate some additional features of the full model, a sensitivity analysis is presented with respect to the weight on on-board delay and the choice of control stop. First, one may examine the effect of holding on the passengers on board at the control point (stop 3). According to Table 6, the average waiting time savings per simulation run ranges from 45.1 to 74.5 minutes across the various holding techniques.

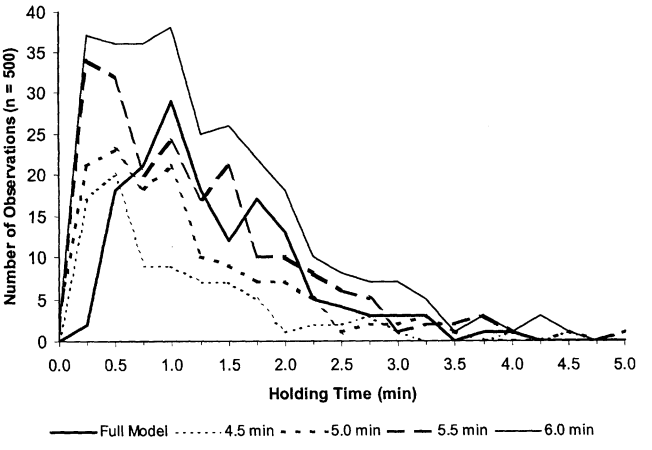

Figure 7 Distribution of Holding Times: Full Model and Threshold Models

In comparison, the average delay for passengers on board ranges from 18.5 to 97.6 minutes per simulation run. This indicates that the delay to passengers on board can be a significant "cost" associated with holding, and that holding actions with this model are likely to be sensitive to the value of θ, the weight on on-board delay.

Table 8 gives the number of holds, the percentage of buses held, and the average holding time for θ values between 0.0 and 1.0. The case $\theta = 0.0$ corresponds to the holding problem where the delay to passengers on board is disregarded, while the case $\theta = 1.0$ corresponds to equal weight to on-board delay and to downstream passenger waiting. As one might expect, as θ increases, the total number and percentage of buses held decreases. For this example, over 37% fewer holds are made with $\theta = 1.0$ versus $\theta = 0.0$, with the percentage of buses held dropping from 38.6% to 24.2%.

One might expect a reduction in both the number of holds and the length of the holds. However, for this example, the average length of a hold does not change noticeably over this range of θ. Additional insight is offered in Figure 8, which gives the distribution of holding times for representative values of θ. In the figure, the distribution of holding times does shift to the left (toward lower values of hold time) as θ increases. However, because the full model discourages very short holds (e.g. under 0.5 minutes), the average holding time does not change substantially. Hence, an increase in the weight θ reduces the number of situations when holding is applied, but has a modest or negligible effect on the duration of a hold.

Table 8 Sensitivity of Full Model to Parameter θ

Weight	Total Holds ($n = 500$)	% Buses Held	Avg. Hold Time (min)
0.0	193	38.6	1.28
0.1	181	36.2	1.29
0.2	173	34.6	1.28
0.3	167	33.4	1.26
0.4	158	31.6	1.27
0.5	150	30.0	1.27
0.6	143	28.6	1.27
0.7	138	27.6	1.25
0.8	134	26.8	1.23
0.9	126	25.2	1.25
1.0	121	24.2	1.24

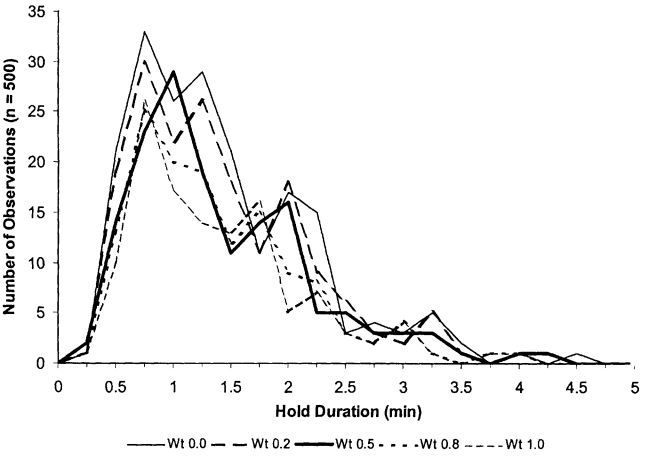

Figure 8 Distribution of Holding Times by Weight θ

Another sensitivity analysis examines the choice of the holding stop. With the original technique suggested by Turnquist and Blume (1980), stops 2 and 3 in this example are candidates for holding. However, adapting their technique to include the weight $\theta = 0.5$, stop 4 in the example also is an eligible control stop. Table 9 compares the holding experience at stop 3 with that at stop 4.

In the example, stop 4 has the highest boarding rate of all stops on the route, and the load leaving stop 4 is also very high. Hence, holding at stop 4 inconveniences a significant number of passengers, and the average on-board delay is more than twice as high as for holding at stop 3. This effect is compounded by the longer average holding time at stop 4 (1.70 minutes versus 1.27 minutes at stop 3). Nonetheless,

Table 9 Sensitivity to Control Stop

	Control Point	
	Stop 3	Stop 4
Avg. waiting time per simulation (min)	2047.5	1963.7
Avg. on-board delay per simulation (min)	49.3	110.3
Total objective (min)	2072.1	2018.8
Total holds ($n = 500$)	150	146
% of buses held	30.0	29.2
Avg hold time (min)	1.27	1.70

the expected downstream waiting time and the total objective value are substantially lower by enacting holding at stop 4. This is noteworthy, in that the intuition from previous studies suggests that holding should be applied *upstream* of the maximum boarding point. In this example, holding at the maximum boarding point is justified and is even more beneficial than at an upstream stop. This effect is due to the discounting of on-board delay using $\theta = 0.5$ and also to the parameters of this example. Nonetheless, this serves to emphasize that the screening technique recommended by Turnquist and Blume should be followed by more detailed analysis of the benefits of control at each candidate stop.

5. Conclusions

This paper provides an analytic model for the vehicle holding problem that explicitly incorporates stochastic service elements. In this respect, it presents a fundamental improvement on the analytic models from the 1970s that explored holding strategies using transit service models that are too simple to be practical for most transit agencies. It also presents a substantial improvement over recent deterministic holding models by explicitly incorporating stochastic effects. Hence, in contrast to previous models, the operations model and holding strategy formulations explicitly consider stochastic running and dwell times, and use these stochastic effects to formulate holding strategies that are sensitive to the variability in vehicle movements on a fixed route.

The example illustrates how the model may be applied, and significant reductions in passenger waiting time are possible through such holding actions. Moreover, the inclusion of stochastic elements in the holding model significantly reduces the number of short holds, over both the deterministic and threshold-based policies. Rather, short holds may provide small improvements when considered deterministically but actually may result in higher headway variability that reduces or eliminates the effectiveness of the hold. Conversely, deterministic and threshold-based policies, when applied for short holds in a truly stochastic transit operating environment, may be ineffective or even detrimental to service.

Finally, the holding model in this paper may be used directly in a decision support context to assist transit operators and/or supervisors in making optimal holding decisions in day-to-day operations. A more detailed case study is planned to validate this model and to show its application in a context where real-time vehicle location information may be available. Naturally, the application of this model for a particular transit agency would be far more illustrative of the potential benefits of the model than the simple example here. Nonetheless, the analytic model presented here is a more general and flexible tool than traditional simulation tools, the results of which may not be useful in a real-time decision-making context.

The model presented in this paper is intended to be an initial effort in developing analytic models of this type. For this reason, it is important to keep in mind that although this holding model is useful for decision support, the model presented in this paper has several critical assumptions, some of which are likely to be violated in practice. The most egregious assumptions, which may serve as areas of future research, include:

- *Scope of holding decision.* The decision model described in this paper covers only the decision to hold a vehicle, made once the vehicle arrives at a control stop. In only solving for a single vehicle, the model is necessarily "myopic" in that it optimizes only a single headway, instead of solving for all subsequent holds (i.e., all subsequent vehicle headways). In this way, each holding decision is justified on its own merits, independently of other control actions (as suggested by Turnquist 1989).

A model incorporating multiple vehicles could also be applied using Marguier's transit operations model, but the initial conditions for the vehicle dynamics in (36–46) become significantly more complicated. In such a case, however, one might wish to apply such a "universal" model sequentially in a real-time decision context. That is, one might determine an optimal set of headways for all vehicles, but only apply a hold (if a hold is optimal) to the current vehicle. A new optimization would be conducted when the next vehicle arrives at the control stop. The value of a "universal" model, versus the model proposed here, deserves further investigation.

• *Vehicle capacity*. In this paper, vehicle capacity is assumed to be infinite. However, in many cases, transit service delays and vehicle bunching can lead to loads up to the effective capacity of a vehicle, and passengers may be left to wait for subsequent vehicles. The effect of capacity restrictions on holding policies, and on transit operational control strategies more generally, significantly complicates the analysis. While the research of O'Dell and Wilson and Shen and Wilson has provided tractable models for rail operations in a deterministic setting, incorporating realistic capacity constraints in an analytic model remains an important area for further research.

• *Passenger boarding and alighting*. The incremental boarding and alighting times per passenger are assumed to be constant, resulting in dwell times that are linear in the number of passengers boarding and alighting. In some cases, particularly with passengers with special needs, the assumption of constant boarding and alighting times may be violated. However, an assumption of stochastic boarding and alighting times would result in a nonlinear service model and, hence, would be much more difficult to deal with analytically. Also, the separation of alighting and boarding processes in this analysis may be violated if these processes occur in parallel. However, parallel processes would result in only minor changes to the model presented here.

A second area of concern is the passenger boarding and alighting distributions. A substantial amount of group travel will affect both the passenger arrival process to bus stops and the assumed alighting behavior at subsequent stops. The passenger arrival and alighting processes may also be related to specific times of the day. There is a need for more research about the characteristics of passenger trip patterns and about how these characteristics may be included in microscopic models of bus operations.

• *Vehicle running times*. The assumptions of a simple distribution of running times between transit stops, and the independence of consecutive vehicle running times, are suspect. These require further empirical study and analysis. However, the assumption of vehicle-specific running time distributions could be captured in the model without loss of linearity. Including correlated running times, however, appears to offer significant complications to this model.

Acknowledgments

The author is grateful to Dr. Philippe Marguier for the insight from his dissertation, without which this research would not have been possible. The author is also grateful to Professor Nigel Wilson of MIT, who brought Dr. Marguier's dissertation to his attention.

References

Abkowitz, M. 1980. The Impact of Service Reliability on Work Travel Behavior. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

———, A. Eiger, I. Engelstein. 1986. Optimal control of headway variation on transit routes. *J. Adv. Transportation* **20**(1) 73–88.

———, I. Engelstein. 1984. Methods for maintaining transit service regularity, *Transportation Res. Record* **961** 1–8.

———, H. Slavin, R. Waksman, L. Englisher, N. Wilson. 1978. *Transit Service Reliability: Final Report*. Technical Report MA-06-0049-78-1, Urban Mass Transportation Administration.

———, J. Tozzi. 1986. Transit route characteristics and headway-based reliability control. *Transportation Res. Record* **1078** 11–16.

Adamski, A. 1979. Optimal Dispatching Control of Bus Lines. K. Iracki, K. Malanowski, S. Walukiewicz, eds. *Optimization Techniques: Proc. of the 9th IFIP Conf. on Optim. Techniques, Part 2*. Springer-Verlag, Berlin, 334–344.

———. 1983. Optimal Dispatching Control of Bus Lines. D. Klamt, R. Lauber, eds. in *Control in Transportation Systems: Preprints from the 4th IFAC/IFIP/IFORS Conf.* VDI/VDE, 67–72.

———. 1992. Probabilistic models of passengers service processes at bus stops. *Transportation Res.* **26B**(4) 253–259.

———. 1993. Real-time computer aided adaptive control in public transport from the point of view of schedule reliability, J. R. Daduna, I. Branco, J. M. P. Piaxao, eds. *Computer-Aided Transit Scheduling: Proceedings.* (July) 278–295.

———. 1996. Flexible dispatching control tools in public transport. L. Bianco, P. Toth, eds. *Advanced Methods in Transportation Analysis*. Springer, Berlin, 481–506.

———, A. Turnau. 1998. Simulation support tool for real-time dispatching control in public transport. *Transportation Res.* **32A**(2) 73–87.

Adebisi, O. 1986. A mathematical model for headway variance of fixed-route buses. *Transportation Res.* **20B**(1) 59–70.

Andersson, P.-A., A. Hermansson, E. Tengvald, G-P. Scalia-Tomba. 1979. Analysis and simulation of an urban bus route. *Transportation Res.* **13A** 439–466.

———, G-P. Scalia-Tomba. 1981. A mathematical model of an urban bus route. *Transportation Res.* **15B**(4) 249–266.

Barnett, A. 1974. On controlling randomness in transit operations. *Transportation Sci.* **8** 102–116.

———. 1978. Control strategies for transport systems with nonlinear waiting costs. *Transportation Sci.* **12**(2) 119–136.

———, D. J. Kleitman. 1978. On two-terminal control of a shuttle service. *SIAM J. Appl. Math.* **35**(2) 229–234.

Bly, P. H., R. L. Jackson. 1974. Evaluation of bus control strategies by simulation. Technical Report 637, Transportation and Road Research Laboratory, Crowthorne, Berkshire, UK.

Bowman, L. A. 1979. Analysis of Network Effects on Bus Transit Reliability and the Potential for Real Time Control. Ph.D. thesis, Northwestern University, Evanston, IL.

———, M. A. Turnquist. 1981. Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops. *Transportation Res.* **15A**(6) 465–471.

Boyd, C. W. 1983. Notes on the theoretical dynamics of intermittent public passenger transportation systems. *Transportation Res.* **17A**(5) 347–354.

Casey, R. 1999. *Advanced Public Transportation Systems Deployment in the United States: Update, January 1999.* Technical Report FTA-MA-26-7007-99-1, Federal Transit Administration, Washington, DC.

———, L. Labell, L. Moniz, J. Royal, M. Sheehan, T. Sheehan, A. Brown, M. Foy, M. Zirker, C. Schweiger. B. Marks, B. Kaplan, D. Parker. 2000. *Advanced Public Transportation Systems: The State of the Art—Update 2000.* Technical Report FTA-MA-26-7007-00-1, Federal Transit Administration, Washington, DC.

Ceder, A., P. Marguier. 1985. Passenger Waiting Time at Transit Stops. *Traffic Engrg. Control* **26**(6) 327–329.

Chapman, R. A., J. F. Michel. 1978. Modelling the tendency of buses to form pairs. *Transportation Sci.* **12**(2) 165–175.

Eberlein, X. J. 1995. Real-time control strategies in transit operations: Models and analysis. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

———, N. H. M. Wilson, D. Bernstein. 2001. The holding problem with real-time information available. *Transportation Sci.* **35**(1) 1–18.

Furth, P. 1995. A headway control strategy for recovering from transit vehicle delays. *Transportation Congress: Civil Engineers—Key to the World's Infrastructure,* vol. 2. 2032–2039.

Goeddel, D. 1996. Benefits assessment of advanced public transportation systems. Technical Report DOT-VNTSC-FTA-96-7, Federal Transit Administration, Washington, DC.

Guenthner R., K. Sinha. 1983. Modeling bus delays due to passenger boardings and alightings. *Transportation Res. Record* **915** 7–13.

Heap R. C., T. H. Thomas. 1976. The modelling of platooning tendencies in public transport. *Traffic Engrg. Control* **17** 360–362.

Jolliffe J. K., T. P. Hutchinson. 1975. A behavioural explanation of the association between bus and passenger arrivals at a bus stop. *Transportation Sci.* **9**(3) 248–282.

Koffman, D. 1978. A simulation study of alternative real-time bus headway control strategies. *Transportation Res. Record.* **663** 41–46.

Levinson, H. 1991. *Supervision Strategies for Improved Reliability of Bus Routes.* Synthesis of Transit Practice 15, National Cooperative Transit Research and Development Program. Transportation Research Board, Washington, DC.

Lin, G., P. Liang, P. Schonfeld, R. Larson. 1995. *Adaptive Control of Transit Operations.* Technical Report FTA-MD-26-7002, Federal Transit Administration.

Marguier, P. H. J. 1985. Bus Route Performance Evaluation under Stochastic Conditions. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

———, A. Ceder. 1984. Passenger waiting strategies for overlapping bus routes. *Transportation Sci.* **18** 207–230.

Newell, G. F. 1974. Control of pairing of vehicles on a public transportation route: Two vehicles, one control point. *Transportation Sci.* **8**(3) 248–264.

———. 1976. Unstable Brownian Motion of a Bus Trip. Uzi Landman, ed. *Statistical Mechanics and Statistical Methods in Theory and Application.* 645–667.

———, R. B. Potts. 1964. Maintaining a bus schedule. *Second Conf. Australian Road Res. Board* **2**(1) 388–393.

O'Dell, S., N. H. M. Wilson. 1999. Optimal real-time control strategies for rail transit operations during disruptions. N. H. M. Wilson, ed. *Computer-Aided Transit Scheduling: Lecture Notes in Economics and Mathematical Systems No. 471.* Springer-Verlag, Berlin, Germany. 299–323.

Osuna, E. E., G. F. Newell. 1972. Control strategies for an idealized public transportation system. *Transportation Sci.* **6** 52–72.

Potts, R. B., E. A. Tamlin, 1964. Pairing of buses. *Second Conf. Australian Road Res. Board* **2**(2) 3–9.

Powell W. B., Y. Sheffi. 1983. A probabilistic model of bus route performance. *Transportation Sci.* **17**(4) 376–404.

Saidi, S., M. Gendreau, F. Soumis. 1994. Using Dynamics in Passenger Behavior to Model Capacity in Congested Transit Networks. Technical Report 94-42, Centre de Recherche sur les Transports, Universite de Montreal, Montreal, Canada, August.

Schiavone, J. 1999. *Understanding and Applying Advanced On-Board Bus Electronics.* TCRP Report 43, Transit Cooperative Research Program, Transportation Research Board, Washington, DC.

Senevirante, P. N. 1990. Analysis of on-time performance of bus services using simulation. *Transportation Engrg.* **116**(4) 517–531.

Shen, S., N. H. M. Wilson. 2000. An optimal integrated real-time disruption control model for rail transit systems. Preprints from *8th Internat. Conf. on Computer-Aided Scheduling of Public Transport.*

Turnquist, M. A. 1978. A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Res. Record* **663** 70–73.

———. 1981. Strategies for improving reliability of bus service. *Transportation Res. Record* **818** 7–13.

———. 1989. Real-time control for improving transit level-of-service. C. Hendrickson, K. Sinha, eds. *First Internat. Conf. on Appl. of Adv. Technologies in Transportation Engrg.* 217–222.

———, S. W. Blume. 1980. Evaluating potential effectiveness of headway control strategies for transit systems. *Transportation Res. Record* **746** 25–29.

Vandebona U., Richardson, A. 1986. Effect of checkpoint control strategies in a simulated transit operation. *Transportation Res.* **20A**(6) 429–436.

Welding, P. I. 1957. The instability of a close-interval service. *Opera. Res. Quart.* **8**(3) 133–148.

Received: July 1998; revisions received: July 1999, June 2000; accepted: March 2001.